Uncovering genomic and transcriptomic complexities in two important Australian aquaculture species; the Sydney rock oyster (*Saccostrea glomerata*) and the banana shrimp (*Fenneropenaeus merguiensis*)

Daniel Edward Powell  
BSc (Microbial Ecology), BSc (Hons), GCIntBus

A thesis submitted for the degree of  
Doctor of Philosophy  
June 2017

Centre for Genetics, Ecology and Physiology  
& School of Science and Engineering  
Faculty of Science, Health, Education and Engineering  

University of the Sunshine Coast  
Sippy Downs, Queensland, Australia
Abstract

Despite some recent progress towards sustainable harvesting, the world’s marine fisheries continue in a state of decline. World food fish production from aquaculture has expanded almost 12 fold over the last three decades as a response, though aquaculture production is vulnerable to adverse impacts such as changing environmental conditions and disease. Methods to secure and improve production and quality are desperately needed if aquaculture is to continue to enhance the world’s available food fish supply. The application of biotechnology in aquaculture is one of the most promising avenues for developing solutions to these challenges. However, a major factor affecting its advancement is the lack of molecular resources. The recent revolution in high-throughput, next-generation sequencing technologies has placed the development of such resources within the reach of most research groups. This PhD study aims to develop new resources for two important species to the Australian aquaculture industry using these next-generation technologies.

The first in-depth examination of the transcriptional content of 8 different tissues from cultured banana shrimp was undertaken using RNA-Seq technologies. A combination of Roche 454FLX and illumina HiSeq 2000 sequencing data was used to produce a reference assembly of 124,631 transcripts from which a total of 59,179 protein sequences were predicted. Further analysis revealed a rich set of transcript sequences exhibiting homology with genes associated with reproduction, sex determination and development and distinguished the tissues responsible for this expression. These data add a substantial contribution to the sequence information made available for *F. merguiensis*.

As viral pathogens pose major challenges to the culture of penaeid shrimps worldwide, the study of cultured banana shrimp infected with hepatopancreatic parvo-like virus (HPV) provides a unique opportunity to explore the host response to viral infection independent of challenge testing. To gain some insight into the genetic mechanisms underlying resistance to high levels of HPV, we examined hepatopancreas tissue from 6 full-sib groups of banana shrimp with differing levels of HPV infection for differences in gene expression. A total of 404 differentially expressed genes were identified among high- and low-HPV full-sib groups. Based on homology analysis, a large proportion of these genes were associated with processes involved in the immune response of crustaceans, including pattern recognition proteins, antimicrobial peptides and antiviral activity. The results indicate shrimp from high-HPV full-sib groups appear to have a lower presence of important immune response elements, yet possess upregulated putative antiviral pathways. Within the differentially expressed genes, over 4,000 sequence variants were identified to be exclusive to either the high- or low-HPV full-sib groups. This work has provided additional understanding of the mechanisms involved in the response of this shrimp species to a naturally occurring viral pathogen. Furthermore, sequence variants identified in this study can be exploited for the development of gene-associated markers of HPV resistance.

Penaeid shrimp exhibit a female-superior sexual dimorphism with respect to growth rate. Uncovering the molecular mechanisms governing sexual dimorphism can assist with the enhancement of profitability of penaeid shrimp culture by enabling the development of a monosex (all-female) populations and help to better
understand nutrient assimilation and utilisation which may lead to improved feeding technologies. In crustaceans, the hepatopancreas is important for the absorption and storage of nutrients for tissue growth and energy consumption during the growth stage. To explore differences in hepatopancreatic gene expression between male and female banana shrimp, a test for differential gene expression was performed to compare the digestive gland tissue from 14 male and 14 female banana shrimp (n=28) of a size within the dimorphic growth range. Female banana shrimp tissue was found to have an increased expression of transcripts coding for chitin catabolism and binding. A novel transcript was also identified that displayed a distinct female-specific pattern of expression, though exhibited no distinguishable features recognisable by alignment against the public sequence databases. Weighted gene co-expression network analysis (WGCNA) involving female samples revealed this novel transcript was highly connected within a gene module containing an abundance of transcripts with homology to transcription factors and genes associated with growth regulation.

The Sydney rock oyster is considered a delicacy in restaurants worldwide and is an economically important cultured marine species in Australia. Contending with disease and environmental stress are considerable challenges to oyster culture, however, like many other shellfish, little is known of the molecular mechanisms underpinning these traits. The generation of a high-quality draft genome can offer insight into these mechanisms, although it also presents difficulties due to the high levels of heterozygosity and repetitiveness observed in oysters. A draft of the estimated 784Mb S. glomerata genome is presented using over 300-fold coverage of illumina small insert paired-end and mate-pair libraries with an additional Chicago library and HiRise scaffolding. The draft assembly has a total sequence length of 788 Mb, a scaffold number of 10,107 and a scaffold N50 of 804 Kb. Analysis of the predicted 29,738 protein-coding genes identifies expansions of gene families associated with innate immunity and stress response and, together with gene expression data, provides insight into the genomic background of this oyster’s resilience to abiotic stress. This Saccostrea genome provides a valuable resource for further studies in molluscan biology, offering support for genetic enhancement programs for commercially produced oyster species.

The work reported in this thesis provides a substantial increase in the molecular resources available for two important Australian aquaculture species and further demonstrates the utility of high-throughput sequencing technologies coupled with bioinformatics analysis to advance molecular genetic studies of non-model species.
Declaration of originality

This thesis describes my original work undertaken at the University of the Sunshine Coast and, to the best of my knowledge and belief, contains no material previously published or written by another person except where due reference has been made. The content of this thesis does not include work that has been submitted to qualify for the award of any other degree in any other tertiary institution.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and immediately made available in accordance with the Copyright Act 1968.

I acknowledge that copyright of all material contained in this thesis resides with the copyright holder(s) of that material.

Signature

...........................................................

Daniel Powell

June 2017
Acknowledgements

Having ambitiously taken a full-time approach to this study, to get it there did require a little help. I would firstly like to thank my primary advisor Professor Abigail Elizur. This work could not have been achieved without your unwavering support and encouragement throughout the entire degree. I greatly appreciate the independence I was offered and to have been invited to join such interesting and leading edge projects that have undoubtedly refined my skills and passion for research.

I would like to thank Associate Professor Wayne Knibb for the opportunity to be a part of the banana shrimp project that ultimately comprised a large component of this thesis. Thanks also to my co-advisor Dr Min Zhao and Dr Nguyen Nguyen for assistance with data analysis and statistics over the course of this study, and to Associate Professor Neil Tindale for helping with advice and admin in the earlier stages of this degree. I would like to thank Professor David Raftos and Dr Wayne O’Connor for their support and contributions to the oyster genome project and for their patience and confidence in my work as it progressed.

My fellow students and colleagues at USC, there are far too many of you to list, but a big thanks to you all for your help and understanding from the very early days right up until these final moments.

I wish to thank Genecology for financial support from publication rewards and to the Faculty and its research support via internal competitive HDR grants from which I was awarded two. This support enabled me to attend world class training and conference events that were highlights of my time as a student and invaluable to my career. I would also like to gratefully acknowledge the University of the Sunshine Coast for the Post Graduate Research Scholarship awarded to support my PhD candidature and the School of Science and Engineering for additional operational funding.

A huge thank you to my family for their support and laughs over these last 4 years, nothing melts away the weeks of being bound to a computer than some stupid moments with you guys.

Finally, I extend my sincere gratitude and appreciation to my gorgeous partner Bec for your unfaltering and considerate support and understanding over these very busy 4 years. I couldn’t have done this without you.
List of publications


This paper (I) forms Chapter 2 of this thesis; **DP** contributed to the sample preparation, coordinated the sequencing, performed all analyses and wrote the manuscript. **WK** and **CR** provided funding and resources and reviewed the manuscript. **AE** conceived the design and contributed to the manuscript.


This paper (II) forms Chapter 3 of this thesis; **DP** assisted in the study design, contributed to sample collection and processing, performed all experimental work, coordinated the sequencing, performed all analyses and wrote the manuscript. **WK** provided funding and reviewed the manuscript. **NHN** provided statistical support and reviewed the manuscript. **AE** conceived the design and contributed to the manuscript.


Chapter 4 of this thesis forms the completed manuscript (III) and has been submitted to a special issue of Hydrobiologia following the Crustacean Society conference where **DP** has been accepted as a symposium speaker. **DP** conceived the study, performed all analyses and wrote the manuscript. **AE** assisted in the study design and contributed to the manuscript.

Additional published contributions relevant to the thesis but not forming part of it


This paper (IV) was supported by the banana shrimp transcriptome analysis performed in Chapter 2. **DP** assisted with sequence data analysis, development of microsatellite markers and contributed to the manuscript.

This paper (V) laid the groundwork for understanding the morphometric complexities of the captive banana shrimp population studied in this thesis and reported the differences in body weight between male and female shrimp that informed the sample selection for Chapter 4. **DP** participated in the design of the DNA markers and data collection.


The above paper (VI) is the antecedent to the work performed in Chapter 3 and formed the basis for the selection of genetic families with varying levels of HPV for experimentation with RNA-Seq. **DP** assisted in the sample collection and processing and analysis of HPV viral titre and contributed to the manuscript.


The above paper (VII) was the first published work to exploit the draft genome of the Sydney rock oyster to complement transcriptome assembly information and to identify genes not found expressed in the transcriptome libraries. **DP** performed the initial genomic draft assembly, assisted in the sequence analysis and contributed to the manuscript.
# Table of contents

Abstract ................................................................................................................................................................. i  
Declaration of originality ....................................................................................................................................... iii  
Acknowledgements ............................................................................................................................................. iv  
List of publications ................................................................................................................................................ v  
  Additional published contributions relevant to the thesis but not forming part of it ..................................... v  
Table of contents ................................................................................................................................................ vii  
List of figures ........................................................................................................................................................ x  
List of tables ......................................................................................................................................................... xi  
List of abbreviations ............................................................................................................................................xii  
Chapter 1: Review of the literature .................................................................................................................. 1  
  1.1 Introduction ................................................................................................................................................ 1  
  1.1.1 Aquaculture in Australia ...................................................................................................................... 1  
  1.1.2 Advances in animal production through the use of genomics ............................................................ 2  
  1.1.3 Availability of genomic resources for Australian aquaculture ............................................................ 6  
  1.1.4 Production of genomic information using next-generation sequencing technologies ....................... 7  
  1.2 Analysis of next-generation sequencing data ............................................................................................ 9  
  1.2.1 Eukaryotic transcriptome assembly and characterisation .................................................................. 9  
  1.2.2 Measuring differential gene expression using next-generation sequencing .................................... 11  
  1.2.3 De-novo genome sequencing and annotation .................................................................................. 12  
  1.3 Research aims and objectives ................................................................................................................... 14  
  1.4 References ................................................................................................................................................ 15  
Chapter 2: De-novo transcriptome analysis of the banana shrimp (*Fenneropenaeus merguiensis*) and identification of genes associated with reproduction and development .......................................................... 19  
  2.1 Abstract .................................................................................................................................................... 19  
  2.2 Introduction .............................................................................................................................................. 19  
  2.3 Materials and Methods ............................................................................................................................ 20  
  2.3.1 Tissue material and RNA extraction .................................................................................................. 20  
  2.3.2 Assembly, quality assessment and protein-coding region prediction .............................................. 21  
  2.3.3 Searches for sequence homologs and annotation ............................................................................ 21  
  2.4 Results ...................................................................................................................................................... 22  
  2.4.1 Transcriptome sequencing and assembly ......................................................................................... 22  
  2.4.2 Estimation of transcriptome completeness ...................................................................................... 22  
  2.4.3 Annotation of predicted protein-coding regions .............................................................................. 23  
  2.4.4 Transcript abundance across tissues ................................................................................................. 25  
  2.4.5 Taxonomy and cross species comparison ......................................................................................... 26  
  2.4.6 Screening of genes associated with reproduction, sex determination and development ............ 27  
  2.5 Discussion ................................................................................................................................................. 31  
  2.5.1 Transcriptome assembly .................................................................................................................... 31
Chapter 3: Transcriptional profiling of banana shrimp *Fenneropenaeus merguiensis* with differing levels of viral load .......................................................... 38

3.1 Synopsis .................................................................................................................... 38
3.2 Introduction ................................................................................................................ 39
3.3 Methods .................................................................................................................... 40
  3.3.1 Sample collection and RNA extraction ................................................................. 40
  3.3.2 Transcriptome assembly, ORF prediction and quality assessment ...................... 40
  3.3.3 Differential gene expression analysis .................................................................. 41
  3.3.4 Variant detection .................................................................................................. 41
3.4 Results ...................................................................................................................... 41
  3.4.1 Sample selection, RNA sequencing and transcriptome assembly ......................... 41
  3.4.2 Differential expression and GO term enrichment analysis .................................. 43
  3.4.3 Identification of sequence variants within DEGs ................................................. 44
3.5 Discussion ................................................................................................................. 47
  3.5.1 Transcriptome assembly and annotation ............................................................ 47
  3.5.2 Differential gene expression ............................................................................... 47
  3.5.3 Identification of sequence variants ...................................................................... 51
  3.5.4 Conclusions ........................................................................................................ 52
3.6 Acknowledgements .................................................................................................. 52
3.7 References ............................................................................................................... 53
S3.1 Supplementary Methods ........................................................................................ 57
  S3.1.1 Functional annotation ..................................................................................... 57
  S3.1.2 RT-qPCR .......................................................................................................... 57
S3.2 Supplementary Results .......................................................................................... 57
  S3.2.1 RT-qPCR validation ........................................................................................ 57
S3.3 Supplemental references ......................................................................................... 59

Chapter 4: Sex-specific transcript expression in the hepatopancreas of the banana shrimp (*Fenneropenaeus merguiensis*) ........................................................................ 60

4.1 Abstract .................................................................................................................... 60
4.2 Introduction ............................................................................................................... 60
4.3 Methods .................................................................................................................... 61
  4.3.1 Sample collection and RNA extraction ................................................................. 61
  4.3.2 Data processing, annotation and differential gene expression analysis ................ 61
  4.3.3 Construction of weighted gene co-expression network ...................................... 62
Chapter 5: The genome of the Sydney rock oyster (*Saccostrea glomerata*)

5.1 Abstract ........................................................................................................................................ 73
5.2 Introduction .................................................................................................................................... 73
5.3 Results .......................................................................................................................................... 74
  5.3.1 Genome sequencing and *de novo* assembly ............................................................................. 74
  5.3.2 Genome annotation and comparative analysis ........................................................................... 75
  5.3.3 Expansion of gene families ........................................................................................................ 76
5.4 Discussion ...................................................................................................................................... 79
5.5 Methods ........................................................................................................................................ 81
  5.5.1 Generation of sequence data .................................................................................................... 81
  5.5.2 Sequence assembly .................................................................................................................... 81
  5.5.3 Genome annotation .................................................................................................................... 81
  5.5.4 Gene family analysis .................................................................................................................. 82
  5.5.5 Quantifying gene expression ...................................................................................................... 82
5.6 References ...................................................................................................................................... 83
S5.1 Supplementary Figures .............................................................................................................. 86
S5.2 Supplementary Tables ................................................................................................................ 90
S5.3 Supplementary Note 1. *S. glomerata* sequencing and *de novo* assembly ......................... 95
  S5.3.1 Oyster sample and sequencing ................................................................................................. 95
  S5.3.2 Initial diploid genome assembly ............................................................................................... 95
  S5.3.3 Construction of primary haploid assembly .............................................................................. 96
  S5.3.4 Chicago library preparation and sequencing ........................................................................... 96
  S5.3.5 Scaffolding input assemblies with HiRise ................................................................................ 96
  S5.3.6 Gap filling and redundancy removal ....................................................................................... 97
S5.4 Supplementary Note 2. Gene model prediction and analysis .............................................. 97
  S5.4.1 Genome annotation ................................................................................................................ 97
  S5.4.2 Pfam enrichment analysis ....................................................................................................... 98
S5.5 Supplementary References ....................................................................................................... 99

Chapter 6: Conclusions .................................................................................................................... 100
6.1 Future directions .......................................................................................................................... 102
6.2 References ..................................................................................................................................... 103

---

4.4 Results ......................................................................................................................................... 62
  4.4.1 Sample selection and RNA sequencing ..................................................................................... 62
  4.4.2 Sex-specific transcript expression ............................................................................................... 63
  4.4.3 Weighted gene co-expression network analysis .......................................................................... 64
4.5 Discussion ..................................................................................................................................... 67
  4.5.1 Conclusions ............................................................................................................................... 68
4.6 Acknowledgements ..................................................................................................................... 69
4.7 References ................................................................................................................................... 70

---

S5.4.1 Genotype annotation .................................................................................................................. 97
S5.4.2 Pfam enrichment analysis ........................................................................................................... 98
S. glomerata gene model prediction and analysis .................................................................................. 97

---

ix
List of figures

Figure 1.1 Total Australian aquaculture production by year since 1990. ............................................................ 1
Figure 2.1 BLAST annotation results.................................................................................................................. 24
Figure 2.2 Distribution of second-level functional annotation based on the gene ontology database............. 25
Figure 2.3 Distribution of predicted protein sequences annotated using the KOG database.......................... 25
Figure 3.1. Workflow diagram ........................................................................................................................... 42
Figure 3.2. Hierarchical clustering.................................................................................................................... 44
Figure 3.3. Fold change and annotation results of additional DEGs implicated in immune related functions and also related to growth and reproduction................................................................................................. 46
Figure 3.4. SNP analysis..................................................................................................................................... 47
Figure 3.5. Putative immune response pathway................................................................................................. 51
Supplementary Figure 3.1. Comparison of the estimated RNA-Seq fold change values with RT-qPCR. ......... 58
Supplementary Figure 3.2. Mean group Log HPV copy number per microgram of DNA............................... 58
Figure 4.1. Heatmap of differentially expressed genes....................................................................................... 63
Figure 4.2. Fold change of differentially expressed transcripts........................................................................ 64
Figure 4.3. Identification of gene co-expression models among the female banana shrimp hepatopancreas samples....................................................................................................................................................... 65
Figure 4.4. Gene ontology analysis.................................................................................................................... 66
Figure 4.5. Gene network map.......................................................................................................................... 66
Figure 5.1. Shared and unique gene families in four species of mollusc........................................................... 76
Figure 5.2. A selection of enriched Pfam domains in S. glomerata................................................................. 77
Figure 5.3. Expansion and expression of S. glomerata fibrinogen_C domain-containing genes...................... 78
Figure 5.4. Inhibitor of apoptosis (IAP) gene expansion in S. glomerata............................................................. 79
Supplementary Figure 5.1. Workflow diagram of the de novo genome assembly and gene model prediction pipeline.................................................................................................................................................. 86
Supplementary Figure 5.2. Insert size distribution for fragment libraries........................................................ 87
Supplementary Figure 5.3. k-mer frequency distribution analysis ................................................................. 87
Supplementary Figure 5.4. k-mer spectra plot of the assembly vs clean short-insert read data using the KAT tool .................................................................................................................................................. 88
Supplementary Figure 5.5. Cumulative annotation edit distance (AED) scores for gene models produced using the MAKER2 pipeline.............................................................................................................. 89
List of tables

Table 2.1 Summary of Roche 454FLX Titanium and illumina HiSeq 2000 sequencing ........................................... 22
Table 2.2 Statistics of hybrid de-novo assembly of the banana shrimp (F. merguiensis) transcriptome and overview of annotation results. .................................................................................................................. 23
Table 2.3 Comparative sequence analysis ............................................................................................................. 26
Table 2.4 Number of hits of selected GO terms based on GO:0032504: multicellular organism reproduction. ........................................................................................................................................................................ 27
Table 2.5 Alignment results of selected homologous proteins found in the F. merguiensis transcriptome associated with the regulation of reproduction ........................................................................................................ 29
Table 2.6 Alignment results of selected homologous proteins found in the F. merguiensis transcriptome known to be involved with sex determination and development ......................................................... 30
Table 3.1. Transcriptome assembly metrics and results from the annotation pipeline. ........................................... 43
Table 3.2. Differentially expressed genes associated with immune response identified when comparing banana shrimp with differing HPV levels. ........................................................................................................ 45
Table 3.3. GO term enrichment for the subset of 404 DEGs against all sequences with GO annotations ............ 46
Supplementary Table 3.1. Details of hepatopancreas tissue samples including HPV concentration, sequencing yield after quality trimming and percentage of clean reads mapping to the unique gene groups. .......... 59
Table 5.1. Features of the draft S. glomerata genome .......................................................................................... 75
Supplementary Table 5.1. Illumina genomic libraries and predicted coverage based on an estimated 784 Mb genome ........................................................................................................................................ 90
Supplementary Table 5.2. Summary of the clean read data used for assembly and scaffolding. ....................... 90
Supplementary Table 5.3. Length distribution and contig composition of the final scaffold output of the draft S. glomerata assembly. ........................................................................................................... 91
Supplementary Table 5.4. Overview of the repetitive content of the S. glomerata genome. .......................... 92
Supplementary Table 5.5. Estimates of genome assembly completeness using the CEGMA and BUSCO tools and read mapping statistics ........................................................................................................ 93
Supplementary Table 5.6. S. glomerata gene model annotation ........................................................................ 94
Supplementary Table 5.7. Completeness estimates of Saccostrea gene models and comparison with other published molluscan genomes. ......................................................................................... 94
List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>aa</td>
<td>amino acids</td>
</tr>
<tr>
<td>ALF</td>
<td>anti-lipopolysaccharide factor</td>
</tr>
<tr>
<td>AMPs</td>
<td>antimicrobial peptides</td>
</tr>
<tr>
<td>bp</td>
<td>base pairs</td>
</tr>
<tr>
<td>CDS</td>
<td>coding sequence</td>
</tr>
<tr>
<td>CEGMA</td>
<td>Core Eukaryotic Genes Mapping Approach</td>
</tr>
<tr>
<td>CHH</td>
<td>crustacean hyperglycemic hormone</td>
</tr>
<tr>
<td>CTLs</td>
<td>C-type lectins</td>
</tr>
<tr>
<td>DEGs</td>
<td>differentially expressed genes</td>
</tr>
<tr>
<td>eQTL</td>
<td>expressed quantitative trait loci</td>
</tr>
<tr>
<td>EST</td>
<td>expressed sequence tag</td>
</tr>
<tr>
<td>FAO</td>
<td>United Nations Food and Agriculture Organization</td>
</tr>
<tr>
<td>FDR</td>
<td>false discovery rate</td>
</tr>
<tr>
<td>FPKM</td>
<td>fragments per kilobase of transcript per million mapped reads</td>
</tr>
<tr>
<td>FREPs</td>
<td>fibrinogen-related proteins</td>
</tr>
<tr>
<td>GIH/VIH</td>
<td>vitellogenin-inhibiting hormone</td>
</tr>
<tr>
<td>GnRH</td>
<td>gonadotropin-releasing hormone</td>
</tr>
<tr>
<td>GO</td>
<td>gene ontology</td>
</tr>
<tr>
<td>HPV</td>
<td>hepatopancreatic parvo-like virus</td>
</tr>
<tr>
<td>IAPs</td>
<td>Inhibitor of apoptosis proteins</td>
</tr>
<tr>
<td>JH</td>
<td>juvenile hormone</td>
</tr>
<tr>
<td>Kb</td>
<td>kilobase</td>
</tr>
<tr>
<td>KEGG</td>
<td>Kyoto Encyclopedia of Genes and Genomes</td>
</tr>
<tr>
<td>KOG</td>
<td>eukaryotic orthologous groups of proteins</td>
</tr>
<tr>
<td>LH</td>
<td>luteinising hormone</td>
</tr>
<tr>
<td>Mb</td>
<td>megabases</td>
</tr>
</tbody>
</table>
MF  methyl farnesoate
MIH  molt-inhibiting hormone
NCBI  National Center for Biotechnology Information
NGS  next-generation sequencing.
nt  nucleotides
ORF  open reading frame
PCR  polymerase chain reaction
PE  paired-end
proPO  prophenoloxidase
QTL  quantitative trait loci
RFLP  restriction fragment length polymorphisms
RPKM  reads per kilobase of transcript per million mapped reads
RT-qPCR  real-time quantitative PCR
RXR  retinoid X receptor
SNP  single nucleotide polymorphism
TLRs  toll-like receptors
TPM  transcripts per one million mapped reads
TSV  Taura Syndrome Virus
UTR  untranslated region
WGCNA  weighted gene co-expression network analysis
WSSV  white spot syndrome virus
Chapter 1: Review of the literature

1.1 Introduction

1.1.1 Aquaculture in Australia

Aquaculture in Australia is small in comparison to the rest of the world, though it is rapidly expanding. Over the last few years, Australia’s aquaculture sector has grown considerably and is currently among the fastest growing primary industries in Australia contributing 34 per cent of the total gross value of food fish production (ABARES 2013). Figure 1.1 depicts the dramatic 6-fold increase in production since 1990 and highlights the sharper growth trend in more recent years. The total production of aquaculture in 2010, 2011 and 2012 reached 71,667, 71,583 and 80,004 tonnes respectively (FAO 2012; ABARES 2013). From the most recent report at the time of writing, total production of aquaculture for 2015 was 89,217 tonnes (ABARES 2015).

The gross value of aquaculture production for the 2011-2012 reporting year increased by around ten percent on the year previous to AUD$1.1 billion, accounting for 46 per cent of the gross value of Australian fisheries production. The volume of aquaculture production increased by 10 per cent to 84,605 tonnes, accounting for 36 per cent of total Australian fisheries production. The top five species groups; shrimp, tuna, salmonids, edible oysters and pearl oysters accounted for 88 per cent of this value, contributing $931 million during this same reporting year (ABARES 2013). The salmonids became Australia’s most valuable species group the year earlier (2010–2011). Salmonid production in aquaculture that year was valued at $512 million, representing almost 22 per cent of the total gross value of fisheries production and yielding the greatest production volume at 43,989 tonnes. This was followed by edible and pearl oysters ($209 million), tuna ($150 million) and shrimp ($59 million). For the 2014-2015 reporting year, the gross value of aquaculture production reached $1.2 billion (ABARES 2015).

![Australian Aquaculture Production](image)

**Figure 1.1 Total Australian aquaculture production by year since 1990.** Data retrieved from the FAO - Fisheries and Aquaculture Information and Statistics Service on the 21st April 2017. (Includes all species groups).
There is increasing demand around the world for Australian-produced aquaculture, including its native species (ABARES 2015). Together, crustacean and mollusc exports accounted for 46 per cent (18,436 tonnes) of total edible export volume and 71 per cent ($711 million) of edible export value in 2011–12 (ABARES 2013). The majority of the volume exported was destined for Asia with the heaviest consumers being Hong Kong ($479 million), Japan ($255 million), Vietnam ($60 million), mainland China ($59 million) and Singapore ($42 million), which combined accounted for 89 per cent of the total value of Australian seafood exports in 2011–12.

The Australian shrimp farming industry produces more than 5,200 tonnes of shrimp a year worth $86 million and making up over 20 per cent of the total Australian shrimp production in 2014–15 (ABARES 2015). The industry is primarily concerned with the farming of three endemic species: the black tiger shrimp (Peneaus monodon), the banana shrimp (Fenneropenaeus merguiensis) and the kuruma shrimp (Peneaus japonicus) (DPI&F 2006; ABARES 2013). Considerable investment from industry and the government into research and development in areas such as genetics, grow-out technology, environmental and health management has provided a steady stimulus for the industry to expand (DPI&F 2006).

Oysters have played a major role in the establishment of Australian aquaculture and continue to be one of the most significant contributors. One of the earliest commercial aquaculture products was the iconic Sydney rock oyster grown in New South Wales from 1872 (ABARES 2013). Edible oysters (15,745 tonnes) were the second most plentiful product produced during 2011–12 and, together with pearl oysters, contributed almost 20 per cent ($209,681) of the total value of Australian aquaculture production (ABS 2012; ABARES 2013). Australia farms three species of edible oyster, the main two being the Pacific oyster (Crassostrea gigas) and the Sydney rock oyster (Saccostrea glomerata) with the Angasi (flat) oyster (Ostrea angasi) grown in smaller volumes across all states. The largest component of the pearling industry in Australia focuses on the silver-lipped pearl oyster (Pinctada maxima), which produces high quality pearl and mother of pearl shell. Other species farmed on a much smaller scale include the shark bay pearl oyster (Pinctada albina) and the akoya pearl oyster (Pinctada imbricata) (ABARES 2013).

1.1.2 Advances in animal production through the use of genomics

Innovations in biotechnology and molecular genetics have had a major impact on aquaculture production. Technologies such as sex-reversal, polyploidy and DNA marker-assisted breeding, developed over the past few decades, have been instrumental in facilitating performance improvements in traits such as growth rates and flesh quality, to name a few (Dunham 2011). Driven by progressive DNA marker and sequencing technologies, research in aquaculture genomics has made remarkable advances in recent years and has already begun to provide a wealth of resources supporting the determination of important economic traits of aquatic organisms.

Genomics is a term coined in the life sciences to describe the scientific study of structure, function and interrelationships of both individual genes and the genome in its entirety (Fadiel et al. 2005). More recent definitions use genomics to describe the study of all the genes and gene products of an organism, including interactions of these genes with each other and with the external environment (Joyce and Palsson 2006). The
genomic era began upon the completion of the Human Genome Project in the year 2001 and has since given rise to genome projects for many other species (Guarnaccia et al. 2014). Comprehensive genomic resources make available to researchers the genetic instructions for the entire repertoire of cellular components and functions of an organism, knowledge that is essential in order to begin unravelling the intricacies of complex biological systems (Collins et al. 2003).

Whilst a small number of the major species produced in aquaculture are beginning to accumulate detailed molecular resources, they are far behind those developed by the livestock industries, in particular, that of beef and dairy cattle. The draft bovine genome was made available in 2005 and just a few years later, the genomic evaluation of dairy cattle became available in the United States. The commercially available BovineSNP50 BeadChip, released in 2007, could measure up to 54,000 single nucleotide variations within economically important farmed cow populations, giving farmers unprecedented control over their breeding programs (Womack 2005; Tassell et al. 2008). The Bovine3K BeadChip, a refined, more economical version of this technology introduced in 2010, was rapidly adopted by the industry. The number of new genotypes being tested each month upon the release of this system quickly reached 5,000 and was continuing to increase (Wiggans et al. 2012). These genomic selection tools can offer information to cattle breeders relating to phenotypic variation in traits which have significant commercial importance, such as birth weight, average daily gain, steak tenderness and marbling scores. The development of this technology has set a benchmark in the level of utility molecular tools can provide towards the improvement of farmed animals. Not surprisingly, the application of this technology for use in different animals has captured the interest of stakeholders in other industries.

Genomic selection is a broader form of marker-assisted selection in which genetic markers covering the whole genome, such as the above mentioned single nucleotide polymorphisms (SNPs), are used (Goddard and Hayes 2007). The advantage of genomic selection over traditional selection methods that are based on pedigree and phenotype data alone is that animals can be selected based on genomic predictions before reaching an age of phenotypic expression. Animals can also be selected for traits that are difficult or expensive to measure, such as disease resistance or milk production, leading to more focused and rapid improvements each generation (Green et al. 2007; Hayes et al. 2013).

Methods to control the level of inbreeding in selected animals have also been developed using a whole-genome approach. These studies of genetic diversity require a well-constructed reference genome in order to catalogue genetic variances present throughout lineages (VanRaden et al. 2011). Studies by Villa-Angulo et al. have estimated that at least 50,000 markers are needed for estimating within-breed effects in cattle, however, the authors postulate 300,000 to 600,000 markers may be required in order to predict effects across breeds (Villa-Angulo et al. 2009). The development of a system to measure markers at this scale can be made possible with access to comprehensive and accurate genome drafts.
Genetic linkage maps are tools devised for investigations of genome organisation and provide a general overview of a genome based on the frequency of recombination between molecular markers. The main application of these genetic maps is to provide a platform for the identification of quantitative trait loci (QTL) which are genomic regions that are associated with a particular phenotypic trait (Cerdà et al. 2010). QTL can be highly informative in breeding programs when linked to traits of commercial interest. Approaches for the development of markers to create genetic linkage maps have traditionally involved restriction fragment length polymorphisms (RFLP) or microsatellites which require a substantial investment in time and resources due to the low success rate of discovering high-resolution polymorphic markers and the added difficulties of identifying candidate genes (Lim et al. 2014). Sequence based genotyping using next-generation technologies can overcome these limitations through the discovery of SNP markers from the comparison of different genomes present in a study population. It is not uncommon to identify several hundred thousand SNPs when comparing entire genomes of individuals which can be done in a single multiplexed experiment, thereby reducing the time and effort required for typical polymorphic marker discovery and significantly increasing the number of candidate markers to drastically improve the accuracy and resolution for selection. RFLP and microsatellite markers generally offer low density coverage of the genome leaving many regions largely unrepresented, making it difficult to obtain accurate and complete information about the numbers and locations of QTL. Increasing marker density can further improve the resolution of genetic maps, thus enhancing the precision of QTL mapping. A study by Yu et al. reported a significant improvement in detection power and resolution of QTL mapping by using an ultra-high density SNP based map (Yu et al. 2011). Houston et al. have made some progress towards the scale of genotype selection that is now common in the cattle industry in aquaculture. This group has reported the use of a method of sequencing fragments of restriction site-associated DNA using high-throughput technology, to identify SNPs linked with QTL in Atlantic salmon. Building upon knowledge gained through studies using traditional marker technology, researchers were able to expand greatly the density of markers across the genome and identify SNPs linked to QTL associated with disease resistance (Houston et al. 2012).

Characterising expressed gene sequences by analysing RNA transcripts can offer insight into the active genetic content of a cell or tissue and provides a functional basis to study the genome. Studies of expressed sequence tags (ESTs), which represent fragments of coding DNA, are useful for describing what genes are expressed in a certain tissue, stage of development or physiological state and can provide information relating to functional characterisation (Fadiel et al. 2005). Gene sequences identified in this way may then be used to develop molecular markers, create targeted sequencing assays, or to inform experiments profiling the expression of these and other closely related genes (Ekblom and Galindo 2011). Studying only the transcribed genes or transcriptome in the absence of a completed genome can provide a rapid method for gene discovery, SNP detection and elucidation of QTL (Le Mignon et al. 2009). Production of gene transcript sequences using high-throughput technologies gives the investigator the possibility, not only to uncover expressed genes of interest, but entire biological pathways associated with complex traits. A study by He & Liu (He and Liu 2013) used a
next-generation sequencing based approach to generate approximately 54,000 expressed transcripts representing novel genes from the longissimus muscle in Qinchuan beef cattle at embryonic and adult stages. This study revealed a large number of genes of both novel and of known function which provided a profile of the developmental processes of this species. These data contributed greatly to the amount of genetic information available for these animals which have benefits for future functional studies on the beef cattle genome and provide resources that have been applied to breeding programs for cattle and closely related mammals (He and Liu 2013). The same approach has also been used successfully for the discovery of genes involved in pathways of interest in the farming of Atlantic cod (Gadus morhua), such as immune response, thermal adaptation and reproduction. This knowledge is important for supporting the developing aquaculture of this species by enabling the application of molecular markers for these pathways within selective breeding programs (Bowman et al. 2011).

Knowledge of the level of expression of RNA transcripts can provide powerful insights into the mechanisms of regulation within the cell. Gene regulators commonly respond to cues from environmental influences such as stress caused by increased temperature or changes in metabolic activity due to infection or starvation. Genes that have similar signatures of regulation in response to a particular extracellular stimulus often have some relationship of function within the cell, tissue or organs. This relationship can be examined to uncover clusters of genes that are responsible for the genetic causes behind certain phenotypic characteristics in animals (Filkov 2005). Determining levels of gene expression between animals exhibiting different phenotypes important for production can lead to the identification of key genes affecting commercial traits. In a recent study by Cui et al., the use of next-generation sequencing technology was employed to examine the complete gene expression profile of mammary gland tissue between two groups of Holstein dairy cattle with extremely high and low milk protein and fat percentage. The investigators detected 31 differentially expressed genes associated with protein metabolism, fat metabolism, and mammary gland development and, together with QTL information, revealed 8 promising candidate genes for milk composition traits (Cui et al. 2014). Cost reductions offered by the widespread adoption of these new sequencing technologies for differential gene expression experiments has enabled its use to become more available to the broader scientific community with studies involving aquaculture species beginning to make use of these methods. Li et al. have reported the identification of changes in the expression of genes related to mucosal immunity in gill and skin tissue in response to starvation in the blue catfish (Ictalurus furcatus). These animals are crossed with the farmed channel catfish (Ictalurus punctatus) to produce faster growing hybrids that are being increasingly used in aquaculture in the United States. The authors observed that fasting altered a total of 530 genes in the surface mucosa which included genes involved in regulating the immune response, energy metabolism, mucus production, cell proliferation, and antioxidant responses. With short term feed deprivation sometimes used as a response to overproduction or as disease control measures, this research highlights the critical interaction between nutrition and immunity (Li et al. 2014). Recent work by van der Merwe and colleagues explored the variability of genes involved the growth rate of the farmed abalone (Haliotis midae). In this study, differential gene expression between two
sibling groups of abalone exhibiting significant growth variation was determined using next-generation sequencing technology. Results obtained from this analysis revealed significant differences in gene expression between large and small abalone. Of these differences, the up-regulation of genes involved in the insulin signalling pathway suggests that insulin-like proteins may be involved in enhanced growth rate for various tissues in this species. Genes involved in this pathway can be further explored as targets for markers of increased growth and for use in selection of future breeding groups in steps to enhance the prominent abalone aquaculture industry in South Africa and potentially to other mollusc breeding programs (van der Merwe et al. 2011).

Clearly, the advancement of genomics and its applications in the farming of animals has already begun to introduce remarkable gains in the knowledge of these animals. The research previously described has brought forth the development of invaluable resources that will provide a basis for biological research well into the future. Aquaculture research has only very recently started to employ this technology, yet significant gains have already been realised. Extending this knowledge is of great importance for furthering the productivity and improvement of the industry and for ensuring a high standard of animal health and wellbeing.

1.1.3 Availability of genomic resources for Australian aquaculture
The importance of aquaculture to the world’s food supply has provided the impetus for much research into the biology of many commercial species. Not surprisingly, the species that comprise the largest portion of the aquaculture industry are some of the best understood. Multi-billion dollar industries, like that of the Atlantic salmon and tilapias, are clearly better positioned to invest the large sums of money required to produce high quality information and research. Traditionally, ventures into biotechnology and other cutting-edge fields of science are expensive in time and capital. Thus, many of the smaller industries have been unable to take advantage of the technologies the larger industries have prospered from. This has led to an underrepresentation of these lesser-produced species in the available research data. Further, not all of the resources are made publically available with some consortiums opting not to upload their data to the public domain in order to maintain a competitive edge and commercialise the knowledge.

Over the past several years, whole genome sequencing has featured prominently in the scientific literature. Many non-model organisms, some of which are used in aquaculture, have now had a draft genome completed with thousands more currently in progress or planned, including those selected as part of the Genome 10K project that proposes to sequence 10,000 vertebrate genomes covering approximately one for each vertebrate genus (Genome 10K Community of Scientists 2009; Zhang et al. 2012). It is clear that the scientific community understands the value of genomic resources in biological research and efforts to obtain this information have been placed at a high priority. The question now posed is which species will gain favour for targeting sooner and which species will be left to wait.

The resources available to date for the main five species groups of Australian aquaculture; salmonids, tuna, oysters (edible and pearl) and shrimp, vary in their depth and quality. Databases such as GenBank maintained
by the NCBI and UniProt by the EMBL are the most extensive source of nucleotide and protein sequences. Types of genetic resources include traditional EST sequences, genome libraries, microarrays and genetic maps. More recently, transcriptome and whole genome sequence data have been deposited in these online libraries for access to the public.

The genomic resources available for the Atlantic salmon are amongst the most extensive of all aquaculture species. These include several genetic maps, an extensive EST database including approximately 500,000 tags and several microarrays from many different tissues and developmental stages. The first draft assembly of the Atlantic salmon genome is currently available to the public (Davidson et al. 2010; Gonen et al. 2014).

A draft tuna genome assembly has recently become available for the Pacific bluefin tuna (*Thunnus orientalis*), (Nakamura et al. 2013). However, Australian aquaculture is primarily concerned with the Southern bluefin tuna (*Thunnus maccoyii*), of which only 142 nucleotide and 130 protein sequences are available in the NCBI databases. As these species are relatively closely related, this genome assembly has been used as an informative reference for the Southern bluefin (Bar et al. 2016).

Being an important production species, not only in Australia but throughout many parts of the world, studies of the Pacific oyster (*C. gigas*) are well advanced. Recently, Zhang et al. reported the sequencing and assembly of the genome of the Pacific oyster (Zhang et al. 2012). In contrast, the Sydney rock oyster (*S. glomerata*) had only 168 nucleotide, 183 EST and 6 protein sequence deposited in the NCBI databases at the commencement of this study. Unlike the bluefin tuna, the divergence of these two species is beyond genus level, leaving some doubt to the usefulness of using the Pacific oyster genome as an effective reference for the Sydney rock oyster. Multiple studies have characterised nucleotide and EST sequences from *P. monodon* and to a lesser extent, *P. japonicus* (Tassanakajon et al. 2010). Despite their worldwide importance to aquaculture, no whole-genome sequences have been made publically available for any marine shrimp. *P. monodon* has been the most extensively studied having over 200,000 nucleotide and EST sequences along with 4 transcriptomic datasets deposited in the NCBI databases yet only a meagre 183 nucleotide and 101 protein sequences were available for *F. merguiensis* at the commencement of this study. Given the amount of sequence information available for aquaculture species is relatively lower than most other domesticated animals, a considerable knowledge gap is especially evident for *F. merguiensis* and *S. glomerata*.

1.1.4 Production of genomic information using next-generation sequencing technologies

The principal behind the generation of genomic sequence data using current technologies is to take the enormous amount of genetic material found in cells and tissues and fragment it into small, 350 – 1,000 bp DNA fragments, which can be read by instruments able to determine the order of nucleotide bases. These small DNA sequences are then pieced back together to form the intact, full-length sequences originating from the cells. Since the early 1990’s, sequencing technology based on the Sanger biochemistry was used almost exclusively for the production of DNA sequences (Shendure and Ji 2008). Adapting this technology using capillary arrays, this method could yield over nine hundred bases of sequence with low error rates and, running
ninety six samples at once, taking around two hours. As Sanger based sequencing relies upon the amplification of the region to be sequenced prior to reading, genetic material was often subjected to either of two different approaches in order to generate adequate quantities. If the genetic material is unknown, DNA would be randomly fragmented and inserted into a high-copy-number plasmid which is then transformed into a bacterial host. If there is some information available for the sequence, PCR amplification with primers that flank the target region are used (Shendure and Ji 2008). This method demands a significant investment in labour and materials leaving its use for high-throughput projects beyond the means and time-span of most grant-funded research (Morozova and Marra 2008).

Recently, the Sanger method has been in many ways surpassed by several next-generation sequencing (NGS) technologies. The success of the human genome project spurred the advent of high-throughput sequencing technologies that could more efficiently handle the amount of sequencing required to complete a genome (Morozova and Marra 2008). In 2005, 454 Life Sciences Corp. reported the simultaneous acquisition of hundreds of thousands of sequence reads per run with equal quality metrics to the Sanger sequencing, albeit with shorter read lengths of 80-120 bases, this system could generate 25 million bases per four-hour run (Margulies et al. 2005). NGS technologies eliminate the need for the bacterial cloning steps necessary for traditional Sanger sequencing and instead amplify fragmented, singly isolated DNA molecules and analyse them in a massively parallel way, making them much more cost efficient (Bubnoff 2008). Since this time, the race has been on to produce cheaper and faster DNA sequencing technologies capable of handling an ever greater number of genomes and in the short time since this technology become commercially available, the cost of sequencing a human genome has fallen from a predicted US$10 Million to the highly anticipated target of US$1,000 (Wetterstrand 2014). This dramatic drop in cost has brought whole genome sequencing within reach of any investigator. At the time of writing, current NGS platforms have the capability to produce reads over 60 Kb in length and 20 billion reads per run totalling up to 6 trillion bases (www.pacb.com, www.illumina.com).

Each of the next-generation sequencing platforms common to the current market use a different approach to generating sequence reads. These differences impose constraints on the data obtained from the run in terms of read length, number of reads and total volume of bases, and also with respect to sequencing accuracy and the production of artefacts in the data (McPherson 2009). There are also differences in the size and additions applied to the nucleic acids during fragmentation and preparation of the samples (Bubnoff 2008). For this reason, selecting which platform to use will depend on the investigators sample type and research question. With technologies being rapidly updated, improved and newly invented, this choice is not always made obvious. Some important considerations may include; the amount of sequence required for the experiment, whether to opt for single or paired-end reads, if paired-end then deciding on insert size, knowing the degree of accuracy required, and to determine if your sample is suitable for the pre-sequencing preparation steps used by the platform (Shendure and Ji 2008). For many researchers, the current platform of choice for most genomics applications is the illumina HiSeq (Quail et al. 2012).
Making genome and transcriptome sequencing more accessible has created other challenges concerning the analysis of such a wealth of data. The speed with which the development of this technology has progressed and the growing capacity of its output has outpaced the ability for researchers to process and analyse the resulting information (McPherson 2009). Computational methods for sequence analysis have generally been developed by bioinformaticians to answer specific research questions. Most of the software developed in this way has been made open-source and is available free of charge by the authors. Slowly, commercially produced software has entered the market and is offered at considerable cost. Whilst this software is extremely user-friendly and can provide time savings, it is usually aimed at research in human or model organisms and often performs at a substandard to the open-source equivalent. Skills in computer programming are often essential for the successful interpretation of next-generation sequence data. As the output of the sequencing platforms increases, so does the requirement for increased computing resources. Access to high-performance computing systems has become crucial for handling the massive data processing, storage and access requirements of these bioinformatics software (Berger et al. 2013).

Sequencing technology has evolved significantly over the relatively short duration of this doctoral study, and along with it, methods and approaches for data analysis. During the undertaking of this degree, new platforms and techniques have emerged that offer types of read data that at the beginning of this study were considered only theoretical, whilst others have been discontinued or made redundant. The first research chapter of this thesis reports analysis of data produced on the 454 FLX platform that has since been retired, whilst platforms such as the PacBio RS II and the Oxford Nanopore ION long-read sequencers have become available and considered to be complementary or viable alternatives (Sakai et al. 2015). Such a dynamic suite of options offers a number of challenges and opportunities. As software tools are frequently developed and improved, some forms of sequence data may become outdated or unsupported yet new tools and pipelines may dramatically improve an assembly or downstream analysis. The release of a new sequencing chemistry may also help to devise a useful sequencing strategy for the more notoriously difficult species.

1.2 Analysis of next-generation sequencing data

1.2.1 Eukaryotic transcriptome assembly and characterisation

The transcriptome is the entire complement of RNA molecules transcribed by an organism or cell type representing the part of the genetic code that is in use at any given point in time (McGettigan 2013). As not every gene is active in every cell (possibly as little as 5 per cent), to understand what component of the genome is contributing to the function of the particular cell or tissue of interest, it is necessary to reveal its transcriptome (Frith et al. 2005). Differences in active genes between cells or tissues underpin the many variations they might exhibit. In eukaryotes, understanding the transcriptome of a tissue can help the investigator elucidate the function and relationship the tissue has within the organism.
A comprehensive study of a transcriptome demands the identification of the full set of transcript sequences, including isoforms and rare transcripts, present in the tissue. This has been difficult to accomplish given the traditionally low volume of capacity of Sanger-based sequencing instruments, leaving our knowledge of the transcriptome confined to a subset of EST libraries which normally comprise only the more abundant transcripts. Recently, whole transcriptome sequencing using next-generation sequencing technologies has offered an almost complete view of the transcriptome with unprecedented accuracy. The high-throughput nature of these sequencing platforms deliver an immense sequencing depth capable of detecting even the more rarely expressed transcripts (Martin and Wang 2011).

There are a number of informatics based methods that can be employed to assemble transcript sequences for reconstructing the transcriptome. Deciding upon which method to use depends upon the type of sequencing platform used, the amount and format of data generated and whether the transcriptome is to be assembled against a reference or de-novo. A transcriptome assembly is considered de-novo if it has no previously described reference genome available to use as a guide for reconstruction (Martin and Wang 2011). A reference-based assembly has the advantage of error correction against a reference genome which can help to control for accuracy, thus, the de-novo assembly approach is often prone to difficulties in resolving transcripts with high similarity or those formed from chimeric artefacts. De-novo assemblies of eukaryotes can be more challenging than prokaryotes due to the difficulties involved in identifying transcribed variations or isoforms of genes produced by alternative splicing. This is further complicated by differing transcript lengths, intronic regions, repeats and varying levels of gene expression that result in an uneven distribution of read abundance (Mundry et al. 2012). A de-novo assembly can also be computationally expensive due to the vast amount of data required which can reach over several hundred gigabases (Gb) and comprise billions of reads. Often large scale computer clusters or purpose built servers are required to analyse data of this scale within a reasonable timeframe. These factors can pose considerable challenges for accurate assembly. It is not uncommon for multiple attempts at assembly to be undertaken in order to optimise the software parameters or to adjust trimming and quality pre-processing of the raw data.

Functional annotation is the next step in the analysis of transcriptomic data post-assembly. Annotation provides an overview of the content of a transcriptome in a format that is easily viewed and searchable. The most commonly cited and perhaps most extensive functional annotation database for gene and protein sequences is the Gene Ontology (GO) (Götz et al. 2008). Manual annotation of the assembled gene sequences offers the most accurate means of gene function assignment. However, manually annotating the tens of thousands of sequences derived from a complete transcriptome is clearly impractical. Datasets of this size require an alternative with some degree of automation in order to achieve a reasonable degree of biological interpretability within an acceptable timeframe (Götz et al. 2008; Dutkowski et al. 2013). GO based annotation is built on the understanding that a high level of functional conservation is observed for core biological processes when comparing orthologous protein sequences of vastly divergent species (Gabaldón and Koonin 2013). This realisation has provided an opportunity for automatic transfer of experimentally confirmed
information from intensively researched model organisms to others based on computationally inferred sequence similarity (Rubin 2000). Although such a semi-automated approach to annotating large sequence datasets offers an attractive alternative, it does have some inherent limitations. Some important concerns are errors in submissions or incomplete records included amongst faithful entries in the databases that can lead to falsely assigned matches, and sequences that are misassembled due to read errors, failing of the software, or that are chimeric, which can cause erroneous inference of function. Thus, some small level of error may be unavoidable, yet such events can be mitigated by a number of measures including adjustments to the stringency of the matching criteria and by investigating the quality of the final assembly.

1.2.2 Measuring differential gene expression using next-generation sequencing

It is understood that differences in gene expression have the potential to exert considerable influence upon the phenotypic diversity within species and even among a population (Skelly et al. 2011). Studies in humans have shown that a considerable amount of this variation in gene expression is heritable with many of these differing expression patterns found to be specific to an allele, and are often associated with some form of sequence variation such as SNPs (Pickrell et al. 2010). Polymorphisms found to affect the expression level of a gene are also most often found near the gene itself, and some are even contained within the gene transcript (Pickrell et al. 2010; McGettigan 2013). Studies estimate that up to 80 per cent of the genes found to have an associated polymorphism within their transcribed region exhibit some degree of allele-specific expression (Skelly et al. 2011). Analysis of gene expression levels between individuals can lead to the identification of genes that vary due to specific environmental conditions. Expressed SNPs found in these genes have the potential to be developed as markers to select for a favourable response to these conditions (McGettigan 2013).

With the recent availability of genome-wide gene expression technologies using next-generation sequencing (RNA-Seq), the full abundance of transcripts present in the cell or tissue can be measured simultaneously. Investigators have transitioned to this technology as it provides advantages over microarray technologies with respect to identification of unknown and rare transcripts, and will resist becoming saturated with those that are more highly expressed (Wang et al. 2009). RNA-Seq enables analysis of the complexity of whole transcriptomes with less bias, greater dynamic range, lower frequency of false-positive signals, and higher reproducibility than that of typical microarray experiments (Cui et al. 2014).

Analysis of gene expression using RNA-Seq will typically involve transcriptome reconstruction followed by read mapping in order to estimate expression levels. Once this information has been generated, differential expression analysis can be measured from the fold change in the number of reads mapping back to the assembled transcripts. Recent studies by Hart et al. on a human cell line are in agreement with other reports that 20–30 million reads are sufficient to detect virtually all active transcripts, and to provide deep enough coverage to undertake analysis of differential expression across the bulk of the transcriptome (Wang et al. 2009; Hart et al. 2013).
There are a number of sources of variability encountered when producing quantitative RNA-Seq data. Some small technical effects are observed due to differences in positions on the sequencing platform and also within the library preparation process. These have been shown not to impact substantially upon the identification of differentially expressed genes if an appropriate normalisation procedure is used (Bullard et al. 2010). Two important sources of these effects is the RNA fragmentation during library construction having a bias for longer transcripts to generate more reads compared to shorter transcripts present at the same abundance in the sample, and the variability in the number of reads produced for each lane on the flow cell of the sequencing platform producing fluctuations in the number of reads produced between samples (Garber et al. 2011). A metric known as Reads Per Kilobase of transcript per Million mapped reads (RPKM) has been used in numerous studies to normalise transcript read counts by both length and by the total number of reads in the sample, permitting comparisons both within and between samples (Sims et al. 2014). Using paired-end reads for RNA-Seq can offer greater accuracy in mapping to the assembly. To support this mechanism, the Fragments Per Kilobase of transcript per Million mapped reads (FPKM) metric has been applied as it accounts for the dependency between the read pairs (Garber et al. 2011; Hart et al. 2013). Independent biological replicates, however, remain the best way to ensure adequate controls for within treatment variability and to ensure reproducibility and should be included as often as practicable (Auer and Doerge 2010).

1.2.3 De-novo genome sequencing and annotation

A genome contains the full suite of genetic information belonging to an organism. Collecting genome sequence information can provide the researcher with the complete set of instructions required for building an organism. The main objective of genome sequencing projects, of course, is to draft and order this information precisely. When planning a de-novo genome sequencing project it is important to determine the acceptable degree of completion required for its intended purpose as the achievability of producing a fully completed, error free genome progressively dissolves as the organism becomes more complex. This is especially relevant for eukaryotic organisms. To highlight this point, consider that the genome sequence of *Drosophila melanogaster* was published in the year 2000 and, although it is thought not to be a particularly difficult genome in comparison with the many others currently being constructed, its most recent release, published some seven years later, still contains gaps and draft regions (Salzberg and Yorke 2005). This is a result of limitations imposed by the short read lengths produced by the current sequencing platforms. A complication that will ensue until available technologies are able to generate continuous error-free reads that span the entire length of a chromosome. Thus, careful attention is required in the choosing of an appropriate sequencing strategy in order to maximise the prospect of faithfully reconstructing a genome.

Reads from the most accurate next-generation platforms are considerably shorter than that obtained from Sanger sequencing yet still contain some level of error. When considered independently, an error in a read is indistinguishable from a sequence variant, potentially misleading the assembler to produce false segments. This obstacle can be somewhat addressed by increasing the coverage of reads spanning the genome (Francis et al. 2013; Sims et al. 2014). The human genome was originally assembled to high quality with only ten-fold
coverage using the longer reads of Sanger sequencing. A coverage of approximately seventy three-fold was generated for the first assembly of the giant panda genome using next-generation technology, yet the assembly produced was of lower quality than that of the human (Li et al. 2010; Sims et al. 2014). This illustrates that whilst increasing coverage does aid error correction, generating greater depth of short reads does not address all issues.

One of the most difficult challenges to overcome is in resolving gaps in assemblies caused by repetitive regions. If parts of the genome contain perfectly repeated sequences, reads obtained from one region may contain the exact same sequence as reads derived from a different region (Miller et al. 2010). Unless more information is available to the assembler, it will be impossible to resolve the correct placement of this read. By using paired-end reads, a DNA fragment is sequenced from each end and, if the fragment insert size is known, one can infer the expected distance between each paired read, therefore giving additional positioning information enabling closer matching of the read to its correct region in the genome. Sample preparation methods can produce insert sizes of up to 20 Kb using mate-pair or up to 40 Kb using fosmid libraries (Zhang et al. 2012). High-quality assemblies are now often produced using hybrid approaches, incorporating different insert sizes and read lengths often from different platforms.

A bias toward the nucleotide bases guanine and cytosine (GC) that is introduced by PCR amplification of DNA during library construction has been identified as a major source of variation in coverage. New methods of sample preparation that eliminate the need for PCR amplification have been developed recently, resulting in improved coverage of high GC regions of the genome (Sims et al. 2014).

The assembly strategy used for genome projects generally involves grouping reads into contigs and the contigs into scaffolds. The scaffolds define the contig order and orientation and informs the overall structure of the size of the gaps between contigs. (Miller et al. 2010). For large datasets of more than a hundred million short reads, De Bruijn (K-mer) graph-based assemblers have been deemed more appropriate and most widely applied in the literature (Zhang et al. 2011). Assemblies are normally measured by the size and accuracy of their contigs and scaffolds. Although there are no set guidelines, an assembly where the shortest scaffold above which 50% of an assembly would be represented (the N50 scaffold length) is roughly the average gene size is a reasonable target for annotation (Miller et al. 2010; Alkan et al. 2011). The protein sequences developed for the Core Eukaryotic Genes Mapping Approach (CEGMA) pipeline can provide an additional means of estimating the completeness and contiguity of an assembly. CEGMA is used to screen the assembly against a collection of largely universal eukaryotic single-copy genes which are expected to be contained in every eukaryotic genome. From this mapping it is possible to determine the percentage of each of these genes locating discretely to an individual scaffold, providing information relating to the degree of artificial repetition (Yandell and Ence 2012). More recently, a software suite termed BUSCO has gained interest and has superseded the CEGMA pipeline to some degree, offering a more comprehensive estimate of gene content and genome completeness (Simão et al. 2015). This pipeline incorporates the use of transcriptomic information.
for building gene models of single copy orthologs instead of the small set of conserved genes used in the CEGMA approach.

DNA contamination from other organisms is another source of error and requires careful scrutiny. These contaminants can lead to the production of artefacts in the assembly that can interfere with scaffold building (Alkan et al. 2011). Measures to identify these sequences include screening the annotated regions for microbial species and others that are distantly related.

1.3 Research aims and objectives

The general aim of this research is to produce comprehensively annotated genomic datasets for two of Australia’s important aquaculture species, *Saccostrea glomerata* and *Fenneropenaeus merguiensis*, in order to address the knowledge gap relating to the availability of such information to the Australian aquaculture industry. This will be addressed through a series of studies consisting of the following four objectives:

i. To uncover the key genes associated with traits important to reproduction, development and sex determination in *F. merguiensis*, the first dataset of transcriptomic sequences from tissues relevant to biological systems governing commercially important traits will be assembled, annotated and characterised.

ii. To explore differences in the expression of genes between families of adult farmed *F. merguiensis* exhibiting extremes of variation in levels of viral load for the purpose of developing molecular markers for disease resistance and immune response to be used in the selective breeding of these animals.

iii. To reveal the sex-specific transcript expression in the digestive gland of female and male *F. merguiensis* bred and grown together in captivity in order to improve our understanding of the molecular basis for the female-superior sexual dimorphism observed in this species.

iv. To investigate the genomic complexity of *S. glomerata*, the first draft genome for this species will be constructed using a combination of sequencing, assembly and annotation methods with the goal of providing information on genomic features and to perform comparisons with the genomes of other species.
1.4 References


Chapter 2: De-novo transcriptome analysis of the banana shrimp (Fenneropenaeus merguiensis) and identification of genes associated with reproduction and development

Daniel Powell 1, Wayne Knibb 1, Courtney Remilton 2, Abigail Elizur 1,

1 Faculty of Science, Health, Education and Engineering, University of the Sunshine Coast, Maroochydore, QLD 4558, Australia
2 Seafarm Pty. Ltd., Bruce Hwy, Cardwell, QLD 4849, Australia

Available online: http://dx.doi.org/10.1016/j.margen.2015.04.006 and presented as it was accepted for publication.

2.1 Abstract
The banana shrimp, Fenneropenaeus merguiensis, is a commercially important marine crustacean for world aquaculture and fisheries. Despite this, limited genetic information is available for it and many other penaeid shrimp species. Here we present the first in-depth analysis of the transcriptional content of 8 different tissues from the banana shrimp using RNA-Seq technologies. A total of over 1 million single-end and over 49 million paired-end reads were obtained from Roche 454FLX and illumina sequencing platforms, respectively, resulting in an assembly of 124,631 transcripts with an N50 of 1,332 and mean length of 514 nt. A total of 59,179 putative protein sequences obtained from the assembled transcripts were annotated using public protein sequence databases and assigned 20,430 BLAST hits, 16,866 GO terms and 13,304 KOG categories. Further analysis revealed a rich set of transcript sequences exhibiting homology with genes associated with reproduction, sex determination and development and distinguished the tissues responsible for this expression. This report adds a substantial contribution to the sequence data currently available for F. merguiensis, providing valuable resources for further research.

2.2 Introduction
The banana shrimp, Fenneropenaeus merguiensis, is an important member of the commercially significant Penaeidae family of marine crustaceans. According to statistics from the United Nations Food and Agriculture Organization (FAO), over 116,000 tonnes of banana shrimp were produced globally from fisheries and aquaculture during 2012. Increased market demand and pressure from over-fishing has led to an expansion of penaeid shrimp culturing (Navakanitworakul et al. 2012). With large-scale aquaculture production of banana shrimp in Indonesia, Vietnam and Australia, worldwide efforts are being made to develop secure captive breeding schemes to mitigate the risks from relying on wild broodstock and to realise benefits from selective breeding programs (Andriantahina et al. 2013; Knibb et al. 2014a).
Effective domestication of a species for aquaculture relies upon good reproductive performance. Thus, improving the efficiency of larval production is of principal importance for penaeid shrimp aquaculture (Ibarra et al. 2007; Marsden et al. 2013). Banana shrimp are known to be more amenable to domestication than that of other penaeid species, such as the more extensively cultured *Penaeus monodon*, being readily matured and breed in captivity without the need for artificial insemination (Nguyen et al. 2014). Minimal differences in spawning performance between wild and pond-reared *F. merguiensis* has also been observed when bred in captivity and if better understood, this characteristic could offer insights into causes for poorer performance in other shrimps (Hoang et al. 2002).

Despite their commercial significance, only limited genetic information is available for many species of penaeid shrimp. Genetic resources for *F. merguiensis* are particularly scarce (Andriantahina et al. 2013). Key to the improvement of technologies for shrimp breeding is elucidating the molecular mechanisms occurring during reproduction (Lo et al. 2007). Understanding the genetic basis underlying the regulation of traits such as gonadal development, maturation and vitellogenesis in this species can provide information for selection programmes that can enhance the development of quality broodstock (Santos et al. 2014).

In the absence of a comprehensive genome sequence, transcriptome sequencing can be used to efficiently identify molecular markers and transcripts involved in important biological processes. The development of the RNA-Seq approach to transcriptome profiling using next-generation sequencing technologies has greatly enhanced the capacity for gene discovery in non-model species such as *F. merguiensis*, due to the vastly improved depth and sensitivity and also the availability of these technologies to researchers (Cahais et al. 2012). Recently, a few studies have successfully used this approach for immune gene annotation and single nucleotide polymorphism (SNP) discovery in *Litopenaeus vannamei* (Ghaffari et al. 2014; Yu et al. 2014). Identification of genes and pathways associated with reproduction and maturation in *F. merguiensis* will build a foundation for further studies exploring gene functionality in detail that may not only lead to improvements in captive broodstock but can also assist in informing domestication practices for other penaeid shrimp.

Here we report the first in-depth analysis of the transcriptional content of 8 different tissues in the marine shrimp *F. merguiensis* using RNA-Seq data with a focus on genes important to the understanding of the reproductive biology and development in penaeid shrimp.

2.3 Materials and Methods

2.3.1 Tissue material and RNA extraction

All samples used in this study were from animals ranging in development stage from juvenile to sexually mature and collected from grow-out ponds at a commercial farm in Cardwell, North Queensland. Tissue samples from the hepatopancreas, stomach, eye stalk, nerve cord, male gonad, female gonad, androgenic gland region and muscle/cuticle were dissected on-site and preserved in RNA/later solution (Life Technologies) before
transportation to the laboratory and storage at -80°C. Total RNA was extracted from each tissue type in pools comprised of up to 10 individuals encompassing each life stage using the RNeasy Plus Mini Kit (Qiagen). Isolation of mRNA was achieved using the Oligotex mRNA Mini Kit (Qiagen). Library preparation and sequencing using the Roche 454FLX Titanium and illumina HiSeq2000 platforms was outsourced to the Australian Genome Research Facility (AGRF).

2.3.2 Assembly, quality assessment and protein-coding region prediction

Quality of the raw sequencing reads were assessed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Trimming of adaptors, filtering of low quality reads and removing contaminants was performed using either the in-built option within Newbler v2.9 (Roche) or Trimmomatic (Bolger et al. 2014) with a custom screening database that included the UniVec database (ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec/), Illumina adaptors and common bacterial 16S sequences. Reads from the 454FLX platform were assembled using Newbler v2.9 with the default settings for cDNA except for: -urt, -minlen 40, -trim, -tr, -vt, and -vs. Clean reads from the HiSeq 2000 platform were assembled using Trinity (release r20140413) (Haas et al. 2013), using the default settings for paired-end reads except for: --min_contig_length 100, --normalize_reads. Output from the assemblers were merged using the CAP3 assembler (Huang 1999) with default settings except for -p 98 and clustered for redundancy using CD-HIT-EST (Fu et al. 2012) with a sequence identity threshold of 0.9. Low complexity transcripts were then removed using PRINSEQ-lite using functions: -lc_method dust, -lc_threshold 7 (Schmieder and Edwards 2011). Quality of the assembly was undertaking using the CEGMA pipeline (Parra et al. 2007) and clean read mapping was performed using the CLCBio Genomics Workbench v 7.0 (Qiagen). The assembled transcripts were screened for putative protein-coding regions using the program TransDecoder (release r20140704) (http://transdecoder.github.io/) with the option of including matches to the Pfam domain database (Finn et al. 2014), and with a minimum size of 30 amino acids (aa), then clustered using CD-HIT with a sequence identity threshold of 1.0.

2.3.3 Searches for sequence homologs and annotation

Software BLAST 2.2.29+ was used for alignment with homologous sequences with a local copy of the NCBI non-redundant database (nr) downloaded in September 2014 (FTP date stamp 09/06/2014). Blast2GO version 2.7 with database b2g_sept13 was used for gene ontology (GO) annotation, enzyme code mapping and InterproScan analysis (Conesa et al. 2005). Protein sequences were aligned against the NCBI KOG database (version 2/2/2011) using RPSBLAST 2.2.15 and the KEGG genes database (version 2/12/2012) using BLASTALL 2.2.15 both using an E-value cut off of 1e-5 and performed via the WebMGA server (http://weizhonglab.ucsd.edu/metagenomic-analysis/) (Wu et al. 2011). Peptide sequences from Daphnia pulex, Drosophila melanogaster, and Tribolium castaneum were obtained from ENSEMBL (release 77) (Flicek et al. 2014). The Zootermopsis nevadensis peptide sequences Official Gene Set OGSv2.2 was obtained from http://termitegenome.org/ during October 2014. Reciprocal BLAST was performed using Proteinortho v5.11 (Lechner et al. 2011), with default parameters except for: -e=1e-10, -conn=0.8, -cov=40, identity=30.
Sequences returning matches to proteins of interest were submitted to BLAST manually in order to confirm alignment specificity. Signal peptide prediction was performed using SignalP v4.0 (Petersen et al. 2011), and prediction of a transmembrane helix by TMHMM v2.0 (Krogh et al. 2001).

2.4 Results

2.4.1 Transcriptome sequencing and assembly

Roche 454FLX Titanium sequencing produced a total of 1,179,165 raw reads from 8 tissues. A total of 49,946,311 paired-end reads were generated from the hepatopancreas tissue of a single male and a single female shrimp using the illumina HiSeq 2000 platform. Sequencing yield and quality statistics of these data are presented in Table 2.1. Given no gold standard exists for the co-assembly of data from different sequencing platforms, a hybrid de-novo assembly approach was undertaken using popular assemblers for each data type. The Newbler and Trinity assemblers were used for Roche 454FLX data and illumina HiSeq data, respectively. Output from these assemblers was then converged, the redundant and low complexity sequences removed, producing a transcriptome of 124,631 non-redundant transcripts with an average length of 514.57 nucleotides (nt), an N50 of 1,332 nt and a total length of 64.13 megabases (Mb) (Table 2.2). Reads returned as singleton were not included in the final assembly. Peptide prediction from the non-redundant nucleotide sequences using Transdecoder software produced a set of 57,640 protein sequences with a maximum size of 6,500 aa and a mean length of 166.89 aa.

Table 2.1 Summary of Roche 454FLX Titanium and illumina HiSeq 2000 sequencing.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>AG</th>
<th>MG</th>
<th>FG</th>
<th>HP</th>
<th>ES</th>
<th>NC</th>
<th>ST</th>
<th>MC</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number raw reads 454</td>
<td>137,246</td>
<td>164,710</td>
<td>170,972</td>
<td>172,047</td>
<td>190,987</td>
<td>58,453</td>
<td>89,093</td>
<td>195,657</td>
<td>1,179,165</td>
</tr>
<tr>
<td>Number clean reads 454</td>
<td>86,877</td>
<td>116,754</td>
<td>148,780</td>
<td>142,345</td>
<td>157,728</td>
<td>45,727</td>
<td>56,468</td>
<td>146,483</td>
<td>901,162 (76.42%)</td>
</tr>
<tr>
<td>Number raw reads HiSeq (PE)</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>49,946,311</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>49,946,311</td>
</tr>
<tr>
<td>Number clean reads HiSeq (PE)</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>43,801,566</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>43,801,566 (87.69%)</td>
</tr>
<tr>
<td>Average clean read length (nt)</td>
<td>319</td>
<td>313</td>
<td>320</td>
<td>367 / 93*</td>
<td>337</td>
<td>315</td>
<td>301</td>
<td>300</td>
<td>/</td>
</tr>
<tr>
<td>Total clean bases 454 (Mb)</td>
<td>27.75</td>
<td>36.60</td>
<td>44.68</td>
<td>52.31</td>
<td>53.24</td>
<td>14.43</td>
<td>17.04</td>
<td>44.00</td>
<td>290.09</td>
</tr>
<tr>
<td>Total clean bases HiSeq (Mb)</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>8,052</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>8,052</td>
</tr>
</tbody>
</table>

* Indicates reads produced using the illumina HiSeq platform. Abbreviations: AG, Androgenic gland region; MG, male gonad; FG, female gonad; HP, hepatopancreas; ES, eye stalk; NC, nerve cord; ST, stomach; MC, muscle/cuticle.

2.4.2 Estimation of transcriptome completeness

Since a reference genome is unavailable for this species to assist in transcriptome reconstruction, an examination of the completeness of the reference assembly by comparison to the set of core eukaryotic genes (CEG) using the CEGMA method was performed. The percentage of complete CEG proteins found in the transcriptome assembly was 97.58%. With partial matches included, the number of CEG proteins found lifted
to 98.93%. Clean read mapping from both sequencing platforms to the assembled transcriptome was performed using an alignment stringency of 90% similarity and 70% length fraction. Of these clean reads, 92.19% and 98.21% were aligned from the Roche 454 and illumina HiSeq platforms respectively, with 89.83% of the illumina HiSeq reads mapping in pairs. This result suggests the transcriptome assembly contained most of the information obtainable from the high quality read data.

Table 2.2 Statistics of hybrid de-novo assembly of the banana shrimp (*F. merguiensis*) transcriptome and overview of annotation results.

<table>
<thead>
<tr>
<th>Assembler output</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total isotig number (Newbler/454)</td>
<td>33,440</td>
</tr>
<tr>
<td>Total contig number (Trinity/HiSeq)</td>
<td>128,424</td>
</tr>
<tr>
<td>Hybrid assembly</td>
<td></td>
</tr>
<tr>
<td>Total non-redundant transcripts</td>
<td>124,631</td>
</tr>
<tr>
<td>Total consensus length (nt)</td>
<td>64,131,367</td>
</tr>
<tr>
<td>Largest transcript</td>
<td>20,192</td>
</tr>
<tr>
<td>Mean transcript length (nt)</td>
<td>514.57</td>
</tr>
<tr>
<td>N50 of transcripts</td>
<td>1,332</td>
</tr>
<tr>
<td>% GC content</td>
<td>42</td>
</tr>
<tr>
<td>Number predicted proteins</td>
<td>57,640</td>
</tr>
<tr>
<td>Longest ORF</td>
<td>6,500</td>
</tr>
<tr>
<td>Mean ORF length</td>
<td>166.98</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Annotation of predicted protein sequences</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number sequences with BLAST hits (nr)</td>
<td>20,430</td>
</tr>
<tr>
<td>Sequences with GO annotations</td>
<td>16,866</td>
</tr>
<tr>
<td>Number of sequences with KOG hits</td>
<td>13,304</td>
</tr>
<tr>
<td>Number of sequences with KEGG hits</td>
<td>13,482</td>
</tr>
<tr>
<td>Sequences assigned enzyme codes</td>
<td>3,482</td>
</tr>
<tr>
<td>Number orphan sequences</td>
<td>37,235</td>
</tr>
</tbody>
</table>

2.4.3 Annotation of predicted protein-coding regions

The set of predicted protein sequences from the banana shrimp were examined for matches to homologs by alignment using the BLASTp algorithm against the NCBI nr protein database returning 18,891 hits at an $E$-value cut-off of 1e-5. These protein sequences were enriched with an additional 1,539 transcripts not included in the Transdecoder output yet returned a BLASTx hit when searched against the same database with the same $E$-value cut-off. From a total of 59,179 sequences, 20,430 (34.52 %) returned a hit from BLAST searches. The top-hit species distribution favoured arthropods with 68.34 % of the top-hit alignments having a similarity higher than 60 % (Figure 2.1).
Annotation of these sequences using the Blast2GO program resulted in 16,866 sequences with assignments to GO terms, of which 14,718 were confirmed with matches from InterproScan. When including all databases, the total number of sequences returning an annotation result was 21,944. Figure 2.2 shows the distribution of sequences across the top 20 GO terms for each of the three categories at Level 2. The majority of the GO annotations were mapped to hits with the IAS evidence code and mapped to the UniProtKB database. Alignment of the predicted *F. merguiensis* protein sequences with the NCBI eukaryotic orthologous groups of proteins (KOG) database resulted in 13,304 sequences with hits being classified into all 26 KOG categories. The most highly represented being in the T category (Signal transduction mechanisms), R category (General function prediction only) and the K category (Transcription), respectively (Figure 2.3). Annotation results were further examined for association with KEGG pathways revealing 13,482 sequences that had hits with genes involved with 282 different pathway maps in the KEGG database (Supplementary File 2). A large number of protein sequences (62.91 %) returned no annotation result after analysis using the above pipeline and were considered orphans. Of these orphan sequences, 993 were predicted to contain a signal peptide using SignalP, 1,821 were predicted to contain a transmembrane helix using TMHMM software.
2.4.4 Transcript abundance across tissues

This study was not designed for quantitative expression analysis and a large disparity is observed among the volume of sequence data obtained from each tissue. However, read mapping revealed a small percentage of transcripts that were found to be uniquely represented within a single tissue type. A high proportion of these
transcripts have been identified as orphan. Mapping of the clean reads to the *F. merguiensis* transcriptome assembly also revealed 1,450 transcripts that were found to be commonly expressed throughout all 8 tissue types.

### 2.4.5 Taxonomy and cross species comparison

Only 84 of the 20,430 BLAST top hits returned a match for *F. merguiensis*. However, the more intensively studied penaeid species *Penaeus monodon* and *Litopenaeus vannamei* were included in the 10 highest ranked species for BLAST top-hits with 518 and 448 sequences, respectively. Separate at the Class rank of taxonomy from the banana shrimp, the most closely related species and only crustacean to have a high-quality published genome sequence available is the water flea (*D. pulex*). In order to determine the number of shared sequences between the two species, the 30,611 proteins sequences predicted from the water flea genome draft were searched for homology with the predicted protein sequences from the banana shrimp transcriptome using reciprocal BLASTp. Only 5,791 hits were identified with the reciprocal criteria applied using an \( E \)-value cut-off of 1e-10 (Table 2.3). To gain a better understanding of the extent proteins are shared between other species of arthropods, the three highest ranking organisms for BLAST top-hits with *D. pulex*, the dampwood termite (*Z. nevadensis*) and the red flour beetle (*T. castaneum*), along with the fruit fly (*D. melanogaster*), were also compared with the predicted proteins from the banana shrimp transcriptome. Table 3 shows the number of hits for each species compared and the number of proteins predicted from each genome draft. A total of 7,656 protein sequences from the banana shrimp transcriptome were identified in at least one of these animals. A subset of 3,584 of these were found across all four species. Of these, 654 sequences were included in the list of 1,450 transcripts found in be expressed in all of the 8 tissues sequenced. The three most abundant GO terms among these 654 sequences were associated with binding (GO:0005488) cellular process (GO:0009987) and metabolic process (GO:0008152), respectively.

#### Table 2.3 Comparative sequence analysis.

Comparison of the predicted proteins from the *F. merguiensis* transcriptome with published protein datasets of other arthropods using reciprocal BLASTp with \( E \)-value cut-off of 1e-10.

<table>
<thead>
<tr>
<th>Species</th>
<th>Number of reciprocal BLAST hits</th>
<th>Number sequences in database</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Zootermopsis nevadensis</em></td>
<td>5,846</td>
<td>15,860</td>
</tr>
<tr>
<td><em>Daphnia pulex</em></td>
<td>5,791</td>
<td>30,611</td>
</tr>
<tr>
<td><em>Tribolium castaneum</em></td>
<td>5,397</td>
<td>16,526</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>4,969</td>
<td>26,950</td>
</tr>
</tbody>
</table>
| Total *F. merguiensis* sequences with BLAST hit | 7,656 | /
| Number *F. merguiensis* sequences common among all | 3,584 | /

---

26
2.4.6 Screening of genes associated with reproduction, sex determination and development

Several GO terms relating to reproduction were identified in the annotation process. A number of sequences were found to have an association with the GO term for multicellular organism reproduction (GO:0032504). A selection of interesting terms relating to this parent category are listed in Table 2.4 together with the number of sequences that were annotated with hits to these Go terms, indicating a broad inclusion of sequences involved in these processes.

Table 2.4 Number of hits of selected GO terms based on GO:0032504: multicellular organism reproduction.

<table>
<thead>
<tr>
<th>GO ID</th>
<th>GO TERM</th>
<th>HIT</th>
<th>EXAMPLE SEQUENCE</th>
<th>E VALUE</th>
<th>HIT ACCESSION NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0048609</td>
<td>Multicellular organismal reproductive process</td>
<td>53</td>
<td>Vitellogenin</td>
<td>0.0</td>
<td>ACV32381.1</td>
</tr>
<tr>
<td>GO:0019953</td>
<td>Sexual reproduction</td>
<td>50</td>
<td>Dmc1</td>
<td>0.0</td>
<td>ADM45305.1</td>
</tr>
<tr>
<td>GO:0022412</td>
<td>Cellular process involved in reproduction in multicellular organism</td>
<td>34</td>
<td>Pickled eggs isoform A</td>
<td>2.03E-116</td>
<td>NP_572331.2</td>
</tr>
<tr>
<td>GO:0009653</td>
<td>Anatomical structure morphogenesis</td>
<td>33</td>
<td>Homeobox protein Nkx-2.6</td>
<td>2.2E-38</td>
<td>NP_001101064.1</td>
</tr>
<tr>
<td>GO:0048519</td>
<td>Negative regulation of biological process</td>
<td>20</td>
<td>Axin 1</td>
<td>4.2E-90</td>
<td>KDR13004.1</td>
</tr>
<tr>
<td>GO:0007548</td>
<td>Sex differentiation</td>
<td>6</td>
<td>Transcription factor SOX-8</td>
<td>5.44E-38</td>
<td>XP_004751055.1</td>
</tr>
<tr>
<td>GO:0021700</td>
<td>Developmental maturation</td>
<td>6</td>
<td>Vascular endothelial growth factor receptor 2</td>
<td>1.33E-21</td>
<td>KDR14705.1</td>
</tr>
<tr>
<td>GO:0019098</td>
<td>Reproductive behaviour</td>
<td>5</td>
<td>Neuroglian-like</td>
<td>2.03E-10</td>
<td>XP_005180380.1</td>
</tr>
<tr>
<td>GO:0048589</td>
<td>Developmental growth</td>
<td>5</td>
<td>Exostosin-2</td>
<td>0.0</td>
<td>NP_001008400.1</td>
</tr>
<tr>
<td>GO:2000241</td>
<td>Regulation of reproductive process</td>
<td>4</td>
<td>Synaptotagmin 1</td>
<td>1.86E-130</td>
<td>NP_995619.2</td>
</tr>
<tr>
<td>GO:0009266</td>
<td>Response to temperature stimulus</td>
<td>3</td>
<td>Regulator of G protein signalling, putative</td>
<td>5.81E-149</td>
<td>XP_002429741.1</td>
</tr>
<tr>
<td>GO:0010817</td>
<td>Regulation of hormone levels</td>
<td>2</td>
<td>Insulin-degrading enzyme</td>
<td>0.0</td>
<td>KDR15953.1</td>
</tr>
</tbody>
</table>

A number of proteins associated with the regulation of reproduction were identified in the *F. merguiensis* transcriptome (Table 2.5). Neuropeptides such as crustacean hyperglycemic hormone (CHH), molt-inhibiting hormone (MIH), vitellogenin-inhibiting hormone (VIH/GIH) and neuroparsin were identified localising predominantly within the eye stalk. Receptors for the neurotransmitters serotonin, dopamine and octapamine, that have been shown to play a role in regulating the release of hormones involved in reproduction (Fingerman 1997), were expressed within the hepatopancreas. Key factors and receptors involved in oogenesis and embryogenesis were also found in many tissues from both male and female shrimp. Some of the cytochrome p450 genes involved in 20-hydroxyecdysone (20E) biosynthesis show expression in the hepatopancreas along with an Ecdysteroid receptor E75 homolog that is expressed throughout all 8 tissues. Farnesoid acid O-methyltransferase, the enzyme that catalyses the formation of Methyl farnesolate (MF), was found to be expressed in all tissues types. Proteins involved with the metabolism of juvenile hormone (JH) were also found within a number of different tissues. Expression of transcripts coding for proteins involved in
gonadotropin metabolism and biosynthesis along with hormone receptors for luteinising hormone (LH), gonadotropin-releasing hormone (GnRH) and estradiol were each detected within the hepatopancreas. A number of interesting proteins associated with sex determination and development were identified in the *F. merguiensis* transcriptome. Apart from the apparently tissue specific insulin-like androgenic gland hormone and the crustacean female sex hormone that has been linked with its function, all other factors listed in Table 2.6 were expressed in the hepatopancreas.
Table 2.5 Alignment results of selected homologous proteins found in the *F. mergiensis* transcriptome associated with the regulation of reproduction.

<table>
<thead>
<tr>
<th>Identity</th>
<th>Accession</th>
<th>Hit Species</th>
<th>E value</th>
<th>Similarity</th>
<th>Example Contig</th>
<th>Tissue expressing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Neurohormones</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VIH/GIH</td>
<td>AGX26044.1</td>
<td>Litopenaeus vannamei</td>
<td>3.77E-60</td>
<td>97%</td>
<td>F.merg119524</td>
<td>ES, NC</td>
</tr>
<tr>
<td>CHH</td>
<td>O97383.1</td>
<td>Peneaus monodon</td>
<td>8.52E-68</td>
<td>95%</td>
<td>F.merg119690</td>
<td>ES, HP</td>
</tr>
<tr>
<td>MIH</td>
<td>ABD73291.1</td>
<td>Litopenaeus vannamei</td>
<td>2.76E-22</td>
<td>100%</td>
<td>F.merg122603</td>
<td>ES</td>
</tr>
<tr>
<td>Neuropeptide Pem-SGP-C1</td>
<td>BAB70610.1</td>
<td>Peneaus monodon</td>
<td>7.4E-55</td>
<td>94%</td>
<td>F.merg120910</td>
<td>ES</td>
</tr>
<tr>
<td>Neuropeptide</td>
<td>AHG06296.1</td>
<td>Peneaus monodon</td>
<td>6.56E-9</td>
<td>100%</td>
<td>F.merg124013</td>
<td>ES</td>
</tr>
<tr>
<td>Neurohormones</td>
<td>AHX39208.1</td>
<td>Metapenaeus ensis</td>
<td>1.63E-18</td>
<td>68%</td>
<td>F.merg5197</td>
<td>AG, ES, FG, MC, NC, ST, HP</td>
</tr>
<tr>
<td><strong>Embryogenesis / Oogenesis associated</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vitellogenin receptor</td>
<td>ABW79798.1</td>
<td>Peneaus monodon</td>
<td>0.0</td>
<td>95%</td>
<td>F.merg118084</td>
<td>FG, MG, ES, NC, HP</td>
</tr>
<tr>
<td>Fushi tarazu-factor 1 (FTZ-F1)</td>
<td>AAD41899.1</td>
<td>Metapenaeus ensis</td>
<td>4.77E-136</td>
<td>87%</td>
<td>F.merg5036</td>
<td>AG, ES, FG, MG, NC, HP</td>
</tr>
<tr>
<td>Chorion peroxidase</td>
<td>XP_001868291.1</td>
<td>Culex quinquefasciatus</td>
<td>7.68E-120</td>
<td>76%</td>
<td>F.merg118996</td>
<td>AG, ES, MG, NC, HP</td>
</tr>
<tr>
<td>Prostaglandin F synthase</td>
<td>AFJ11397.2</td>
<td>Peneaus monodon</td>
<td>0.0</td>
<td>95%</td>
<td>F.merg8926</td>
<td>AG, ES, FG, MG, MC, HP</td>
</tr>
<tr>
<td>Prostaglandin E synthase</td>
<td>AFJ11396.1</td>
<td>Peneaus monodon</td>
<td>0.0</td>
<td>99%</td>
<td>F.merg33353</td>
<td>FG, MG, MC, HP</td>
</tr>
<tr>
<td><strong>Neurotransmitter receptors</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serotonin receptor</td>
<td>AAS05316.1</td>
<td>Metapenaeus ensis</td>
<td>6.54E-28</td>
<td>84%</td>
<td>F.merg38775</td>
<td>HP</td>
</tr>
<tr>
<td>Dopamine receptor (DA$_1$)</td>
<td>ABI64137.1</td>
<td>Panulirus interruptus</td>
<td>1.37E-48</td>
<td>95%</td>
<td>F.merg63463</td>
<td>HP</td>
</tr>
<tr>
<td>Octopamine receptor beta-3R-like</td>
<td>NP_001280505.1</td>
<td>Tribolium castaneum</td>
<td>6.32E-84</td>
<td>76%</td>
<td>F.merg27646</td>
<td>NC, HP</td>
</tr>
<tr>
<td><strong>Ecdysteroid biosynthesis and receptors</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phantom (CYP306A1)</td>
<td>BAH24005.1</td>
<td>Marsupenaeus japonicus</td>
<td>1.23E-33</td>
<td>98%</td>
<td>F.merg83694</td>
<td>HP</td>
</tr>
<tr>
<td>Disembodied (CYP302A1)</td>
<td>EFX63066.1</td>
<td>Daphnia pulex</td>
<td>6.52E-63</td>
<td>72%</td>
<td>F.merg80003</td>
<td>HP</td>
</tr>
<tr>
<td>Ecdysteroid receptor E75</td>
<td>AGS94407.1</td>
<td>Litopenaeus vannamei</td>
<td>0.0</td>
<td>97%</td>
<td>F.merg29504</td>
<td>All Tissues</td>
</tr>
<tr>
<td>Retinoid X receptor 2</td>
<td>ACN78602.1</td>
<td>Fenneropenaeus chinensis</td>
<td>6.38E-77</td>
<td>100%</td>
<td>F.merg35996</td>
<td>NC, ST ,HP</td>
</tr>
<tr>
<td><strong>Hormone enzymes and receptors</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estradiol 17-beta-dehydrogenase 12</td>
<td>XP_971720.1</td>
<td>Tribolium castaneum</td>
<td>3.42E-95</td>
<td>75%</td>
<td>F.merg4342</td>
<td>All Tissues</td>
</tr>
<tr>
<td>5-alpha-reductase alpha polypeptide 1</td>
<td>XP_004340242.1</td>
<td>Orzyias latipes</td>
<td>2.95E-54</td>
<td>48%</td>
<td>F.merg31938</td>
<td>AG, HP</td>
</tr>
<tr>
<td>Estradiol receptor-like protein 1</td>
<td>AFV74663.1</td>
<td>Portunus trituberculatus</td>
<td>1.4E-60</td>
<td>70%</td>
<td>F.merg3370</td>
<td>MG, NC, HP</td>
</tr>
<tr>
<td>Gonadotropin-releasing hormone receptor</td>
<td>XP_001655249.1</td>
<td>Aedes aegypti</td>
<td>1.18E-28</td>
<td>74%</td>
<td>F.merg84401</td>
<td>HP</td>
</tr>
<tr>
<td>Lutropin-choriogonadotropic hormone receptor-like</td>
<td>XP_003243835.1</td>
<td>Acrythosiphon pisum</td>
<td>2.35E-23</td>
<td>62%</td>
<td>F.merg119239</td>
<td>AG, NC</td>
</tr>
<tr>
<td>Thyroid hormone receptor beta-A</td>
<td>EKC30867.1</td>
<td>Crassostrea gigas</td>
<td>5.27E-29</td>
<td>60%</td>
<td>F.merg9336</td>
<td>FG, HP</td>
</tr>
<tr>
<td>Nuclear progesterone receptor</td>
<td>ADY38993.1</td>
<td>Peneaus monodon</td>
<td>1.33E-85</td>
<td>99%</td>
<td>F.merg3829</td>
<td>ES, FG, MG, MC, NC, ST, HP</td>
</tr>
<tr>
<td><strong>MF and JH metabolism</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Farnesoic acid O-methyltransferase</td>
<td>AAX24111.1</td>
<td>Litopenaeus vannamei</td>
<td>0.0</td>
<td>99%</td>
<td>F.merg28364</td>
<td>All Tissues</td>
</tr>
<tr>
<td>Juvenile hormone epoxide hydrolase 1</td>
<td>KDR10172.1</td>
<td>Zootermopsis nevadensis</td>
<td>1.11E-140</td>
<td>68%</td>
<td>F.merg29956</td>
<td>ES, FG, MG, MC, NC, HP</td>
</tr>
<tr>
<td>Juvenile hormone-inducible protein</td>
<td>XP_001648737.1</td>
<td>Aedes aegypti</td>
<td>4.08E-40</td>
<td>50%</td>
<td>F.merg4882</td>
<td>AG, ES, FG, MG, MC, ST, HP</td>
</tr>
<tr>
<td>Juvenile hormone esterase</td>
<td>EFX78605.1</td>
<td>Daphnia pulex</td>
<td>2.54E-68</td>
<td>59%</td>
<td>F.merg33208</td>
<td>ST, HP</td>
</tr>
<tr>
<td>3-hydroxy-3-methylglutaryl-CoA reductase</td>
<td>KDR13783.1</td>
<td>Zootermopsis nevadensis</td>
<td>2.56E-18</td>
<td>62%</td>
<td>F.merg14131</td>
<td>HP</td>
</tr>
</tbody>
</table>

Abbreviations: AG, Androgenic gland region; ES, eye stalk; FG, female gonad; MG, male gonad; NC, nerve cord; MC, muscle/cuticle; ST, stomach; HP, hepatopancreas.
Table 2.6 Alignment results of selected homologous proteins found in the *F. merguiensis* transcriptome known to be involved with sex determination and development.

<table>
<thead>
<tr>
<th>Identity</th>
<th>Accession</th>
<th>Species</th>
<th>E value</th>
<th>Similarity</th>
<th>Example Contig</th>
<th>Tissue expressing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insulin-like androgenic gland hormone</td>
<td>AFU60547.1</td>
<td><em>Fenneropenaeus chinensis</em></td>
<td>5E-50</td>
<td>97%</td>
<td>F.merg121390</td>
<td>AG</td>
</tr>
<tr>
<td>Crustacean female sex hormone</td>
<td>AEI72264.1</td>
<td><em>Carcinus maenas</em></td>
<td>1.49E-29</td>
<td>57%</td>
<td>F.merg121663</td>
<td>ES</td>
</tr>
<tr>
<td>VASA</td>
<td>ABQ00071.1</td>
<td><em>Fenneropenaeus chinensis</em></td>
<td>0.0</td>
<td>97%</td>
<td>F.merg8657</td>
<td>FG, MG, ST, HP</td>
</tr>
<tr>
<td>Tudor and KH domain-containing protein</td>
<td>KDR15021.1</td>
<td><em>Zooptermopsis nevadensis</em></td>
<td>1.92E-108</td>
<td>62%</td>
<td>F.merg29878</td>
<td>AG, FG, MG, MC, HP</td>
</tr>
<tr>
<td>Piwi-1</td>
<td>AGV15455.1</td>
<td><em>Portunus trituberculatus</em></td>
<td>0.0</td>
<td>84%</td>
<td>F.merg1737</td>
<td>FG, MG, HP</td>
</tr>
<tr>
<td>Piwi-2</td>
<td>AGV15456.1</td>
<td><em>Portunus trituberculatus</em></td>
<td>9.31E-84</td>
<td>74%</td>
<td>F.merg118277</td>
<td>AG, FG, MG, HP</td>
</tr>
<tr>
<td>Piwi-3</td>
<td>AGV15457.1</td>
<td><em>Portunus trituberculatus</em></td>
<td>0.0</td>
<td>81%</td>
<td>F.merg33191</td>
<td>ST, HP</td>
</tr>
<tr>
<td>Maternal protein pumilio</td>
<td>KDR20699.1</td>
<td><em>Zooptermopsis nevadensis</em></td>
<td>0.0</td>
<td>78%</td>
<td>F.merg8833</td>
<td>All Tissues</td>
</tr>
<tr>
<td>DMT1</td>
<td>AHI47024.1</td>
<td><em>Macrobrachium</em></td>
<td>1.15E-12</td>
<td>100%</td>
<td>F.merg74818</td>
<td>HP</td>
</tr>
<tr>
<td>Transformer-2b</td>
<td>AFU60544.1</td>
<td><em>Fenneropenaeus chinensis</em></td>
<td>1.71E-80</td>
<td>99%</td>
<td>F.merg5835</td>
<td>FG, MG, MC, NC, ST, HP</td>
</tr>
<tr>
<td>Transformer-2c</td>
<td>AFU60545.1</td>
<td><em>Fenneropenaeus chinensis</em></td>
<td>1.36E-86</td>
<td>93%</td>
<td>F.merg31878</td>
<td>FG, MC, NC, HP</td>
</tr>
<tr>
<td>FEM-1</td>
<td>AHA90856.1</td>
<td><em>Litopenaeus vannamei</em></td>
<td>0.0</td>
<td>99%</td>
<td>F.merg4023</td>
<td>All Tissues</td>
</tr>
<tr>
<td>Sex-lethal 1</td>
<td>AGI44577.1</td>
<td><em>Macrobrachium</em></td>
<td>2.42E-133</td>
<td>86%</td>
<td>F.merg29925</td>
<td>All Tissues</td>
</tr>
<tr>
<td>Argonaute 1</td>
<td>AFQ31557.1</td>
<td><em>Marsupenaeus japonicus</em></td>
<td>0.0</td>
<td>99%</td>
<td>F.merg6296</td>
<td>AG, ES, FG, MG, MC, NC, HP</td>
</tr>
<tr>
<td>Extra sex combs</td>
<td>AGI50961.1</td>
<td><em>Macrobrachium</em></td>
<td>0.0</td>
<td>92%</td>
<td>F.merg2825</td>
<td>ES, FG, MG, MC, HP</td>
</tr>
<tr>
<td>Forkhead box protein L2</td>
<td>XP_004917868.1</td>
<td><em>Xenopus tropicalis</em></td>
<td>2.22E-17</td>
<td>76%</td>
<td>F.merg111286</td>
<td>HP</td>
</tr>
<tr>
<td>Sex determination protein fruitless-like</td>
<td>XP_005186912.1</td>
<td><em>Musca domestica</em></td>
<td>1.45E-26</td>
<td>66%</td>
<td>F.merg2143</td>
<td>FG, MG, NC, HP</td>
</tr>
<tr>
<td>Transcription factor SOX-5</td>
<td>EFN64823.1</td>
<td><em>Camponotus floridanus</em></td>
<td>2.31E-59</td>
<td>91%</td>
<td>F.merg7309</td>
<td>AG, ES, MG, HP</td>
</tr>
<tr>
<td>Star protein</td>
<td>KDR15857.1</td>
<td><em>Zooptermopsis nevadensis</em></td>
<td>6.27E-90</td>
<td>74%</td>
<td>F.merg2984</td>
<td>ES, NC, HP</td>
</tr>
</tbody>
</table>

Abbreviations: AG, Androgenic gland region; ES, eye stalk; FG, female gonad; MG, male gonad; NC, nerve cord; MC, muscle/cuticle; ST, stomach; HP, hepatopancreas.
2.5 Discussion

2.5.1 Transcriptome assembly

In this manuscript, we present a draft transcriptome of 8 tissues from the important marine shrimp *F. merguiensis*. The objective of this research was to characterise the transcriptional content of these tissue types and to identify interesting gene sequences in order to provide a resource for future research in penaeid biology and to support the enhancement of commercial breeding programs.

The assembly strategy used in this study to maximise the value of sequence read data from different platforms is not uncommon in the literature. The Newbler software package is commonly reported as a top performing assembler for 454 generated reads (Mundry et al. 2012; Ren et al. 2012). Likewise, the Trinity assembler has been cited as an effective method to assemble illumina generated RNA-Seq reads (Finseth and Harrison 2014). Successful hybrid assemblies employing a clustering approach and involving sequencing reads from different platforms have also been previously reported (Melicher et al. 2014; Roulin et al. 2014). Our assembly resulted in above 90 % of all clean reads mapping, an N50 of 1,332, a mean read length of 514.57 and almost 99 % completeness with the CEGMA pipeline core gene set, allowing us confidence in the down-stream analysis of the transcriptome assembly. The more relaxed alignment stringency of 0.9 % for clustering of the transcripts was used to address the high levels polymorphisms expected from using multiple animals per sample.

2.5.2 Annotation and comparative analysis

Alignments to public databases were successful in annotating just over a third of the predicted protein sequences from the transcriptome assembly. As expected, the highest contributors of BLAST top-hits came from organisms with published genome sequences. To date, only 105 protein sequences have been deposited for *F. merguiensis* (Taxonomy ID: 71412) in the NCBI nr database explaining the low number of BLAST top-hits for this species. The majority of the banana shrimp sequences had no evident homolog or predicted function further highlighting the lack of genomic resources for this species exemplified by Andriantahina and colleagues (Andriantahina et al. 2013), and consequently offering a large collection of perhaps novel transcripts inviting further exploration. GO terms such as growth, immune system process, reproduction, reproductive process and developmental process were featured in the functional annotation of the *F. merguiensis* transcriptome. Each of these terms were represented in the top 20 categories for Biological Process, suggesting these data include a rich source of information of interest for aquaculture. As discussed in a recent review by Santos and colleagues, RNA-Seq data such as these may inform further research seeking to identify quantitative trait loci (QTL) for traits in these GO categories (Santos et al. 2014). Even though *D. pulex* shares a closer phylogenetic relationship, the annotation results of the gene content from the *F. merguiensis* transcriptome appears to share a higher level of similarity to the protein sequences in the *Z. nevadensis* database. This result is supported by the outcome of the reciprocal BLAST alignments with these species. This may be due to the *D. pulex* genome supporting a high gene count with a large proportion of its genes reported to be unique among any other available proteome (Colbourne et al. 2011).
A group of 654 sequences were found commonly expressed across all tissue types and were also identified in 4 other broadly related arthropod species compared using a reciprocal BLAST alignment. Considering GO terms for many of these sequences relate to basic cellular functions, this group of 654 sequences may perhaps be a set of highly conserved genes responsible for essential cellular processes or “housekeeping” genes that could be explored further for use in gene expression experiments. High-throughput sequencing data are often used as a rich source of genetic markers for use in genotype studies. Although the content of candidate marker regions are not reported here, a comprehensive set of single sequence repeats (SSRs) were identified in this transcriptome data. A number of these markers were developed and employed for research described previously (Knibb et al. 2014a; Knibb et al. 2014b; Nguyen et al. 2014).

2.5.3 Expression of genes associated with reproduction and development

Gene sequences related to reproductive performance and gonadal maturation were identified in the *F. merguiensis* transcriptome. The CHH neuropeptide family have been demonstrated to have a regulatory role in reproduction in penaeid shrimp (Huberman 2000; Ibarra et al. 2007). The identification of transcripts coding for CHH family peptides in this study may assist in further research into alternatives to eyestalk ablation methods for spawning induction. Consistent with other reports, no homolog of the mandibular organ-inhibiting hormone (MOIH) was detected in the *F. merguiensis* transcriptome data (Ibarra et al. 2007). However, it is difficult to confirm if MOIH is absent in this species using an alignment approach to peptides from other crustaceans given it has a high sequence similarity with other members of the CHH family (Webster et al. 2012). Ecdysteroids have also been shown to play an important role in reproductive processes such as ovarian growth (Swetha and Sainath 2011). The hormone 20E is thought to be primarily involved in moultng of crustaceans, though has been shown to also stimulate vitellogenesis in decapods such as *Macrobrachium rosenbergii* (Ibarra et al. 2007). In this study, transcripts coding for an ecdysteroid receptor E75, which is induced downstream by the 20E-retinoid X receptor (RXR) complex, were expressed in all tissues including the male and female gonad, further supporting the widespread role of ecdysteroid control in shrimp (Qian et al. 2014).

Methyl farnesoate (MF) is thought of as the crustacean equivalent of juvenile hormone III (JH) and has been observed to stimulate ovarian development in marine shrimp (Rotllant et al. 2000; Nagaraju 2007). MF has also been demonstrated to stimulate vitellogenesis in the shrimp, *Sicyonia ingentis* (Paran et al. 2010). Transcripts encoding sequences for enzymes in the MF metabolic pathway have been identified here in the gonad tissues and hepatopancreas of *F. merguiensis*. A study by Makkapan and colleagues describes the influence of serotonin and MF in the stimulation of ovarian development of *F. merguiensis* (Makkapan et al. 2011). Serotonin has been shown to significantly inhibit MF synthesis in crustaceans (Laufer et al. 1993). Interestingly, sequences encoding a serotonin receptor and many of the enzymes associated with MF catabolic pathways were found to be expressed in the hepatopancreas, suggesting this organ may be involved in facilitating the interaction between serotonin and MF.
Enzymes with the capacity to synthesize steroids more typically associated with vertebrates have been reported in the gonads of crustaceans (Benzie 1998; Swetha and Sainath 2011). Progesterone and estradiol have been shown to stimulate ovarian development and vitellogenesis in shrimp and peak levels of progesterone and 17β-estradiol has been observed in the hepatopancreas of mature stage *P. monodon* (Quinitio et al. 1994; Swetha and Sainath 2011). Sequences encoding enzymes and receptors linked with steroid hormones such as progesterone, estradiol and testosterone were found to be expressed in the androgenic gland region, gonad tissue and hepatopancreas in the *F. merguiensis* transcriptome, providing some support for a role for these hormones in the regulation of reproduction in this species.

Transcripts coding for an insulin-like androgenic gland hormone, widely accepted to be a key factor in masculinisation (Ventura et al. 2011), was found to be expressed specifically in the male androgenic gland region. A transcript homologous to a recently reported crustacean female sex hormone involved in the regulation of female reproductive development (Zmora and Chung 2014), was found uniquely expressed in the banana shrimp eye stalk. Other sex determining gene homologs for the transformer-2 (Tra-2) and doublesex (DMRT-1) proteins, reported to be associated with female sex determination in *F. chinensis* (Li et al. 2012), were found to be expressed in the gonad tissues and hepatopancreas. Pond-reared female banana shrimp have demonstrated faster growth than their male counterparts (Hoang et al. 2002), thus, the identification of sequences encoding these sex determining proteins could assist in research seeking to develop a monosex culture of this species.

RNA-Seq data from the hepatopancreas offered a much deeper coverage when compared to the other 7 tissues, owing to the contribution of reads from the illumina HiSeq platform. This likely explains the majority of interesting transcripts found in this study being detected in this tissue. However, it also confirms this organ has an extensive role in the reproductive biology of this species.

### 2.6 Conclusion

The successful assembly of a transcriptome from multiple tissues has enabled the identification and characterisation of genes of importance for the understanding of penaeid shrimp reproductive biology and development whilst also revealing a large number of unknown transcripts inviting further examination. The present study adds a substantial contribution to the sequence data available for *F. merguiensis*, providing valuable resources for further studies.

### 2.7 Availability of Supporting Data

Read data is available from the NCBI Short Read Archive under BioSample submission numbers: SAMN03198704, SAMN03198705, SAMN03198706, SAMN03198707, SAMN03198708, SAMN03198709, SAMN03198710, SAMN03198711, and SAMN03198712. Assemblies are available from the authors by request.
2.8 Acknowledgments
We gratefully acknowledge the support of the Australian Seafood CRC (project 2009/724), the Fisheries Research Development Cooperation, the Australian Prawn Farmers Association and a Marie Curie International Research Staff Exchange Scheme Fellowship within the 7th European Community Framework Programme (612296-DeNuGReC). We thank Anna Kuballa, Paul Whatmore, Nicole Ertl, Daryle Sullivan, Angela Herd and David Bright for their contributions to sample collection and processing.

Available online
Supplementary File 1. Results of putative peptide sequence alignment with KOG genes.
Supplementary File 2. Results of putative peptide sequence alignment with KEGG pathway genes.
2.9 References


Chapter 3: Transcriptional profiling of banana shrimp 
*Fenneropenaeus merguiensis* with differing levels of viral load

Daniel Powell1, Wayne Knibb1, Nguyen Hong Nguyen2 and Abigail Elizur1

1Genecology Research Centre, Faculty of Science, Health, Education and Engineering, University of the Sunshine Coast, Maroochydore DC, Queensland 4558, Australia

From the symposium “Tapping the Power of Crustacean Transcriptomics to Address Grand Challenges in Comparative Biology” presented at the annual meeting of the Society for Integrative and Comparative Biology, January 3–7, 2016 at Portland, Oregon.

Published in *Integrative and Comparative Biology* (2016) 56 (6): 1131‐1143.

Available online: [http://dx.doi.org/10.1093/icb/icw029](http://dx.doi.org/10.1093/icb/icw029) and presented as it was accepted for publication.

### 3.1 Synopsis

Viral pathogens are of serious concern to the culture of penaeid shrimps worldwide. However, little is known about the molecular response of shrimp to viral infection. Selective breeding has been suggested as an effective long-term strategy to manage viral disease, though more information on gene function is needed to help inform breeding programs. The study of cultured banana shrimp (*F. merguiensis*) infected with hepatopancreatic parvo-like virus (HPV) provides a unique opportunity to explore the host response to viral infection independent of challenge testing. To gain insight into the genetic mechanisms underlying resistance to high levels of HPV, we examined hepatopancreas tissue from 6 full-sib groups of banana shrimp with differing levels of HPV infection for differences in gene expression. A total of 404 differentially expressed genes were identified with 180 being over-expressed and 224 under-expressed among high-HPV full-sib groups.

Based on homology analysis, a large proportion of these genes were associated with processes reported to be involved in the immune response of crustaceans, including pattern recognition proteins, antimicrobial peptides, components of the prophenoloxidase system and antiviral activity. The results indicate shrimp from high-HPV full-sib groups appear to have a lower presence of important immune response elements, yet possess upregulated putative antiviral pathways. Within the differentially expressed genes, over 4,000 sequence variants were identified to be exclusive to either the high or low-HPV full-sib groups. To our knowledge, this is the first report of differential expression analysis using RNA-Seq to explore differences in viral load among high and low-HPV full-sib groups of cultured shrimp. This research has provided additional insight into our understanding of the mechanisms involved in the response of this shrimp species to a naturally occurring viral pathogen. Sequence variants identified in this study offer an exceptional resource for mining gene-associated markers of HPV resistance.
3.2 Introduction

The control of disease is currently one of the most important challenges concerning the rapidly expanding cultured shrimp industry. Viral pathogens are among the most serious causes of disease in penaeid shrimps, generating huge economic losses worldwide (Knibb et al. 2015). Selective breeding has been suggested as an effective long term strategy to manage viral infection wherein selection methods for development of pathogen resistant stocks are often based on survivability when challenged with a particular pathogen (Robinson et al. 2014). However, microbial challenge assays are often unreliable due to bias associated with the introduction of the pathogen and survivability may, in part, be due to factors different to that of resistance to the initial infection (Gitterle et al. 2005). Furthermore, selection of survivors from extremely virulent diseases can rapidly reduce the genetic variation in the population if not carefully managed (Cock et al. 2009) and for some species, heritability for survival to pathogen challenge has been reported as extremely low and negatively correlated with growth (Robinson et al. 2014). Nevertheless, progress has been made in the selection of some shrimp species resistant to certain pathogens, such as Taura Syndrome Virus (TSV), although the factors responsible for this resistance are still poorly understood (Moss et al. 2013).

Recent studies have used high-throughput sequencing technologies (e.g., RNA-Seq) to examine the transcriptional profile of penaeid shrimp challenged with infection from white spot syndrome virus (WSSV) (Chen et al. 2013; Xue et al. 2013) and TSV (Sookruksawong et al. 2013; Zeng et al. 2013) revealing valuable information relating to gene functions associated with immune response, resistance and host-virus interactions. However, the molecular mechanisms underlying many of these processes remain largely unknown. RNA-Seq studies have been used successfully to identify sequence variants that co-locate with differentially expressed genes. The gene function information from studies such as these can be coupled with quantitative trait loci (QTL) to provide markers for selective breeding programs. The absence of an informative QTL for WSSV and other important shrimp viruses may, among other things, be due to challenge methods not accurately reporting resistance phenotypes (Robinson et al. 2014). Approaches other than challenge testing may be needed to improve the availability of markers for disease resistance.

Hepatopancreatic parvo-like virus (HPV) is a single-stranded DNA virus infecting penaeid shrimp species and is believed to cause slow growth that has been observed to impact production (Flegel 2006; Sukhumsirichart et al. 2006). Banana shrimp (F. merguiensis) are economically valuable cultured shrimp species to Australia, Indonesia and Vietnam (Nguyen et al. 2014). HPV infection was first reported in cultured banana shrimp in 1982 from Singapore and has been present in a number of shrimp farms culturing banana shrimp in Queensland, Australia (Yan et al. 2010; Owens et al. 2011). Transmission of the virus is thought to be via ingestion, though there are some reports of vertical transmission from parental broodstock (Flegel 2006). A recent study by Knibb et al. reported estimates of heritability for resistance of cultured banana shrimp to HPV infection as moderate to large when measuring levels of virus concentration (copy number) within family groups (Knibb et al. 2015). The authors propose that a method of selection based on the viral copy number from an animal rather than its propensity to survive infection may offer advantages by avoiding difficulties in
challenge testing, and suggest that the interplay between banana shrimp and HPV might be useful as a model to understand the genetic mechanisms of resistance, independent of survivability. Thus, the main objective of this study was to enable some insight into the mechanisms underlying resistance to HPV at a genetic level. For this purpose, hepatopancreas tissue from shrimp belonging to full-sib groups with significantly different HPV load were examined for differences in levels of gene expression using RNA-Seq.

3.3 Methods

3.3.1 Sample collection and RNA extraction

All samples used in this study were from banana shrimp showing no visible sign of disease bred and grown together that were collected over a period of 4 days in August from the same grow-out pond at a commercial farm in Cardwell, North Queensland. All animals sampled were in intermolt. Details of microsatellite genotyping for pedigree assignment and measurement of HPV concentration using qPCR have been previously reported (Knibb et al. 2015). Samples were chosen from a cohort of 278 animals assigned to 27 full-sib groups based on genotype data, each with a detectable HPV load. Four individuals from 3 full-sib groups with the highest viral load and 4 individuals from 3 full-sib groups with the lowest viral load (n=24) were sequenced from this cohort. Full-sib groups with too few members for adequate replication were not considered. Given each sample contained a detectable HPV load, there were no control or non-infected animals included in this study. Thus, conclusions drawn from the comparisons are made relative to these two conditions. The hepatopancreas was excised from banana shrimp on-site and stored in RNALater (Life Technologies) for transport and for storage at -80 °C. RNA was extracted and purified from 20 mg of tissue taken from a longitudinal section of the caudal region using the RNeasy Plus Universal kit (Qiagen) as per the manufacturer’s instructions. Library preparation using the illumina TruSeq RNA library prep kit and sequencing of the randomly pooled RNA libraries across 4 lanes on the illumina HiSeq2500 platform was outsourced to the Australian Genome Research Facility (AGRF). The data sets supporting the results of this article are available in the NCBI SRA repository under BioSample accession numbers SAMN04270865 and SAMN04270866, http://www.ncbi.nlm.nih.gov/biosample/.

3.3.2 Transcriptome assembly, ORF prediction and quality assessment.

Quality of the raw read data was assessed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Trimming of adaptors, filtering of low quality reads and removing contaminants were performed using Trimmomatic (Bolger et al. 2014) with a custom screening database. Clean reads were assembled using Trinity (release r20140717) using the default settings for paired-end reads except for --normalize_reads --min_kmer_cov 2 --bfly_opts --CuffFly --extended_lock. Contigs from the Trinity output were clustered for redundancy using CD-HIT-EST (Li and Godzik 2006) with a sequence identity threshold of 0.9. The non-redundant transcript assembly was screened for putative protein-coding regions using the program TransDecoder (release 2.0.1) (http://transdecoder.github.io/) with the
option of including matches to the Pfam domain database (Finn et al. 2014), and BLASTx hits to the Uniref90 database downloaded in February 2015 (EMBL date stamp 07/01/2015) and with a minimum size of 30 amino acids (aa), then clustered using CD-HIT using a sequence identity threshold of 0.9. Quality of the assessment was undertaken using the CEGMA (Parra et al. 2007) pipeline and clean read mapping to the full non-redundant assembly was performed using Bowtie2 v2.2.3 (Langmead and Salzberg 2012) using the default parameters. Functional annotation of the protein-coding sequences including gene ontology enrichment analysis was performed as described in Supplementary Methods S3.1.1.

3.3.3 Differential gene expression analysis

Clean read data from each individual sample was mapped to the subset of gene group transcripts using the align_and_estimate_abundance.pl script from the Trinity v2.0.6 software package using default parameters except for --est_method RSEM --aln_method bowtie2 --trinity_mode. Reads not mapping in pairs were discarded from DE analysis. The output was analysed using the Bioconductor (Huber et al. 2015) package DESeq2 (v1.6.3) (Love et al. 2014) in R. Differential gene expression was determined using a multifactor design incorporating conditions for high or low viral load and for the lane of the HiSeq flow cell from which the samples were sequenced. We considered only the gene groups calculated to have at least a 1.1 fold difference in expression and an FDR corrected p-value of <0.05 as significant in line with other recent reports from non-mammalian studies (Castillo et al. 2015; Koltes et al. 2015; McTaggart et al. 2015).

3.3.4 Variant detection

To search for variants that are present consistently throughout either the high or low virus full-sib groups, read mapping was performed using Bowtie2 against the transcript list of DEGs using the default parameters except for --no-mixed --no-discardant. Variant calls were made using the SAMtools/BCFtools v1.2 (Li 2011) software packages. Variants were filtered for those unique to either the high or low virus full-sib groups using the filter against known variants function of the CLC Genomics Workbench v8.0.2 (Qiagen).

3.4 Results

3.4.1 Sample selection, RNA sequencing and transcriptome assembly

The mean Log10 HPV gene copies per µg of DNA was determined to be significantly different among the 27 full-sib groups of banana shrimp (ANOVA $F_{1,26} = 3.4$, $P < 0.001$) (Supplementary Figure 3.2). An overview of the sample information for each animal sequenced is provided in Supplementary Table 1 showing the HPV gene copy numbers between shrimp selected from high versus low full-sib groups to differ by 3 to 6 orders of magnitude. Each of the 24 sequencing libraries produced a total yield of 106.37 Gb of raw sequence data (Figure 3.1). The average number of paired-end (PE) reads per library after quality trimming was 19.4 million. The output of the Trinity software and subsequent redundancy removal resulted in 130,677 non-redundant transcripts yielding an overall consensus length of 122 Mb (Table 3.1). The average transcript length was 939.5 bp ranging from 200 to 37,222 bp with the N50 calculated to be 1,881 bp. Over 96% of reads mapped to
the assembly with 89.26% mapping concordantly in pairs. Comparison of the assembled transcriptome using the CEGMA pipeline returned 97.98% and 99.19% complete and partial matches, respectively. A total of 32,960 transcripts were predicted to contain protein-coding regions that were clustered into 18,559 unique components as defined by the Trinity pipeline, which we refer to as unique gene groups. The average number of PE reads per sample mapping to this gene group subset of the transcriptome was 78.55%, indicating the majority of read data was captured within this subset.

**Figure 3.1. Workflow diagram.** Schematic presenting an overview of the pipeline for sample selection, RNA-Seq output and the bioinformatics analysis and data processing used in this study.
Table 3.1. Transcriptome assembly metrics and results from the annotation pipeline.

<table>
<thead>
<tr>
<th>Transcriptome assembly output</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total non-redundant transcripts</td>
<td>130,677</td>
</tr>
<tr>
<td>Total consensus length (nt)</td>
<td>122,771,252</td>
</tr>
<tr>
<td>Largest transcript</td>
<td>37,222</td>
</tr>
<tr>
<td>Mean transcript length (nt)</td>
<td>939.50</td>
</tr>
<tr>
<td>N50 of transcripts</td>
<td>1,881</td>
</tr>
<tr>
<td>Overall mapping (%)</td>
<td>96.63</td>
</tr>
<tr>
<td>Concordantly mapping pairs (%)</td>
<td>89.26</td>
</tr>
<tr>
<td>CEGMA pipeline complete (%)</td>
<td>97.98</td>
</tr>
<tr>
<td>CEGMA pipeline partial (%)</td>
<td>99.19</td>
</tr>
<tr>
<td>Number predicted proteins</td>
<td>38,694</td>
</tr>
<tr>
<td>Longest ORF</td>
<td>12,406</td>
</tr>
<tr>
<td>Mean ORF length</td>
<td>306.39</td>
</tr>
<tr>
<td><strong>Number transcripts containing predicted proteins</strong></td>
<td><strong>32,960</strong></td>
</tr>
<tr>
<td>Number of unique gene groups</td>
<td>18,559</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Annotation of transcripts containing predicted proteins</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number sequences with BLASTx hits (nr)</td>
<td>22,759 (69.05%)</td>
</tr>
<tr>
<td>Sequences with GO annotations</td>
<td>19,489 (59.12%)</td>
</tr>
<tr>
<td>Number of sequences with KOG hits</td>
<td>22,189 (67.32%)</td>
</tr>
<tr>
<td>Number of sequences with KEGG hits</td>
<td>22,383 (67.90%)</td>
</tr>
</tbody>
</table>

3.4.2 Differential expression and GO term enrichment analysis

Of the 18,559 gene groups identified in the transcriptome assembly, 404 were determined significantly differentially expressed; 180 were over-expressed in high-HPV full-sib groups while 224 were under-expressed. A subset of 91 genes from the 404 were unable to be annotated and were considered novel. A number of immune-related genes comprised the remaining 313 DEGs that could be functionally annotated. Hierarchical cluster analysis of the expression profiles of all the DEGs show shrimp with high and low-HPV level separating into two distinct groups. However, within these distinct groups, samples clustered closely but not strictly within their family assignment (Figure 3.2). A number of genes in a wide array of processes reported to be involved in the immune response of crustaceans including pattern recognition proteins, antimicrobial peptides (AMPs), the prophenoloxidase (proPO) activating system, signalling and antiviral activity were found to be differentially expressed between the high and low-HPV shrimp (Table 3.2). The GO term enrichment analysis (Table 3.3) revealed an enrichment of GO categories involving metabolic processes and oxidoreductase and catalytic activity in genes found to be under-expressed in the high-HPV shrimp. Transcription factor activity was enriched in the GO terms associated with genes under-expressed in high-HPV shrimp and a number of transcripts coding for reverse transcriptase and transposase were over-expressed in these animals. Components of the RNAi pathway along with zinc finger cch-type antiviral proteins were over-expressed in the high-HPV shrimp. A HPV capsid protein was found to exhibit the greatest fold change difference among all DEGs supporting the differences in infection levels between the groups (Figure 3.3). Interestingly, several genes related to growth and reproductive maturation were identified as under-expressed in the high-HPV shrimp. Supplementary Figure 3.1 shows validation of gene expression using RT-qPCR.
3.4.3 Identification of sequence variants within DEGs

A large number of sequence variants, including single nucleotide polymorphisms (SNPs), insertions and deletions, were identified within the 404 DEGs. After filtering of variants that were common across the two conditions, 4,240 of these were observed to be present exclusively in either the high or low-HPV full-sib groups (Figure 3.4). Many of the DEGs with an identity associated with immune response contained several of these variants. A total of 535 homozygous SNPs were detected in transcripts coding for genes such as crustin, C-type lectin and argonaut2.
Table 3.2. Differentially expressed genes associated with immune response identified when comparing banana shrimp with differing HPV levels. A positive fold change value signifies over-expression in high-HPV shrimp.

<table>
<thead>
<tr>
<th>Identity</th>
<th>Accession</th>
<th>Hit species</th>
<th>E-value</th>
<th>Similarity (%)</th>
<th>Contig ID</th>
<th>FDR</th>
<th>FC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Transcription factors</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transcription factor ATF-b</td>
<td>AHY82567.1</td>
<td>Litopenaeus vannamei</td>
<td>2.19E-157</td>
<td>76</td>
<td>c39964_g1</td>
<td>3.11E-02</td>
<td>-1.97</td>
</tr>
<tr>
<td>c-Jun protein</td>
<td>AIB28949.1</td>
<td>L. vannamei</td>
<td>3.92E-159</td>
<td>99</td>
<td>c49654_g1</td>
<td>2.81E-02</td>
<td>-1.55</td>
</tr>
<tr>
<td>Transcription factor kayak (c-fos homolog)</td>
<td>XP_006564216.1</td>
<td>Apis mellifera</td>
<td>4.59E-7</td>
<td>63</td>
<td>c52223_g1</td>
<td>4.42E-03</td>
<td>-1.58</td>
</tr>
<tr>
<td><strong>PRPs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c-type lectin 3</td>
<td>AGZ95687.1</td>
<td>Palaeon modestus</td>
<td>7.19E-22</td>
<td>56</td>
<td>c54472_g1</td>
<td>5.94E-03</td>
<td>-2.04</td>
</tr>
<tr>
<td>c-type lectin 4</td>
<td>AGS42195.1</td>
<td>Fenneropenaeus</td>
<td>3.49E-13</td>
<td>47</td>
<td>c40858_g1</td>
<td>3.85E-02</td>
<td>-1.93</td>
</tr>
<tr>
<td>c-type lectin 1</td>
<td>AE092001.1</td>
<td>Scylla paramamosain</td>
<td>1.62E-25</td>
<td>57</td>
<td>c48186_g1</td>
<td>1.26E-03</td>
<td>-1.87</td>
</tr>
<tr>
<td>c-type lectin 2</td>
<td>AF59946.1</td>
<td>Marsupenaeus japonicus</td>
<td>2.15E-158</td>
<td>89</td>
<td>c51298_g1</td>
<td>3.81E-02</td>
<td>-1.74</td>
</tr>
<tr>
<td>TLR 3</td>
<td>AEK86517.1</td>
<td>L. vannamei</td>
<td>0.00</td>
<td>98</td>
<td>c49171_g1</td>
<td>2.04E-02</td>
<td>-1.62</td>
</tr>
<tr>
<td>Techylectin-5B (fibrinogen-related domain)</td>
<td>KDR18776.1</td>
<td>Zooteormopsis nevadensis</td>
<td>2.07E-77</td>
<td>68</td>
<td>c50158_g2</td>
<td>3.43E-02</td>
<td>-1.87</td>
</tr>
<tr>
<td><strong>AMPS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anti-lipopolysaccharide factor</td>
<td>ADE27980.1</td>
<td>Fenneropenaeus indicus</td>
<td>5.06E-49</td>
<td>96</td>
<td>c49309_g2</td>
<td>4.46E-03</td>
<td>-1.99</td>
</tr>
<tr>
<td>Penaedin</td>
<td>AAV85945.1</td>
<td>Fenneropenaeus chinensis</td>
<td>8.4E-28</td>
<td>82</td>
<td>c47796_g1</td>
<td>9.91E-03</td>
<td>-1.81</td>
</tr>
<tr>
<td>Crustin 1</td>
<td>ACZ43781.1</td>
<td>F. chinensis</td>
<td>8.19E-52</td>
<td>95</td>
<td>c44685_g1</td>
<td>5.33E-03</td>
<td>-1.75</td>
</tr>
<tr>
<td>Crustin 2</td>
<td>ACZ43782.1</td>
<td>F. chinensis</td>
<td>9.42E-69</td>
<td>97</td>
<td>c46040_g1</td>
<td>4.66E-02</td>
<td>-1.72</td>
</tr>
<tr>
<td>Hemocyte lysozyme</td>
<td>ABO33316.1</td>
<td>Peneus monodon</td>
<td>6.59E-95</td>
<td>98</td>
<td>c45236_g2</td>
<td>2.46E-03</td>
<td>-2.26</td>
</tr>
<tr>
<td>Invertebrate-type lysozyme</td>
<td>AET36867.1</td>
<td>L. vannamei</td>
<td>2.53E-47</td>
<td>74</td>
<td>c9050_g1</td>
<td>5.52E-04</td>
<td>2.62</td>
</tr>
<tr>
<td><strong>ProPO/clotting system</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hemolymph clottable protein</td>
<td>Q9US72.1</td>
<td>P. monodon</td>
<td>0.00</td>
<td>94</td>
<td>c54022_g2</td>
<td>1.13E-03</td>
<td>-1.89</td>
</tr>
<tr>
<td>Hemocyte kazal-type proteinase inhibitor</td>
<td>ABC33915.1</td>
<td>F. chinensis</td>
<td>2.02E-58</td>
<td>94</td>
<td>c49067_g2</td>
<td>2.33E-07</td>
<td>-3.18</td>
</tr>
<tr>
<td>Hemoycanin</td>
<td>ABV58636.1</td>
<td>Metapenaeus ensis</td>
<td>8.02E-102</td>
<td>95</td>
<td>c43304_g2</td>
<td>4.57E-04</td>
<td>-2.66</td>
</tr>
<tr>
<td>Laccase 1</td>
<td>EH67706.1</td>
<td>Danus plexippus</td>
<td>5.78E-163</td>
<td>60</td>
<td>c53467_g1</td>
<td>2.75E-02</td>
<td>-1.88</td>
</tr>
<tr>
<td>prophenoloxidase</td>
<td>AGT80129.1</td>
<td>F. merguiensis</td>
<td>0.00</td>
<td>99</td>
<td>c45188_g1</td>
<td>3.79E-02</td>
<td>-1.62</td>
</tr>
<tr>
<td>Clip-domain serine proteinase 3</td>
<td>AFA42361.1</td>
<td>Portunus trituberculatus</td>
<td>1.66E-100</td>
<td>58</td>
<td>c50672_g3</td>
<td>4.20E-02</td>
<td>-1.53</td>
</tr>
<tr>
<td><strong>Antioxidant enzymes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thioredoxin</td>
<td>ACX30746.1</td>
<td>F. chinensis</td>
<td>1.33E-64</td>
<td>96</td>
<td>c42023_g1</td>
<td>1.55E-02</td>
<td>-1.94</td>
</tr>
<tr>
<td>Glutathione peroxidase</td>
<td>ACB42236.1</td>
<td>Metapenaeus ensis</td>
<td>1.84E-62</td>
<td>73</td>
<td>c45680_g1</td>
<td>1.19E-03</td>
<td>-1.78</td>
</tr>
<tr>
<td>Cytosolic manganese superoxide dismutase</td>
<td>ACK44125.1</td>
<td>P. monodon</td>
<td>1.24E-102</td>
<td>99</td>
<td>c44749_g2</td>
<td>1.04E-02</td>
<td>-1.43</td>
</tr>
<tr>
<td>Selenoprotein 1</td>
<td>NP_001177306.1</td>
<td>Ciona intestinalis</td>
<td>7.99E-39</td>
<td>50</td>
<td>c50622_g1</td>
<td>4.96E-02</td>
<td>-1.32</td>
</tr>
<tr>
<td><strong>RNAi pathway</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dicer 2</td>
<td>AGL08684.1</td>
<td>P. monodon</td>
<td>0.00</td>
<td>97</td>
<td>c52190_g2</td>
<td>3.22E-02</td>
<td>1.52</td>
</tr>
<tr>
<td>Argonaut 2</td>
<td>ADB25181.1</td>
<td>L. vannamei</td>
<td>0.00</td>
<td>85</td>
<td>c55174_g2</td>
<td>1.40E-02</td>
<td>1.51</td>
</tr>
<tr>
<td>Arsenite-resistance protein 2-like (Ars-2)</td>
<td>AEI38215.1</td>
<td>L. vannamei</td>
<td>0.00</td>
<td>99</td>
<td>c51655_g1</td>
<td>4.87E-02</td>
<td>1.21</td>
</tr>
<tr>
<td><strong>JAK/STAT pathway</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein inhibitor of activated STAT</td>
<td>AHF29321.1</td>
<td>Scylla paramamosain</td>
<td>5.16E-59</td>
<td>56</td>
<td>c55150_g2</td>
<td>1.62E-02</td>
<td>1.38</td>
</tr>
<tr>
<td>Phosphoinoside 3-kinase PI3K</td>
<td>ADE44091.1</td>
<td>Panulirus argus</td>
<td>0.00</td>
<td>92</td>
<td>c55177_g2</td>
<td>3.79E-02</td>
<td>1.21</td>
</tr>
<tr>
<td>Cytokine receptor domeless</td>
<td>AY46351.1</td>
<td>L. vannamei</td>
<td>7.39E-46</td>
<td>41</td>
<td>c55085_g1</td>
<td>2.75E-02</td>
<td>1.39</td>
</tr>
<tr>
<td>Ras GTP exchange factor, son of sevenless</td>
<td>XP_002428152.1</td>
<td>Pediculus humanus corporis</td>
<td>0.00</td>
<td>76</td>
<td>c36640_g2</td>
<td>1.26E-03</td>
<td>-1.88</td>
</tr>
</tbody>
</table>

**FC** fold change; **FDR** false discovery rate adjusted **P**-value.
Table 3.3. GO term enrichment for the subset of 404 DEGs against all sequences with GO annotations.

<table>
<thead>
<tr>
<th>GO-ID</th>
<th>Term</th>
<th>Category*</th>
<th>FDR</th>
<th>Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0006091</td>
<td>generation of precursor metabolites and energy</td>
<td>P</td>
<td>3.54E-07</td>
<td>OVER</td>
</tr>
<tr>
<td>GO:0003735</td>
<td>structural constituent of ribosome</td>
<td>F</td>
<td>6.88E-07</td>
<td>OVER</td>
</tr>
<tr>
<td>GO:0016491</td>
<td>oxidoreductase activity</td>
<td>F</td>
<td>1.09E-06</td>
<td>OVER</td>
</tr>
<tr>
<td>GO:0005198</td>
<td>structural molecule activity</td>
<td>F</td>
<td>3.59E-04</td>
<td>OVER</td>
</tr>
<tr>
<td>GO:1901576</td>
<td>organic substance biosynthetic process</td>
<td>P</td>
<td>4.55E-03</td>
<td>OVER</td>
</tr>
<tr>
<td>GO:0005975</td>
<td>carbohydrate metabolic process</td>
<td>P</td>
<td>4.55E-03</td>
<td>OVER</td>
</tr>
<tr>
<td>GO:0003824</td>
<td>catalytic activity</td>
<td>F</td>
<td>3.27E-02</td>
<td>OVER</td>
</tr>
<tr>
<td>GO:0001071</td>
<td>nucleic acid binding transcription factor activity</td>
<td>F</td>
<td>4.98E-02</td>
<td>OVER</td>
</tr>
<tr>
<td>GO:0016746</td>
<td>transferase activity, transferring acyl groups</td>
<td>F</td>
<td>4.98E-02</td>
<td>OVER</td>
</tr>
<tr>
<td>GO:0043167</td>
<td>ion binding</td>
<td>F</td>
<td>2.72E-03</td>
<td>OVER</td>
</tr>
<tr>
<td>GO:1901360</td>
<td>organic cyclic compound metabolic process</td>
<td>P</td>
<td>4.55E-03</td>
<td>UNDER</td>
</tr>
<tr>
<td>GO:0090304</td>
<td>nucleic acid metabolic process</td>
<td>P</td>
<td>2.34E-02</td>
<td>UNDER</td>
</tr>
</tbody>
</table>

*F; molecular function, P; biological process.

Figure 3.3. Fold change and annotation results of additional DEGs implicated in immune related functions and also related to growth and reproduction. A positive fold change indicates over-expression and a negative fold change indicates under-expression in high-HPV shrimp.
Figure 3.4. SNP analysis. The distribution of the 4,240 SNP and indel types identified among the 404 DEGs that were determined to be exclusive to either the high- or low-HPV full-sib groups in this study.

3.5 Discussion

3.5.1 Transcriptome assembly and annotation

The \textit{F. merguiensis} hepatopancreas transcriptome assembly resulted in almost 90% of all clean reads mapping concordantly in pairs and above 99% completeness with the CEGMA pipeline core gene set. This result, coupled with the high sequencing depth, provides us with confidence in the completeness and quality of the transcriptome assembly. The total number of predicted proteins annotated with a BLAST result from this assembly was 22,759 compared with 20,430 from an earlier report (Powell et al. 2015), likely due to the rate of expansion of the NCBI nr database from the previous year.

3.5.2 Differential gene expression

The rationale for selecting high and low-HPV full-sib groups, rather than representative high and low animals, is that the full-sib average could be a better estimation of the genetic value. Sibs can be considered genetic replicates, therefore, by working with family groups, one may get clearer results than without the pedigree information. Hierarchical clustering of expression within the DEGs among individual samples separated high and low-HPV shrimp into two clear clusters based on this condition. Though, animals were observed to cluster closely but not strictly within their respective full-sib groups, which shows genetic factors are influencing these patterns of expression.

A clear bias in the expression of genes involved in important immune-related mechanisms is evident between the two groups of animals in this study. Genes comprising the more notable immune response mechanisms were predominately under-expressed in high-HPV shrimp. Several genes encoding antioxidant enzymes, including glutathione peroxidase (GPx), selenoprotein L, cytosolic manganese superoxide dismutase (MnSOD), were found to be under-expressed in high-HPV shrimp. Reactive oxygen species (ROS) are an important part of the crustacean immune defence system that are produced to help eliminate invading microbes
Viruses are known to inhibit the synthesis of antioxidant enzymes and some, such as MnSOD, have been linked to anti-WSSV responses in crustaceans (Mohankumar and Ramasamy 2006; Liu et al. 2009). Furthermore, the tissue antioxidant defence system of the black tiger shrimp *P. monodon* has been reported to operate at a lower rate under WSSV infection (Mathew et al. 2007). The under-expression of these antioxidant enzymes in high-HPV shrimp may have resulted in the higher concentration of virus in these animals.

Responses to oxidative stress are known to be governed by redox-sensitive transcriptional factors which include AP-1 (Wang et al. 2012b). AP-1 has been shown to play an important role in stimulating the transcription of immune effector molecules, such as AMPs and in the cellular response to a broad range of stimuli, including cytokines, growth factors and microbial infection (Li et al. 2015). A putative cis-regulator element for AP-1 was found upstream of the anti-lipopolysaccharide factor (ALF) genes in *P. monodon* (Tharntada et al. 2008) and in the white leg shrimp *L. vannamei*, components of the AP-1 transcription complex, c-Fos and c-Jun, were seen interacting to activate AMPs, such as penaeidin (Li et al. 2015). Additionally, the c-Fos and c-Jun AP-1 complex has been observed to be down-regulated after posttranslational sumoylation by certain protein inhibitor of activated STAT (PIAS) proteins in mammalian cells (Bossis et al. 2005; Tempe et al. 2013). If this interaction extends to crustaceans, the over-expression of the PIAS protein observed in the high-HPV shrimp may somewhat account for the accompanying reduction in expression of AP-1 activated immune effectors, such as ALF and penaeidin.

AMPs are small molecules that exhibit potent antimicrobial activity against a wide range of microorganisms, including bacteria, viruses and fungi (Tassanakajon et al. 2013). Four of the most common AMP family members described in shrimp, the crustins, penaeidins, lysozyme and ALF, were each found to be differentially expressed between high and low-HPV shrimp. All four of these AMP families were under-expressed in high-HPV shrimp with only one invertebrate-type lysozyme found over-expressed in these animals. Whilst numerous studies have described the antibacterial activities of AMPs in crustaceans, there are a growing number of reports of an antiviral role for some of these molecules. Recent studies have demonstrated that the AMPs crustin-3 and ALF were upregulated in WSSV challenged *P. monodon* (Antony et al. 2011) and that ALF can interfere with WSSV propagation in this species and also in the Crayfish *P. leniusculus* (Liu et al. 2006; Tharntada et al. 2009). A lysozyme from the blue shrimp *L. stylirostris* was found upregulated in WSSV-infected animals and treatment with the lysozyme protein was effective at blocking WSSV infection (Mai and Wang 2010). AMPs are known to be induced by several pattern recognition proteins (PRPs) which are a group of proteins that can recognise pathogens by binding to certain molecular patterns such as polysaccharides, glycoproteins and also polynucleotides associated with bacteria and viruses (Wang and Wang 2013). It is widely believed that this recognition process activates the host defence system leading to rapid humoral and cellular immune responses (Sun et al. 2008; Wang and Wang 2013). Several groups of PRPs have been identified in invertebrates including lipopolysaccharide and β-1,3-glucan binding proteins (LGBP), C-type lectins (CTLs), fibrinogen-related proteins (FREPs) and toll-like receptors (TLRs) (Wang and Wang 2013). CTLs are a large
family of proteins considerably diverse among invertebrates that have been demonstrated to couple pathogen-associated molecular pattern (PAMP) recognition to the initiation of several innate response actions such as induction of the proPO system and respiratory burst (Robinson et al. 2006; Wang et al. 2011; Wang and Wang 2012). One particular CTL was shown to modulate the expression of antimicrobial peptides including ALF, lysozyme, penaeidin and crustin in the kuruma shrimp, *M. japonicas* (Wang et al. 2014), whereas other reports describe the involvement of CTLs in the viral resistance of shrimp (Junkunlo et al. 2012; Wang and Wang 2012). TLRs are transmembrane proteins that can mediate the recognition of PAMPs (Aderem and Ulevitch 2000). In mammals, transcription factor AP-1 is also known to be activated downstream of TLR signalling (O’Neill et al. 2013). Recent studies report TLRs from *L. vannamei* and *P. monodon* were upregulated in WSSV challenged shrimp (Wang et al. 2012a; Deepika et al. 2014). Five CTLs, a FREP and a toll-like receptor were identified as differentially expressed in this study. Interestingly, all of these PRPs were under-expressed in high-HPV shrimp, which can also be considered over-expressed in shrimp exhibiting lower levels of HPV infection.

CTLs, along with other PRPs such as LGBPs, are known activators of melanin synthesis via the proPO cascade. The proPO system results in the generation of cytotoxic products and in cellular melanotic encapsulation which are some of the most effective mechanisms of invertebrate cellular defence (Amparyup et al. 2013). Proteinases and proteinase inhibitors have well-characterised roles in the immune function of invertebrates (Kong et al. 2009; Li et al. 2012). Clip-domain serine proteinases have been implicated in the proPO activation cascades as an activating enzyme converting proPO into the key enzyme phenoloxidase (Amparyup et al. 2013). Kazal-type SPIs have been proposed as promising candidates for limiting proteolytic activity in coagulation and proPO activation (Jiménez-Vega and Vargas-Albores 2005; Kong et al. 2009). Recently, a five-domain Kazal-type SPI from *P. monodon* has been shown to be up-regulated upon infection with WSSV and consequently implicated in antiviral response (Donpudsa et al. 2010). These proteinases and proteinase inhibitors each display modulated expression, adding support for a role in response to viral infection in this species. Previous studies in several species of shrimp have reported changes in levels of expression in proPO-related genes in response to viral infection, many of which report a downregulation in response to the pathogen (Roux et al. 2002; Ai et al. 2008; Yeh et al. 2009). A recent study has demonstrated that melanisation has an antiviral role in *P. monodon* infected with WSSV (Sutthangkul et al. 2015). In crustaceans, hemocyanin is produced by the hepatopancreas and contained in the plasma and can be converted by hemocyte components to a phenoloxidase-like enzyme (Adachi et al. 2003; Cerenius et al. 2008). Hemocyanin has also been shown to function as an antimicrobial peptide and has displayed non-specific antiviral properties (Zhang et al. 2004; Liu et al. 2009). All DEGs associated with the proPO system were identified as under-expressed in the high-HPV shrimp. The downregulation of these genes supports previous findings suggesting lower activity of the proPO system may promote higher levels of viral infection.

Important progress has been made in the understanding of invertebrate antiviral responses in recent years. The RNA interference (RNAi) pathway is one such response that can be triggered by dsRNA produced by
replicating viruses which are processed into shorter fragments by the endonuclease Dicer and incorporated into the RNA-induced silencing complex (RISC) to be degraded (Liu et al. 2009). A report by La Fauce and Owens (La Fauce and Owens 2009), using an HPV strain isolated from *F. merguiensis*, demonstrated a reduction in HPV titre in a cricket model after injection of dsRNA specific for the virus capsid protein. The Argonaout protein family are key components of the RISC complex (Labreuche and War 2013) and Ars2 has been shown to interact with Dicer2 in *L. vannamei*, suggesting it may be involved in the regulation of this pathway in shrimp (Chen et al. 2012). The over-expression of Dicer2, Argonaout2 and Ars2 in the high-HPV shrimp suggests this pathway is likely upregulated in response to the higher concentration of virus.

Increasing evidence suggests JAK/STAT signalling pathways are involved in the antiviral response of shrimp (Chen et al. 2008). Knockdown of a cytokine receptor analogous to Domeless in *L. vannamei* resulted in a lower mortality and in less WSSV copies, leading to the suggestion that this receptor may be hijacked by WSSV to benefit virus replication (Yan et al. 2015). Another study detected an increase in dicer2 and argonaout2 when knocking down a STAT in WSSV infected *L. vannamei* (Wen et al. 2014). Although, the STAT gene was not identified as differentially expressed in this study, over-expression of PIAS transcripts, that are known to deactivate STAT, were observed in high-HPV shrimp. This may be exerting some influence on STAT regulation and may be assisting the elevated expression of Dicer2 and Argonaout2.

Taken together, these results indicate shrimp from high-HPV full-sib groups appear to have a somewhat reduced presence of these important immune response gene transcripts, yet possess upregulated putative antiviral pathways. Given the association of PRPs with the activation of the immune mechanisms described here (Figure 3.5), the reason behind this could lie in the increased expression of the PRP repertoire in shrimp with lower viral load, enabling them to be resistant to HPV infection via a more sensitive activation of humoral defence. If these gene products have the capability to bind virus particles, as suggested in the literature (Sritunyalucksana et al. 2013), it is tempting to then hypothesise that shrimp with a propensity for increased expression of these molecular pattern binding proteins may fair better when exposed not only to HPV infection, but to infection from other recognisable pathogens. Alternatively, as many of these factors associate with hemocytes, an inherent decrease in hematopoiesis in the high-HPV shrimp could also explain the downregulation of some of these mechanisms. As circulating hemocytes were no doubt contained within the hepatopancreas tissue prior to sequencing, further study to investigate the tissue specificity of these factors would be useful. The presence of a secondary bacterial infection was also considered as the cause of this increased immune response in the low-HPV animals, however, the absence of a detectable signal from these samples by PCR using universal bacterial 16S rRNA primers indicated to the contrary (data not shown).

There is some evidence of a reduction in expression of genes related to growth and reproductive maturation in the high-HPV shrimp, such as an insulin-like growth factor-binding protein, which has been implicated in reproductive regulation (Chandler et al. 2015), an ovary development-related protein and an estrogen sulfotranferase, as listed in Figure 3. A reduced growth rate of juveniles is the primary symptom of acute HPV infection in penaeid shrimp (Safeena et al. 2012) and as fecundity has also been shown to increase with size.
(Hoang et al. 2002), inhibition of growth would likely result in reduced reproductive activity. This observation may point to evidence of the molecular pathogenesis of HPV infection with regard to reduced growth and maturation.

3.5.3 Identification of sequence variants

Molecular markers that associate with traits for disease resistance can become powerful tools for selective breeding programs. Some recent studies have demonstrated the usefulness of SNPs linked with gene expression data to help comprehend mechanisms determining complex traits (Santos et al. 2014). For example, a set of transcribed SNPs were used to identify loci associated with resistance to WSSV in challenged *P. monodon*, some of which mapped closely to candidate genes involved in immune response (Robinson et al. 2014). Here we report over 4,000 sequence variants localising to genes identified as differentially expressed and occurring uniquely in either low or high-HPV full-sib groups. Over 500 of these sequence variants were identified as homozygous, wherein only the variant different from the reference was detected in animals particular to low or high-HPV load, indicating all animals from each of the 3 full-sib groups share this variant exclusively. These homozygous SNPs were identified in varying abundance within transcripts coding for immune-associated genes such as C-type lectins, crustin and argonaut2. The positioning of SNPs within these genes include open reading frames and 3’ untranslated regions. Whether the presence of such variants has any effect on the function or regulation of the resulting gene products beckons additional investigation. These

---

**Figure 3.5. Putative immune response pathway.** *a* pathways associated with invertebrate immune response represented by genes over-expressed in the high-HPV animals. *b* the putative interrelationship between some of the important immune mechanisms according to reports in the literature that are represented by genes under-expressed in the high-HPV animals.
data offer an exceptional resource for mining gene associated markers and further research is anticipated that will explore the effectiveness of these sequence variants for use as functional markers of HPV resistance.

3.5.4 Conclusions

To our knowledge, this is the first report of a differential expression analysis using RNA-Seq to explore gene expression with regard to viral load among genetic families of cultured shrimp. A total of 404 DEGs were identified, 180 being over-expressed in high-HPV full-sib groups where 224 were under-expressed. Of these DEGs, 313 could be functionally annotated revealing a large proportion having associations with processes reported to be involved in the immune response of crustaceans. These processes include pattern recognition proteins, AMPs, the proPO activating system, signalling and antiviral activity. The differential expression analysis indicates shrimp from high-HPV families appear to have lower expression of some of these important immune response elements, yet possess upregulated putative antiviral pathways. Hierarchical clustering of expression values suggests some degree of genetic influence may be driving these differences. Sequence variants that were unique to either low or high-HPV animals were identified among these DEGs, offering an exceptional resource for discovering gene-associated SNPs for use as functional markers of HPV resistance.

3.6 Acknowledgements

We thank Ms Jane Quinn, Mr Courtney Remilton, Dr Josephine Nocillado, Mr Thoa Ngo and Ms Nicole Ertl for their contributions to sample collection and processing.
3.7 References


Aquaculture 246:139–149.


Li Q, Cui Z, Liu Y, Wang S, Song C (2012) Three clip domain serine proteases (cSPs) and one clip domain serine protease homologue (cSPH) identified from haemocytes and eyestalk cDNA libraries of swimming crab *Portunus trituberculatus*. Fish Shellfish Immunol 32:565–571.


complex occurs on target promoters to limit transcriptional activation. Oncogene 33:921–7.


S3.1 Supplementary Methods

S3.1.1 Functional annotation

Software BLAST 2.2.30+ was used for alignment against a local copy of the NCBI non-redundant (nr) protein database downloaded on the 13\textsuperscript{th} of March 2015. Blast2GO version 2.8 with database b2g_jan15 was used for gene ontology (GO) annotation and InterproScan analysis (Conesa et al. 2005). The GO terms assigned by Blast2GO were subject to GO term enrichment. Gene Ontology enrichment of the DEGs was calculated using the GOSSIP algorithm within the Blast2GO package based on the frequency of GO terms associated with DEG list relative to total predicted protein background with an FDR correction for multiple testing (FDR < 0.05). Protein sequences were aligned against the NCBI KOG database (version 2/2/2011) using RPSBLAST v2.2.15 and the KEGG genes database (version 2/12/2012) using BLASTALL v2.2.15 both using an E value cut off of 1x10\textsuperscript{-5} and performed via the WebMGA server (Wu et al. 2011) (http://weizhonglab.ucsd.edu/metagenomic-analysis/).

S3.1.2 RT-qPCR

Three biological replicates were chosen from each condition (n=6) from total RNA extracted as mentioned above. Input total RNA was normalised using the QuantiFluor\textsuperscript{®} dsDNA System and Quantus\textsuperscript{®} fluorometer (Promega) and cDNA was produced using the Quantitect\textsuperscript{®} Reverse Transcription kit (Qiagen) according to the manufacturer’s instructions incorporating the genomic DNA elimination reaction. All qPCR reactions (15 μl) were performed in duplicate using the Quantifast\textsuperscript{®} Probe PCR kit (Qiagen) with 20 μM forward and reverse primers and 10 μM probe, then analysed using the Corbett Rotorgene 6000 real-time PCR system (Corbett Life Science). Custom qPCR probes were labelled with a FAM 5’ reporter and ZEN / Iowa Black FQ quenchers (Integrated DNA Technologies). Probe sets were designed using the PrimerQuest\textsuperscript{®} Design Tool (Integrated DNA Technologies) and efficacy was determined by 10-fold serial dilution standard curve analysis (probe sequences are detailed in Additional file 2). Cycling conditions were: activation at 95°C for 3 min, followed by 40 repeats of denaturation at 95°C for 5s and annealing/extension at 60°C for 30s. Threshold cycle (C\textsubscript{T}) values reported by the Rotorgene 6000 real-time PCR system were normalized to the reference gene EF1-α and converted to relative log-fold differences between low and high-HPV shrimp. The expression of each gene was analysed using the comparative C\textsubscript{T} (ΔΔC\textsubscript{T}) method.

S3.2 Supplementary Results

S3.2.1 RT-qPCR validation

The expression of transcripts c49067_g2; kazal-type serine proteinase inhibitor (SPI), c43304_g2; hemocyanin and c55174_g2; Dicer2, were measured using RT-qPCR and analysed with the ΔΔC\textsubscript{T} method normalised to endogenous reference gene, elongation factor 1-alpha (EF-1α) (Supplementary Figure 1). The kazal-type SPI and hemocyanin genes were determined to be under-expressed in high HPV individuals whilst Dicer2 was over-expressed, in agreement with the RNA-Seq analysis.
Supplementary Figure 3.1. Comparison of the estimated RNA-Seq fold change values with RT-qPCR. Relative fold change data is normalised to reference gene EF-1α. Relative log fold change between high and low-HPV shrimp was calculated for the qPCR data using the ΔΔCt method.

Supplementary Figure 3.2. Mean group Log HPV copy number per microgram of DNA. Material extracted from hepatopancreas tissue of infected *F. merguiensis* (Figure modified from Knibb et al, 2015). Full-sib groups with too few members for adequate replication (such as 26, 24, 23 and 22) were not considered for inclusion in this study.
Supplementary Table 3.1. Details of hepatopancreas tissue samples including HPV concentration, sequencing yield after quality trimming and percentage of clean reads mapping to the unique gene groups.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Family ID</th>
<th>Sex</th>
<th>Size g</th>
<th>HPV designation</th>
<th>HPV gene copies / ug</th>
<th>Number clean PE reads</th>
<th>Mapping*</th>
</tr>
</thead>
<tbody>
<tr>
<td>448</td>
<td>1</td>
<td>M</td>
<td>17</td>
<td>High</td>
<td>8.28E+08</td>
<td>21,677,501</td>
<td>78.36</td>
</tr>
<tr>
<td>381</td>
<td>1</td>
<td>F</td>
<td>10.1</td>
<td>High</td>
<td>3.54E+08</td>
<td>20,063,977</td>
<td>78.03</td>
</tr>
<tr>
<td>322</td>
<td>1</td>
<td>M</td>
<td>18.6</td>
<td>High</td>
<td>1.35E+08</td>
<td>17,359,321</td>
<td>79.17</td>
</tr>
<tr>
<td>608</td>
<td>1</td>
<td>M</td>
<td>17.8</td>
<td>High</td>
<td>1.33E+07</td>
<td>20,825,830</td>
<td>79.31</td>
</tr>
<tr>
<td>397</td>
<td>2</td>
<td>F</td>
<td>10.9</td>
<td>High</td>
<td>2.29E+08</td>
<td>17,978,947</td>
<td>78.46</td>
</tr>
<tr>
<td>88</td>
<td>2</td>
<td>M</td>
<td>9.6</td>
<td>High</td>
<td>1.18E+08</td>
<td>18,724,010</td>
<td>77.77</td>
</tr>
<tr>
<td>206</td>
<td>2</td>
<td>M</td>
<td>16.6</td>
<td>High</td>
<td>2.16E+07</td>
<td>18,986,204</td>
<td>78.52</td>
</tr>
<tr>
<td>402</td>
<td>2</td>
<td>F</td>
<td>14</td>
<td>High</td>
<td>1.35E+07</td>
<td>19,781,450</td>
<td>77.62</td>
</tr>
<tr>
<td>168</td>
<td>3</td>
<td>F</td>
<td>9.1</td>
<td>High</td>
<td>2.57E+08</td>
<td>18,967,527</td>
<td>78.03</td>
</tr>
<tr>
<td>855</td>
<td>3</td>
<td>M</td>
<td>18.4</td>
<td>High</td>
<td>2.25E+08</td>
<td>17,660,793</td>
<td>77.54</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
<td>F</td>
<td>8.3</td>
<td>High</td>
<td>4.08E+07</td>
<td>18,700,846</td>
<td>78.45</td>
</tr>
<tr>
<td>53</td>
<td>3</td>
<td>M</td>
<td>9.3</td>
<td>High</td>
<td>2.35E+07</td>
<td>18,851,633</td>
<td>77.76</td>
</tr>
<tr>
<td>806</td>
<td>21</td>
<td>M</td>
<td>12.4</td>
<td>Low</td>
<td>3.66E+04</td>
<td>20,205,685</td>
<td>78.79</td>
</tr>
<tr>
<td>690</td>
<td>21</td>
<td>F</td>
<td>20.9</td>
<td>Low</td>
<td>1.59E+04</td>
<td>19,634,341</td>
<td>78.79</td>
</tr>
<tr>
<td>678</td>
<td>21</td>
<td>F</td>
<td>24</td>
<td>Low</td>
<td>6.76E+03</td>
<td>18,293,738</td>
<td>80.36</td>
</tr>
<tr>
<td>568</td>
<td>21</td>
<td>F</td>
<td>22.4</td>
<td>Low</td>
<td>3.71E+03</td>
<td>20,341,562</td>
<td>78.64</td>
</tr>
<tr>
<td>445</td>
<td>25</td>
<td>M</td>
<td>9.6</td>
<td>Low</td>
<td>7.29E+04</td>
<td>18,878,325</td>
<td>77.18</td>
</tr>
<tr>
<td>298</td>
<td>25</td>
<td>F</td>
<td>16.2</td>
<td>Low</td>
<td>5.51E+04</td>
<td>19,676,404</td>
<td>78.07</td>
</tr>
<tr>
<td>738</td>
<td>25</td>
<td>M</td>
<td>24.2</td>
<td>Low</td>
<td>9.50E+03</td>
<td>19,807,545</td>
<td>79.03</td>
</tr>
<tr>
<td>517</td>
<td>25</td>
<td>M</td>
<td>22.5</td>
<td>Low</td>
<td>1.68E+03</td>
<td>19,612,126</td>
<td>79.84</td>
</tr>
<tr>
<td>768</td>
<td>27</td>
<td>F</td>
<td>12.8</td>
<td>Low</td>
<td>9.76E+02</td>
<td>20,839,876</td>
<td>79.95</td>
</tr>
<tr>
<td>838</td>
<td>27</td>
<td>M</td>
<td>13.8</td>
<td>Low</td>
<td>9.38E+02</td>
<td>21,119,664</td>
<td>78.55</td>
</tr>
<tr>
<td>564</td>
<td>27</td>
<td>F</td>
<td>12.7</td>
<td>Low</td>
<td>6.14E+02</td>
<td>18,714,576</td>
<td>78.82</td>
</tr>
<tr>
<td>792</td>
<td>27</td>
<td>F</td>
<td>18.2</td>
<td>Low</td>
<td>3.83E+02</td>
<td>19,090,880</td>
<td>79.22</td>
</tr>
</tbody>
</table>

Total 465,792,761

*Mapping in pairs to transcripts containing predicted proteins.

S3.3 Supplemental references


Chapter 4: Sex-specific transcript expression in the hepatopancreas of the banana shrimp (*Fenneropenaeus merguiensis*)

Daniel Powell* and Abigail Elizur

Centre for Genetics, Ecology and Physiology, Faculty of Science, Health, Education and Engineering, University of the Sunshine Coast, Maroochydore DC, Queensland 4558, Australia

Submitted to a special issue of *Hydrobiologia* following the presentation at mini-symposium “Case studies on Genomics & Physiology” The Crustacean Society conference, Barcelona, Spain, June 2017.

4.1 Abstract

The banana shrimp, *Fenneropenaeus merguiensis*, is a commercially important species of marine crustacean that, like other penaeids, exhibit a female-superior sexual dimorphism with respect to growth rate, making this species an attractive candidate for all-female (monosex) culture. To gain insight into the molecular mechanisms underpinning this characteristic, a test for differential gene expression was performed to compare digestive gland tissue from 14 male and 14 female banana shrimp (n=28) within the dimorphic growth range. Female banana shrimp samples were found to have an increased expression of transcripts coding for chitin catabolism and binding. A novel transcript was also identified that displayed a distinct female-specific expression. Weighted gene co-expression network analysis (WGCNA) from female samples revealed this novel transcript was highly connected within a gene module containing an abundance of transcripts with homology to transcription factors and genes associated with growth regulation.

4.2 Introduction

Most cultured crustacean species exhibit sexually dimorphic growth patterns (Ventura and Sagi 2012). The commercially important banana shrimp, *Fenneropenaeus merguiensis*, like other penaeid shrimp, exhibit a female-superior sexual dimorphism with respect to growth rate. A study by Moss et al. in the Chinese white shrimp (*Fenneropenaeus chinensis*) revealed that the growth of females diverges males of the same age at a body weight of approximately 10 grams (Moss et al. 2002). To date, a detailed understanding of the mechanisms underpinning this characteristic has yet to be established. Uncovering the molecular mechanisms governing sexual dimorphism can assist with the enhancement of profitability of penaeid shrimp culture by enabling the development of monosex (all-female) populations (Argue et al. 2002) and by contributing to a better understanding of nutrient assimilation and utilisation which may lead to improved feeding strategies or may uncover a trait useful for selective breeding (Andriantahina et al. 2013). In decapods, the digestive gland or hepatopancreas is a major organ that performs many of the important functions of the vertebrate intestine, pancreas and liver concerning the production and secretion of digestive enzymes and the absorption and storage of nutrients for tissue growth and energy consumption (Hu and Leung 2007; Wang et al. 2014).
Furthermore, it is an important organ for the synthesis and metabolism of steroid hormones and vitellogenin that govern development and maturation (Vaca and Alfaro 2000; Powell et al. 2015). Whilst a number of studies have investigated sex-biased gene expression in crustaceans (for examples see Gao et al. 2014; Liu et al. 2015; Huylmans et al. 2016), few have focused on the differences within the same discreet tissue type in both sexes. To explore the differences in tissue-specific gene expression between male and female banana shrimp, a test for differential gene expression was undertaken to compare hepatopancreas tissue from 14 male and 14 female banana shrimp (n=28) with a body weight of 11 grams or larger, at which point a sexually dimorphic growth pattern becomes apparent in other *Fenneropenaeus* species.

### 4.3 Methods

#### 4.3.1 Sample collection and RNA extraction

All samples used in this study were from captive banana shrimp showing no visible sign of disease, bred and grown together and part of a cohort of 1,998 animals collected over a period of 4 days in August 2014 from the same grow-out pond at a commercial farm in Cardwell, North Queensland. All animals sampled appeared to be in intermolt and were of the same age. The hepatopancreas was excised from the animals on-site and stored in RNAlater (Life Technologies) for transport and storage at -80°C. RNA was extracted and purified from 20 mg of tissue taken from a longitudinal section of the caudal region using the RNeasy Plus Universal kit (Qiagen) as per the manufacturer’s instructions. Library preparation using the illumina TruSeq RNA library prep kit and sequencing of the randomly pooled RNA libraries across 4 lanes on the illumina HiSeq2500 platform was outsourced to the Australian Genome Research Facility (AGRF). The data sets are available in the NCBI SRA repository under BioSample accession numbers SAMN04270865 and SAMN04270866.

#### 4.3.2 Data processing, annotation and differential gene expression analysis

Quality of the raw read data was assessed using FastQC ([http://www.bioinformatics.babraham.ac.uk/projects/fastqc/](http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)). Trimming of adaptors, filtering of low quality reads and removing contaminants were performed using Trimmomatic (Bolger et al. 2014) with a custom screening database.

The reference transcriptome used in this study has been described previously (Powell et al. 2016). A total of 32,960 transcripts were predicted to contain protein-coding regions that were clustered into 18,559 unique components or gene groups as defined by the Trinity pipeline, which we refer to as genes for simplicity.

The reference transcriptome was re-annotated to include the latest submissions using BLAST 2.5.0+ software for alignment with homologous sequences with a local copy of the NCBI non-redundant database (nr) using an E-value cut off of 1e-5. Sequences with homology to proteins of interest were submitted to BLAST manually in order to confirm alignment specificity. Signal peptide prediction was performed using SignalP v4.0 (Petersen et al. 2011), and prediction of a transmembrane helix by TMHMM v2.0 (Krogh et al. 2001). Blast2GO version...
3.1 was used for gene ontology (GO) annotation, enzyme code mapping and InterproScan searches (Conesa et al. 2005). Clean read data from each individual sample was mapped to the gene group transcripts using the align_and_estimate_abundance.pl script from the Trinity v2.0.6 software package using default parameters except for --est_method RSEM --aln_method bowtie2 --trinity_mode. Reads not mapping in pairs were discarded from DE analysis. The output was analysed using the Bioconductor (Huber et al. 2015) package DESeq2 (v1.6.3) (Love et al. 2014) in R. Differential gene expression was determined using a multifactor design incorporating conditions for male or female and for the lane of the HiSeq flow cell from which the samples were sequenced. We considered only the gene groups calculated to have at least a 1.3 fold difference in expression and an FDR corrected \( p \)-value of \(<0.05\) as significant, in line with other reports from non-mammalian studies (Castillo et al. 2015; Koltes et al. 2015; McTaggart et al. 2015). Heatmap was produced using the gplots v3.0.1 package in R.

4.3.3 Construction of weighted gene co-expression network

To identify molecular interactions between genes potentially associated with novel transcripts, we performed a weighted gene co-expression network analysis (WGCNA) using female hepatopancreas RNA-Seq count data (n=14). WGCNA seeks to identify modules of densely interconnected genes with highly correlated expression levels across samples, providing a means to uncover novel connections between genes (Oldham et al. 2009; Sasagawa et al. 2016). Count data was normalised using the variance-stabilizing transformation from the DESeq2 package as described elsewhere (Dobson et al. 2016). A soft-thresholding power of 8 was selected using the scale-free topology criterion determined by the pickSoftThreshold function in the WGCNA package v1.51 (Langfelder and Horvath 2008) in R. The minimum size of the modules were set to 30 with cut height of 0.25 following the parameters outlined in the tutorial found at http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/Tutorials. A signed network was used to report only positively correlated genes. Gene networks were visualised with Cytoscape v3.2.1 (Shannon et al. 2003).

4.4 Results

4.4.1 Sample selection and RNA sequencing

A higher average body weight for females compared to males was previously reported for the full cohort of 1,998 animals drawn upon for this study, indicating sexually dimorphic growth is present in the captive population (Nguyen et al. 2014). The total body weight of sequenced female animals ranged from 12.7 – 24.5 g with an average of 17.92 g and the total body weight of sequenced male animals ranged from 11.7 – 24.2 g with an average of 17.25 g. The combined sequencing output from each of the 28 libraries produced a total of 126.5 Gb of raw data. The average number of paired-end (PE) reads per library was 22.5 million.
### 4.4.2 Sex-specific transcript expression

Of the 18,559 genes tested, only 22 were identified as differentially expressed. Of these differentially expressed genes (DEGs), 7 were found to be under-expressed and 15 over-expressed in the female banana shrimp when compared to male (Figure 4.1). Just 13 out of the 22 DEGs could be annotated when searched using publically available databases. Interestingly, genes associated with the catabolism and binding of chitin, such as chitinase I, beta-n-acetylglucosaminidase and a periophin-44-like, comprised 4 of the 10 annotated genes identified as over-expressed in female shrimp. The 5 annotated transcripts identified as under-expressed in female shrimp included a low-density lipoprotein receptor and others involved in membrane transport.

Hierarchical clustering of each shrimp sample presented in Figure 4.2 shows a clear division of male and female samples into 2 groups with only 1 male sample (male 14) clustering more closely with the female samples, indicating the expression profile of the set of transcripts identified as differentially expressed were highly sex-specific.

**Figure 4.1. Heatmap of differentially expressed genes.** A heatmap presenting each differentially expressed transcript identified below an adjusted P-value threshold of $<0.05$ when comparing female with male banana shrimp of approximately 11 grams and above body weight.

A novel transcript (ID: c40155_g2) was identified as over-expressed in female shrimp and with a strikingly higher fold change (8.2 fold) than any other DEG in this study (Figure 4.2). This novel transcript is predicted to contain a coding sequence of 330 aa that has no homology to any other sequence in the NCBI non-redundant protein database, the UniProt database or among the PFam domain models. This transcript was predicted not to contain a signal peptide, a transmembrane domain or other gene features using an InterproScan search. An
examination of the raw count data confirmed an absence of sequencing reads aligning to this novel transcript for all but 1 male sample (male 3), which revealed an alignment of only 2 reads compared with a total of 939 reads for all female samples combined giving an average of 67 reads per sample.

**Figure 4.2. Fold change of differentially expressed transcripts.** A positive fold change indicates over-expression and a negative fold change indicates under-expression in female banana shrimp.

### 4.4.3 Weighted gene co-expression network analysis

Normalised count data from the 14 female hepatopancreas samples were analysed using WGCNA to identify gene co-expression patterns with a high positive correlation with the novel transcript c40155_g2. A total of 55 modules, excluding the module grey that includes unassigned genes, were identified and are depicted labelled by colours in the hierarchical clustering dendrogram (Figure 4.3). The co-expression module *skyblue* containing transcript c40155_g2 is comprised of 160 gene groups, of which 74 could be annotated. GO annotations of gene groups in the module *skyblue* revealed that protein and DNA binding were the most common molecular functions whilst biological processes involving transport and transcription were among the most highly represented (Figure 4.4), suggesting this module may represent a functional network of genes involved in transcriptional regulation.

Novel transcript c40155_g2 was found to be highly connected within the module *skyblue* revealing a number of connections to transcription factors and several other transcripts related to growth and nutrient uptake such as carboxypeptidase a2, insulin-like growth factor-binding protein and a proton-coupled folate
transporter-like. A gene network displaying the degree of connectivity between selected representative transcripts is presented in Figure 4.5.

Figure 4.3. Identification of gene co-expression models among the female banana shrimp hepatopancreas samples.
Figure 4.4. Gene ontology analysis. The multi-level node score distribution of the gene ontology categories comprising the module *skyblue* that contains the novel transcript c40155_g2. MF; Molecular function, BP; Biological Process, CC; Cellular Component.

Figure 4.5. Gene network map. A selection of annotated genes within the co-expression module *skyblue* and the degree of connectivity with these genes and novel transcript c40155_g2.
4.5 Discussion

In this study, we identified differences in gene expression within the hepatopancreas tissue of female and male banana shrimp of a size range known to exhibit sexually dimorphic growth; a trait of potential commercial importance. Whilst a small number of only 22 genes were identified as differentially expressed, we believe the size of the dataset provides a high degree of reliability for this observation. The large number of biological replicates per condition used in this study (n=14) is not commonly reported for RNA-Seq experiments involving crustaceans. This number of replicates together with an individual sequencing depth of over 20 million PE reads obtained per sample offers additional power in determining differences from transcripts with inherently low levels of expression.

Slightly more than half of the DEGs identified could be annotated using homology searches of the public databases, highlighting a familiar problem encountered when working with sequence data from non-model organisms. Of the 13 DEGs that could be annotated, minimal information was available at the time of writing regarding the function of these genes in crustaceans. A GO enrichment analysis of the DEGs against the full gene set also failed to return a significant result. Interestingly, 4 of the DEGs found over-expressed in female shrimp have been previously described with functions that could be somewhat linked. Chitinolytic enzymes in crustaceans are involved in essential functions including nutrient digestion, defence against pathogens and ecdysis (Huang et al. 2010). Chitin is degraded to soluble oligosaccharides by chitinases and subsequently hydrolysed to n-acetylglucosamine by the enzyme beta-n-acetylglucosaminidase (Watanabe et al. 1998). In penaeid shrimp, chitinase I is found predominately localized to the hepatopancreas and thought to be involved in the digestion of chitin-containing food and the modification of the peritrophic membrane (PM) in the intestine (Tan et al. 2000; Rocha et al. 2012; Salma et al. 2012). PMs are secreted from the gut epithelium in crustaceans and are primarily composed of chitin (Martin et al. 2006). A peritrophic membrane protein, peritrophin-44, was demonstrated to bind chiton and is thought to have roles in the maintenance of the PM structure in other arthropods (Elvin et al. 1996). While the functions of shrimp PMs are still poorly understood, there has been a number of studies undertaken in insects (Wang et al. 2012). A recent study reports that increased chitin content in the PM of the European corn borer (O. nubilalis) larvae dramatically decreased the larval body weight (Khajuria et al. 2010), suggesting that chitinolytic enzymes are important for regulating the chitin content and thus the porosity of the PM, influencing the rate of food digestion and nutrient absorption (Wang et al. 2012). The elevated expression of chitinase I, beta-n-acetylglucosaminidase and peritrophin-44-like transcripts in the hepatopancreas of female shrimp in this study may be contributing to modulation of the PM for improved nutrient absorption by reducing chitin content and thus increasing permeability. This would support the findings from previous studies reporting enhanced feed utilisation for female penaeid shrimp compared to males (S.W. and D.R.Hewitt 1994; Moss 2006).

Given the animals sampled for this study had not yet reached the stage of sexual maturity, it was not surprising that genes important for sexual maturation, such as vitellogenin, while present in the sequence data, were not determined to be differentially expressed. However, a cyclin B encoding transcript was identified as over-
expressed in female banana shrimp. Cyclin B is a well-known regulatory factor that plays a crucial role in mitosis and meiosis and is a component of a maturation promoting factor for ovarian development in crustaceans (Qiu and Yamano 2005; Visudtiphole et al. 2009). Higher levels of cyclin B could be an indication of increased cell proliferation in female shrimp or may be a signal of the early stages of maturation.

WGCNA has been used previously to identify connections between genes based on their profile of expression throughout a set of samples (Sasagawa et al. 2016). Sample sizes for WGCNA reported in the literature are often larger than 12 (Rodenas-Cuadrado et al. 2015), suggesting the number of replicates (n=14) used for the analysis of co-expression in this study, though relatively small, had sufficient power to identify robust gene networks. WCGNA was employed with the aim of gaining some insight into the function of novel transcript c40155_g2, which displayed strict sex-specific expression. Given that this transcript was almost exclusively expressed in female tissue, only the female samples were included in the analysis. Exploring the connectivity within the gene module skyblue that contained novel transcript c40155_g2, the strongest connections were identified between the following 5 highly connected transcripts. A Glutamate NMDA receptor homolog (c13456_g1) was among the most highly connected transcripts in the module. NMDA receptors have been shown to act as a primary upstream regulator of Methyl Farnes ate (MF) signalling in D. pulex (Toyota et al. 2015). Studies suggest that MF is involved in the regulation of reproduction and molting (Nagaraju 2007). Insulin-like growth factor-binding proteins (c41454_g1) are important for movement of insulin-like growth factors (IGFs) in the blood. The accessibility of these binding proteins can regulate IGF-based signalling which can result in an exaggeration or reduction of growth (Rosen et al. 2013). Trypsin (c4345_g1) is one of the most abundant digestive proteases found in shrimp (Navarrete del Toro et al. 2006). Studies suggest that trypsin levels can regulate metabolism and can be controlled via ecdysteroid hormones that modulate expression during molt cycles (Sánchez-Paz et al. 2003). Chondroitin sulfate proteoglycan (c64641_g1) is an extracellular matrix (ECM) protein that has been shown to be involved in the regulation and homeostasis of the pyloric circuit in the crab (C. borealis) (Hudson et al. 2015), which governs the rhythmic movements of the pyloric chamber (Tazaki and Tazaki 1997). Hemicentin (c17072_g1) is another ECM protein known to form tracks that regulate the attachment and movement of intestine on the epidermal basement membrane (Vogel et al. 2001). Taken together, the most highly connected genes with novel transcripts c40155_g2 appear to be associated with the regulation of digestion and growth. Considering a large proportion of GO categories for transcripts within the skyblue module relate to DNA and protein binding and that the majority of genes in this module identify as transcription factors, there is support for the postulate that this novel transcript may function as a gene for growth regulation, though additional research is needed.

4.5.1 Conclusions

This study reports a robust analysis of sex-specific gene expression within hepatopancreas tissue in the banana shrimp, supported by a considerable number of biological replicates. These data suggest an upregulation of transcripts involved with chitin metabolism present in the female hepatopancreas that could potentially influence food uptake. A novel transcript was also identified with no distinguishable features recognisable by
homology searches with the public sequence databases. WGCNA was useful in providing insight into putative functions of this novel transcript by revealing a number of connections with it and genes associated with regulating digestion and growth. Further studies involving the silencing of this novel transcript in developing animals would provide additional insight into its function.

4.6 Acknowledgements
We thank Nguyen Hong Nguyen for support with morphometric data analysis and a Marie Curie International Research Staff Exchange Scheme Fellowship within the 7th European Community Framework Programme (612296-DeNuGReC).
4.7 References


Toyota K, Miyakawa H, Yamaguchi K, Shigenobu S, Ogino Y, Tatarazako N, Miyagawa S, Iguchi T (2015) NMDA receptor activation upstream of methyl farnesoate signaling for short day-induced male offspring production in the water flea, Daphnia pulex. BMC Genomics 16:186.


Chapter 5: The genome of the Sydney rock oyster (*Saccostrea glomerata*)

5.1 Abstract

Oysters perform important roles in estuarine ecosystems and are of substantial economic value to fisheries and aquaculture worldwide. Contending with disease and environmental stress are considerable challenges to oyster culture. The Sydney rock oyster, *Saccostrea glomerata*, is an iconic and commercially important species of edible oyster in Australia. This study decodes the estimated 784 Mb *Saccostrea* genome using over 300-fold sequence coverage resulting in a total scaffold number of 10,107 and a scaffold N50 of 804 Kb. Analysis of the predicted 29,738 protein-coding genes identifies expansions of gene families associated with innate immunity and stress response and, together with gene expression data, offers insight into the genomic background of this oyster’s resilience to abiotic stress. This draft genome provides a valuable resource for further studies in molluscan biology and offers support for genetic enhancement programs involving commercially produced oyster species.

5.2 Introduction

Oysters of the family Ostreidae are a group of bivalve molluscs that include over 70 extant members considered to be keystone species widely distributed in estuarine ecosystems, performing important roles in mitigating turbidity and improving water quality. Edible oysters have established commercial significance in fisheries and aquaculture industries being among the most highly produced mollusc species in the world. Oysters have evolved an extraordinary resilience to the harsh conditions of intertidal marine environments and the persistent exposure to microbes encountered by filter-feeding. A rich and diverse set of immune and stress response genes in the oyster genome are thought to be pivotal to the remarkably effective host defence system that enables these animals to thrive in estuaries and coastal oceans worldwide. Despite these inherent adaptations, oyster populations both wild and captive are threatened with mass mortalities caused by epizootic infections and by factors associated with environmental change. Understanding and ameliorating susceptibility to these threats is essential for the establishment of secure mariculture and effective conservation.

The Sydney rock oyster (*Saccostrea glomerata*) is an economically important species of edible oyster in Australia, naturally populating the shorelines of the eastern coast. Its cultivation contributes substantially to an aquaculture industry dating back to the 19th century, supported by selective breeding programmes that have been operating for over 25 years. The Sydney rock oyster, in contrast to the more widely grown Pacific oyster (*Crassostrea gigas*), grows around 60% slower under favourable conditions, yet has a higher tolerance.

---

1. Note: This chapter is formatted in the style of the Nature Communications journal.
to abiotic stress, surviving up to three times longer out of water. It also appears to be resistant to the devastating Pacific Oyster Mortality Syndrome caused by the virus OsHV-1. These characteristics call for investigations into the nature of the higher resilience observed in *S. glomerata* and offer a unique opportunity for comparative studies given the recent availability of the Pacific oyster genome. Decoding the complete *Saccostrea* genome will enable a deeper understanding of the biology and evolution of the Ostreidae and will serve as a valuable resource for genetic improvement within the oyster farming industry. This study presents an annotated draft genome for *S. glomerata* and explores comparisons between genes relevant to survivability among a close relative and other more evolutionarily distant molluscs.

5.3 Results

5.3.1 Genome sequencing and de novo assembly

The genome of *S. glomerata* was estimated to be 784 Mb in size based on an analysis of k-mer frequency distribution and was sequenced to over 300-fold coverage based on this estimate by a whole-genome shotgun approach using the illumina HiSeq platform (Supplementary Table 5.1). Over 200 Gb of quality filtered short-insert and mate-pair read data (Supplementary Table 5.2) was initially assembled and scaffolded in accordance with the strategy outlined in Supplementary Figure 5.1. Further contiguity improvements were made by Chicago library sequencing, HiRise scaffolding and gap filling (Supplementary Note 1). The final *Saccostrea* draft assembly included 788 Mb in 10,107 scaffolds with a scaffold N50 of 804.2 Kb and a contig N50 of 39.8 Kb (Table 5.1). This genome is ~44% larger than that estimated for the closely related Pacific oyster *Crassostrea gigas* (545 Mb) yet smaller than the pearl oyster *Pinctada fucata* (1.14 Gb). Many invertebrate genomes, including that of oysters, are known for high levels of heterozygosity and repetitiveness which can complicate the assembly process. The repeat content of the draft genome was estimated to be 45.03% (Supplementary Table 5.3). The *Saccostrea* genome exhibits a relatively high level of heterozygosity based on the observation of 3.2 million SNPs and 354,373 short insertion/deletions (indels) in 703,199,470 eligible positions for a polymorphism rate of 0.51%, comparable with the inbred sequenced *Crassostrea* (0.73%) and in contrast with the octopus (0.08%). Almost 82% complete and 96% partial matches to the 248 core eukaryotic gene set could be identified within the draft assembly using the CEGMA pipeline (Supplementary Table 5.5), which is comparable with other invertebrate genome assemblies reported previously. Using the BUSCO tool we could detect a total of 787 (93.3%) of the 843 genes in the metazoan library with 672 (79%) of these being complete matches. Over 91% of the clean short-insert read data could be aligned to the draft assembly. The *S. glomerata* assembly achieved a high level of contiguity and completeness for an illumina-only invertebrate assembly without the use of bacterial artificial chromosome (BAC) sequencing. Half the assembly was contained in the longest 241 scaffolds ranging from 0.8 to 7.1 Mb. Over 90% of the assembled bases were covered by the longest 1,321 (13%) scaffolds. The assembly metrics of the *Saccostrea* genome are comparable to those of other published mollusc genomes and together with completeness measures, indicate the production of a comprehensive draft.
Table 5.1. Features of the draft *S. glomerata* genome.

<table>
<thead>
<tr>
<th><strong>Final Scaffolds</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of scaffolds</td>
<td>10,107</td>
</tr>
<tr>
<td>Number of contigs</td>
<td>39,887</td>
</tr>
<tr>
<td>Total scaffold length (Mb)</td>
<td>788.1</td>
</tr>
<tr>
<td>Total contig length (Mb)</td>
<td>746.5</td>
</tr>
<tr>
<td>% of bases in gaps</td>
<td>5.3</td>
</tr>
<tr>
<td>Scaffold N50 (Kb)</td>
<td>804.2</td>
</tr>
<tr>
<td>Contig N50 (Kb)</td>
<td>39.4</td>
</tr>
<tr>
<td>% main genome in scaffolds &gt; 50 Kb</td>
<td>94.3</td>
</tr>
<tr>
<td>Largest scaffold (Mb)</td>
<td>7.14</td>
</tr>
<tr>
<td>% GC content</td>
<td>33.31</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Gene Models</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number protein-coding gene models</td>
<td>29,738</td>
</tr>
<tr>
<td>Mean gene length (bp)</td>
<td>8,737</td>
</tr>
<tr>
<td>Mean number of exons per mRNA</td>
<td>8</td>
</tr>
<tr>
<td>% genome covered by genes</td>
<td>33.0</td>
</tr>
<tr>
<td>% genome covered by CDS</td>
<td>5.8</td>
</tr>
</tbody>
</table>

5.3.2 Genome annotation and comparative analysis

A total of 29,738 protein-coding genes from the *S. glomerata* genome were annotated using homology-based and *ab initio* predictions (Supplementary Note 2.1). Almost 88% of these were supported by RNA-Seq evidence derived from 6 different tissues. The gene content spanned a third of the genome with a mean gene length of 8,737 bp averaging 8 exons per transcript (Table 5.1). The number of gene predictions presented here is similar to that reported for the two published oysters *C. gigas* (28,027) and *P. fucata* (29,353)\(^5,13\). Given their close phylogenetic relationship, it was not surprising that of the 29,738 *S. glomerata* protein-coding genes, 26,741 (89%) returned a BLASTp top-hit (*E*-value \(\leq 10^{-5}\)) to *C. gigas* from searches of the NCBI non-redundant (nr) database, suggesting a high degree of similarity may exist among the gene sets in the Ostreidae family. Only 182 (0.61%) genes showed no similarity to any known sequence. Comparing BUSCO search results with the gene models from *Crassostrea* and *Pinctada* show that the predictions from this study have the greatest number of complete (82.4%) and least number of missing genes (6.5%) of the three, indicating this study has produced the most complete gene model set for an oyster species published to date (Supplementary Table 5.7). Of the entire *S. glomerata* gene set, 7,932 (26.7%) genes could be assigned to one or more Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways by homology inference and 21,116 (71%) were found to contain matches to the Pfam domain database (*E*-value \(\leq 10^{-6}\)). Comparison of gene families shared among molluscs *Saccostrea*, *Crassostrea*, *Pinctada* and *Lottia* show that *S. glomerata* has 972 unique gene families and shares 13,035 gene families with *C. gigas* (Figure 5.1).
5.3.3 Expansion of gene families

To better understand the genome-wide similarities among molluscs, an examination of the distribution of 8,629 Pfam domains across a diverse set of 23 metazoan genomes was performed. This analysis identified a number of significantly enriched Pfam domains in *Saccostrea* that were considered to represent gene family expansions. These included toll-like receptors (TLRs), immunoglobulin domain-containing genes, thrombospondins, complement (C1q) subunits, G-protein-coupled receptors (GPCRs) and heat-shock proteins (Figure 5.2). Of these expansions, some of the most pronounced have occurred in gene families associated with innate immune response. While an expansion of immune-related genes has been described previously in studies of *Crassostrea*, being attributed to a pathogen-rich and dynamic intertidal habitat, several protein families in *Saccostrea* however, have greatly exceeded the domain counts of those found in the other oysters. The *Saccostrea* genome contains 182 toll-interleukin receptor (TIR) domains which are notably expanded beyond the relatively broad repertoire of 61 and 91 found in *Crassostrea* and *Pinctada*, respectively. This level of expansion has extended to other families of non-self recognition proteins such as C-type lectins and fibrinogen-related proteins.

The fibrinogen_C gene family is the C-terminal globular domain of fibrinogen-related proteins (FREPs) which can function as a molecular recognition unit playing a significant role in immunological defence. There are a total of 576 genes in the *S. glomerata* genome containing at least one fibrinogen_C domain, more than in any of the 22 genomes included in the enrichment analysis. This number is triple that of *Crassostrea* (192) and *Pinctada* (163) and over one hundred more than *B. floridae* which has the highest compliment of FREPs (395).
of any other species included in the comparison (Figure 5.3c). The *Saccostrea* gene models containing fibrinogen_C domains vary considerably in their length, composition and domain arrangement and the majority appear to be indiscriminately distributed across the genome apart from a few main clusters, the largest being a set of 17 genes arranged in the same orientation ranging over 478 Kb on a single scaffold (Figure 5.3b). The majority of FREPs were not expressed among the tissue types sequenced in this study, however, the expression profile in 7 tissues suggests these genes are distinctly tissue-specific, occurring to a lesser extent within the hemolymph (Figure 5.3a). The broadest expression of FREPs can be seen in gill tissue, likely due to the increased exposure of this organ to external factors.


Apoptosis is vital for the survival of eukaryotes through the gentle removal of damaged cells that may interfere with normal function\textsuperscript{17}. Precise regulation of apoptosis is essential for an organism's ability to maintain homeostasis and adapt to ever changing environments\textsuperscript{18}. Inhibitor of apoptosis proteins (IAPs) are characterised by the presence of at least one copy of the baculovirus IAP repeat (BIR) domain which facilitate binding between IAPs and caspases, controlling apoptotic signalling\textsuperscript{19}. Oysters are thought to maintain a powerful anti-apoptosis system given the repertoire of 48 IAPs annotated from the genome of *Crassostrea*\textsuperscript{5}, and further supported by the 61 found in *Pinctada* in this study. The *Saccostrea* genome encodes a further 80 IAP genes, almost half of which form two clusters, based on phylogenetic analysis with the major group of 24 genes, suggesting a recent expansion (Figure 5.4a). Interestingly, this major cluster of genes does not appear to be localised to a particular region in the genome, being distributed over a several scaffolds. However, the
smaller cluster of 9 genes reside on a single scaffold 218 Kb in size. The expression profile of these 80 IAPs in 7 oyster tissues appears to be somewhat tissue-specific, suggesting the possibility of functional specialisations (Figure 5.4b). Although, the phylogenetic relationships of these IAPs do not appear to accord with tissue-specific patterns of expression.

Figure 5.3. Expansion and expression of S. glomerata fibrinogen_C domain-containing genes. (a) Expression profiles of 262 fibrinogen domain containing genes in 7 S. glomerata tissues. Cell colours indicate the number of standard deviations from the mean expression level. Expression was not detected in these tissues for the remaining 314 fibrinogen domain containing genes. (b) Scaffold SGL_242 and Scaffold SGL_221 contain the 2 largest clusters of fibrinogen_C domain-containing genes. The largest cluster is on Scaffold SGL_242 and contains 17 genes that vary in genomic span and are each orientated in the anti-sense direction. Scaffold SGL_221 contains 15 genes and is orientated with the majority in a set in the sense direction flanked by 5 genes in the anti-sense direction. (c) Distribution of fibrinogen_C domain-containing gene models detected using hmmsearch among 17 selected metazoan genomes.
Figure 5.4. Inhibitor of apoptosis (IAP) gene expansion in *S. glomerata*. (a) Phylogenetic tree of molluscan IAP genes in Sgl (blue), Pfu (green), Cgi (yellow), Aca (orange), Lgi (purple) and Obi (red). I, putative IAP gene expansion (24 genes); II, an IAP cluster (9 genes) in *Saccostrea* representing the largest co-located set of genes on a single scaffold (Scaffold SGL_814). (b) Expression profile of IAP genes in 7 oyster tissues.

5.4 Discussion

This study presents the sequenced and assembled genome of the Sydney rock oyster, *S. glomerata*, an important species for Australian aquaculture. This is the second reference genome from the family Ostreidae to be reported, enabling comparative studies between *Saccostrea* and *Crassostrea* that can improve our understanding of the mechanisms these organisms have evolved for survival in the highly stressful intertidal environment, facilitating selective breeding and enhancement of oyster cultivation practices. The draft assembly achieved comparably high levels of completeness and contiguity. The statistics for scaffold and contig N50 improved upon the two oyster genomes currently available due largely to technological advances in sequencing. The use of 250 bp read lengths and “Chicago” libraries contributed heavily to the production of a highly contiguous assembly, offering a more affordable alternative to BAC or fosmid sequencing.

The larger genome size of *Saccostrea* did not appear to contain a proportionately higher number of protein-coding genes, nor larger average exon or intron sizes. However, 354 Mb of repetitive sequence (45% of the genome) was identified compared with only 202 Mb (36% of the genome) found in *Crassostrea*. An enrichment of exonuclease and transposase domains such as PF01498 (15 out of 89 across 23 species) in
Saccostrea, that were absent in Crassostrea, suggests this may be the result of increased transposable element activity.

The types of gene families found expanded in the Saccostrea genome are not dissimilar to those seen in the other bivalve genomes. Genes associated with stress response and immune defence are thought to have expanded in response to the environmental variability and the exposures of filter-feeding confronted by the oyster. Some of these gene families are amplified up to 3 fold higher in Saccostrea than the other published molluscs and are more abundant than in any of the other 23 species investigated here. This observation leads to the suggestion that either the selective pressures placed on this species are more extreme than in other oysters or perhaps the mechanisms for gene expansion are intrinsically more active. Whichever the case, it is hypothesised that the degree of expansion is, at least in part, contributing to a phenotype for higher resilience.

The diversification of toll-like receptors (TLRs) is thought to be important for the recognition of Pathogen Associated Molecular Patterns (PAMPs) and the subsequent activation of an immunological response via MYD88 signalling. The expansions of TLR receptors may be an indication they are favoured over other types of immune receptors.

Large expansions of FREPs have been extensively studied in the snail Biomphalaria and are speculated to be due to lineage-specific selective pressure from trematode parasites like Schistosoma mansoni. Protozoan parasites are major threats for a variety of bivalve species. Commercial production of S. glomerata has been severely impacted by mortalities arising from infection with the protozoan pathogens Marteilia sydneyi causing QX disease and Bonamia roughleyi causing Winter Mortality, which infect S. glomerata exclusively. The repertoire of fibrinogen_C domain-containing genes is distinctly higher in Saccostrea than any of the 23 metazoan genomes included in this study (Figure 5.3c). This extraordinary expansion of FREPs in Saccostrea may be in response to selective pressure from persistent challenge by protozoan pathogens and a need for broad recognition capacity. Clustering of these genes within the genome offers some evidence of local tandem duplications and a clear pattern of tissue-specific expression indicate these genes may have evolved to perform specialised functions. Interestingly, the expression of FREPs appear somewhat concentrated in the gill tissue which coincides with the site of infection of both pathogens. However, only 262 (45%) of the 576 genes identified with a fibrinogen_C domain were found to be expressed among the set of 7 tissues included in this study. The tissue-specific pattern of expression suggests the analysis of a more extensive selection of tissue types and developmental stages would likely detect additional transcripts. Nonetheless, it is possible some proportion of these genes may no longer be functional and will need to be examined by further studies involving additional analysis of expression and manual curation of each gene model, even though the gene prediction pipeline was set to exclude models that did not contain a complete open reading frame.

In Australia, wild populations of S. glomerata have been under pressure from the introduction of the faster growing C. gigas, which rapidly overgrows and displaces the native oyster at the low to mid-intertidal zones. At higher intertidal areas, however, S. glomerata is able to persist due to its higher tolerance to the tougher
A highly developed anti-apoptosis system is thought to be important for the endurance of oysters to abiotic stress ascribed to the expansion of IAP genes and the propensity of air exposure to induce the dramatic upregulation of IAPs in *C. gigas*. The repertoire of 80 IAP genes in the *Saccostrea* genome compared to the 48 annotated in *Crassostrea* may help to explain why *S. glomerata* can withstand higher levels of air exposure. The highest expression of IAPs appear in the gill and hemolymph (Figure 5.4b), tissues which experience more severe stress compared to others from exchanges with the environment, yet a distinctly different set appear to express more highly in the gonad, indicating role specialisation among this gene family. A phylogenetic tree of the molluscan IAPs reveals two major clusters of 24 and 9 *Saccostrea* genes (Figure 5.4a) that together would constitute the extended expansion of genes found in this species over that of *Crassostrea*. These two clusters of genes would therefore be interesting targets for experiments measuring the response to air exposure.

The *Saccostrea* genome presented here provides a valuable resource for further studies in molluscan biology and offers support for genetic enhancement programs for commercially produced oyster species.

### 5.5 Methods

#### 5.5.1 Generation of sequence data
Mantle and gill tissues were dissected from a single female oyster for high molecular weight DNA extraction and library preparation. PCR-free short-insert libraries of 210 and 450 bp along with mate-pair libraries of 3, 6 and 9 Kb were sequenced on the illumina HiSeq 2500 (Supplementary Table 5.1). A Chicago library was produced from an additional oyster and sequenced on the illumina platform.

#### 5.5.2 Sequence assembly
Raw reads were quality filtered and trimmed using either Trimmomatic or Skewer to produce over 200 Gb of clean data (Supplementary Table 5.2). Clean reads were assembled and initially scaffolded *de novo* using Meraculous. For detailed information see Supplementary Note 1. Construction of a primary haploid assembly was performed using HaploMerger and further scaffolding performed by Dovetail Genomics (Santa Cruz, CA). Read mapping of genomic reads was performed using Bowtie. Rate of heterozygosity was estimated using the SAMtools/BCFtools (mpileup) pipeline to call SNPs and short InDels. Estimates of completeness were undertaken using both the CEGMA pipeline and BUSCO searches of the OrthoDB metazoan library.

#### 5.5.3 Genome annotation
Automated gene annotation was performed using MAKER (Supplementary Note 2.1). Repetitive sequences were soft-masked with RepeatMasker using the RepBase library and a custom repeat library generated with RepeatModeler. Protein-coding sequences from the genomes of *C. gigas*, *P. fucata*, *L. gigantea*, *O. bimaculoides*, *D. melanogaster* and *H. sapiens* were used for homology-based gene prediction. Stranded
RNA-Seq data was aligned to the genome for use with the BRAKER1 pipeline for training an AUGUSTUS model which was included with a stranded transcriptome assembly produced by Ertl et al., CEGMA derived proteins and the features of the RNA-Seq alignment in the MAKER2 pipeline. Predictions were filtered from the final gene set if they displayed no alignment to either the protein-coding genes from other species using BLAST (E-value <10\(^{-10}\)) or to the Pfam database.

5.5.4 Gene family analysis
Protein sequences from _S. glomerata_, _C. gigas_, _P. fucata_ and _L. gigantea_ were compared for orthology using all-against-all BLASTP alignment (E-value of 10\(^{-5}\)) and clustered using OrthoMCL (inflation value of 1.5) and visualised using ClusterVenn. Protein domain enrichment analysis was performed with Hidden Markov Model (HMM) searches of the Pfam database (Supplementary Note 2.1). Phylogenetic trees were generated using full length protein sequences aligned with MUSCLE and constructed with FastTree using the Jones-Taylor-Thornton model then visualised with FigTree v1.4.3 and MEGA 7. Functional annotation of the protein-coding genes from the _Saccostrea_ genome was undertaken using BLASTp 2.5.0+ (E-value 10\(^{-5}\)) to search for homologs with a local copy of the NCBI non-redundant database (nr). Protein-coding sequences were aligned against the NCBI KOG database (version 3/28/2017) using RPSBLAST v2.2.15 performed via the WebMGA server and for analysis of biological pathways present within the _Saccostrea_ genome, KEGG orthologous gene information was obtained using the KEGG Automatic Annotation Server (KAAS) using a bi-directional best hit (BBH) approach with the eukaryote representative gene set as reference.

5.5.5 Quantifying gene expression
Paired-end transcriptome libraries from gill, mantle, male gonad, female gonad, muscle, hemolymph and digestive system used in this study were generated previously. Expression levels were measured by aligning quality processed RNA-Seq reads to the genome assembly using HiSat. Mapped reads were sorted with SAMtools and counts reported as transcripts per one million mapped reads (TPM) using StringTie. TPM values visualised in heatmaps were transformed to log\(^2\) (TPM+1) and normalised using the scale() function in R.
5.6 References


40. Ertl, N. G., O’Connor, W. A., Papanicolaou, A., Wiegand, A. N. & Elizur, A. Transcriptome Analysis of the


Supplementary Figure 5.1. Workflow diagram of the *de novo* genome assembly and gene model prediction pipeline.
Supplementary Figure 5.2. Insert size distribution for fragment libraries. (a) Average insert size for the 210 library is 182 bp. (b) Average insert size for the 450 library is 403 bp.

Supplementary Figure 5.3. k-mer frequency distribution analysis (a) k-mer frequency histogram for k=55 shows a major heterozygous peak at 61 and a much smaller homozygous peak at 122. k-mers of depth 18 and below were not used in the assembly. (b) Cumulative fraction of 55-mers ($f_s$) shows that approximately 81% of the genome is present in 2 haploid copies.
Supplementary Figure 5.4. *k*-mer spectra plot of the assembly vs clean short-insert read data using the KAT tool. This graph shows a 0x coverage heterozygous peak (black) towards half that of the 1x heterozygous peak (red). A small amount of 2 fold *k*-mer coverage can be seen over the homozygous peak (orange). This result suggests there is a small amount of the second haplotype remaining in the assembly.
Supplementary Figure 5.5. Cumulative annotation edit distance (AED) scores for gene models produced using the MAKER2 pipeline. Almost 70% of the gene models have an AED score of 0.25 or less. MAKER2 predictions with an AED score above 0.5 were not included in the final gene model set.
Supplementary Table 5.1. Illumina genomic libraries and predicted coverage based on an estimated 784 Mb genome. The 210 bp short-insert libraries were run over 3 lanes and the mate-pair libraries run over 1 lane of the HiSeq 2500 in high-throughput mode. The 450 bp short-insert libraries were run over a single lane of the HiSeq 2500 in rapid mode.

<table>
<thead>
<tr>
<th>Library insert size</th>
<th>Estimated number clean reads</th>
<th>Average read length</th>
<th>Estimated total clean sequence</th>
<th>Estimated average insert size</th>
<th>Average read Q-score</th>
<th>% bases ≥Q20</th>
</tr>
</thead>
<tbody>
<tr>
<td>210 bp</td>
<td>1,081,100,000</td>
<td>144.3</td>
<td>156.0Gb</td>
<td>182</td>
<td>35.7</td>
<td>98.18</td>
</tr>
<tr>
<td>450 bp</td>
<td>216,380,000</td>
<td>218</td>
<td>47.2Gb</td>
<td>403</td>
<td>38.6</td>
<td>99.19</td>
</tr>
<tr>
<td>9 Kb</td>
<td>46,740,000</td>
<td>108.1</td>
<td>5.1Gb</td>
<td>9,637</td>
<td>35.5</td>
<td>96.23</td>
</tr>
<tr>
<td>6 Kb</td>
<td>55,540,000</td>
<td>107.6</td>
<td>6.0Gb</td>
<td>7,779</td>
<td>35.4</td>
<td>95.77</td>
</tr>
<tr>
<td>3 Kb</td>
<td>46,520,000</td>
<td>107.4</td>
<td>5.0Gb</td>
<td>3,110</td>
<td>35.4</td>
<td>95.81</td>
</tr>
</tbody>
</table>

Supplementary Table 5.2. Summary of the clean read data used for assembly and scaffolding.

<table>
<thead>
<tr>
<th>Library insert size</th>
<th>Read pairs</th>
<th>Read length</th>
<th>Raw bases</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>210 bp</td>
<td>294,332,697</td>
<td>2x151</td>
<td>88,888,474,494</td>
<td>113.3x</td>
</tr>
<tr>
<td>450 bp</td>
<td>276,801,509</td>
<td>2x151</td>
<td>83,594,055,718</td>
<td>106.6x</td>
</tr>
<tr>
<td>9 Kb</td>
<td>148,815,415</td>
<td>2x250</td>
<td>74,407,707,500</td>
<td>94.9x</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>246,890,237,712</td>
<td>314.8x</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Library insert size</th>
<th>Read pairs</th>
<th>Read length</th>
<th>Raw bases</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 Kb</td>
<td>44,887,683</td>
<td>2x151</td>
<td>13,556,080,266</td>
<td>17.3x</td>
</tr>
<tr>
<td>6 Kb</td>
<td>52,254,854</td>
<td>2x151</td>
<td>15,780,965,908</td>
<td>20.1x</td>
</tr>
<tr>
<td>9 Kb</td>
<td>43,259,524</td>
<td>2x151</td>
<td>13,064,376,248</td>
<td>16.6x</td>
</tr>
<tr>
<td>Chicago</td>
<td>205,716,319</td>
<td>2x151</td>
<td>62,126,328,338</td>
<td>79.24x</td>
</tr>
</tbody>
</table>
Supplementary Table 5.3. Length distribution and contig composition of the final scaffold output of the draft *S. glomerata* assembly.

<table>
<thead>
<tr>
<th>Minimum scaffold length</th>
<th>Number of scaffolds</th>
<th>Number of contigs</th>
<th>Total scaffold length (bp)</th>
<th>Total contig length (bp)</th>
<th>Scaffold contig coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>10,107</td>
<td>39,888</td>
<td>788,123,928</td>
<td>746,522,037</td>
<td>94.72%</td>
</tr>
<tr>
<td>1 Kb</td>
<td>8,887</td>
<td>38,659</td>
<td>787,021,822</td>
<td>745,420,983</td>
<td>94.71%</td>
</tr>
<tr>
<td>2.5 Kb</td>
<td>4,522</td>
<td>34,231</td>
<td>780,651,012</td>
<td>739,059,605</td>
<td>94.67%</td>
</tr>
<tr>
<td>5 Kb</td>
<td>3,545</td>
<td>33,069</td>
<td>777,145,808</td>
<td>735,700,524</td>
<td>94.67%</td>
</tr>
<tr>
<td>10 Kb</td>
<td>2,962</td>
<td>32,185</td>
<td>773,142,203</td>
<td>732,148,321</td>
<td>94.70%</td>
</tr>
<tr>
<td>25 Kb</td>
<td>2,340</td>
<td>31,023</td>
<td>762,915,764</td>
<td>722,764,252</td>
<td>94.74%</td>
</tr>
<tr>
<td>50 Kb</td>
<td>1,802</td>
<td>29,660</td>
<td>743,359,192</td>
<td>704,362,509</td>
<td>94.75%</td>
</tr>
<tr>
<td>100 Kb</td>
<td>1,314</td>
<td>27,746</td>
<td>708,442,865</td>
<td>671,338,074</td>
<td>94.76%</td>
</tr>
<tr>
<td>250 Kb</td>
<td>732</td>
<td>23,527</td>
<td>615,296,776</td>
<td>583,216,698</td>
<td>94.79%</td>
</tr>
<tr>
<td>500 Kb</td>
<td>414</td>
<td>18,974</td>
<td>502,826,883</td>
<td>476,747,558</td>
<td>94.81%</td>
</tr>
<tr>
<td>1 Mb</td>
<td>175</td>
<td>12,446</td>
<td>334,471,441</td>
<td>317,125,181</td>
<td>94.81%</td>
</tr>
<tr>
<td>2.5 Mb</td>
<td>32</td>
<td>4,401</td>
<td>121,469,582</td>
<td>115,113,814</td>
<td>94.77%</td>
</tr>
<tr>
<td>5 Mb</td>
<td>5</td>
<td>967</td>
<td>27,705,694</td>
<td>26,280,189</td>
<td>94.85%</td>
</tr>
</tbody>
</table>
Supplementary Table 5.4. Overview of the repetitive content of the *S. glomerata* genome.

<table>
<thead>
<tr>
<th>Repetitive element</th>
<th>Number of elements</th>
<th>Length occupied (bp)</th>
<th>Percentage of genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retroelements</td>
<td>1,160,841</td>
<td>341,108,100</td>
<td>43.28%</td>
</tr>
<tr>
<td>SINEs</td>
<td>100</td>
<td>8,607</td>
<td>0.00%</td>
</tr>
<tr>
<td>LINEs:</td>
<td>100,574</td>
<td>27,053,381</td>
<td>3.43%</td>
</tr>
<tr>
<td><em>RTE/Bov-B</em></td>
<td>4,354</td>
<td>1,420,195</td>
<td>0.18%</td>
</tr>
<tr>
<td><em>L2/CR1/Rex</em></td>
<td>1,749</td>
<td>349,408</td>
<td>0.09%</td>
</tr>
<tr>
<td>LTR elements:</td>
<td>20,307</td>
<td>10,678,615</td>
<td>1.35%</td>
</tr>
<tr>
<td><em>Gypsy/DIRS1</em></td>
<td>8,229</td>
<td>4,298,645</td>
<td>0.55%</td>
</tr>
<tr>
<td><em>BEL/Pao</em></td>
<td>1,478</td>
<td>756,084</td>
<td>0.19%</td>
</tr>
<tr>
<td>Unclassified</td>
<td>1,027,277</td>
<td>297,648,657</td>
<td>37.77%</td>
</tr>
<tr>
<td>DNA transposons:</td>
<td>27,280</td>
<td>5,377,254</td>
<td>0.68%</td>
</tr>
<tr>
<td><em>hobo-Activator</em></td>
<td>1,931</td>
<td>268,545</td>
<td>0.03%</td>
</tr>
<tr>
<td><em>Tc1-IS630-Pogo</em></td>
<td>2,838</td>
<td>555,625</td>
<td>0.07%</td>
</tr>
<tr>
<td>Low complexity</td>
<td>22,767</td>
<td>1,161,624</td>
<td>0.15%</td>
</tr>
<tr>
<td>Simple repeats</td>
<td>122,515</td>
<td>6,328,625</td>
<td>0.80%</td>
</tr>
<tr>
<td><strong>Total bases masked</strong></td>
<td><strong>354,917,563</strong></td>
<td><strong>45.03%</strong></td>
<td></td>
</tr>
</tbody>
</table>
### Supplementary Table 5.5. Estimates of genome assembly completeness using the CEGMA and BUSCO tools and read mapping statistics

<table>
<thead>
<tr>
<th></th>
<th>Number of Proteins</th>
<th>Completeness (%)</th>
<th>Total Number</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Complete</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete</td>
<td>203</td>
<td>81.85</td>
<td>229</td>
<td>1.13</td>
</tr>
<tr>
<td>Group 1</td>
<td>56</td>
<td>84.85</td>
<td>64</td>
<td>1.14</td>
</tr>
<tr>
<td>Group 2</td>
<td>43</td>
<td>76.79</td>
<td>47</td>
<td>1.09</td>
</tr>
<tr>
<td>Group 3</td>
<td>50</td>
<td>81.97</td>
<td>56</td>
<td>1.12</td>
</tr>
<tr>
<td>Group 4</td>
<td>54</td>
<td>83.08</td>
<td>62</td>
<td>1.15</td>
</tr>
<tr>
<td><strong>Partial</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete</td>
<td>238</td>
<td>95.97</td>
<td>285</td>
<td>1.2</td>
</tr>
<tr>
<td>Group 1</td>
<td>65</td>
<td>98.48</td>
<td>74</td>
<td>1.14</td>
</tr>
<tr>
<td>Group 2</td>
<td>53</td>
<td>94.64</td>
<td>66</td>
<td>1.25</td>
</tr>
<tr>
<td>Group 3</td>
<td>58</td>
<td>95.08</td>
<td>71</td>
<td>1.22</td>
</tr>
<tr>
<td>Group 4</td>
<td>62</td>
<td>95.38</td>
<td>74</td>
<td>1.19</td>
</tr>
<tr>
<td>Complete BUSCOs</td>
<td>672</td>
<td></td>
<td></td>
<td>79.72%</td>
</tr>
<tr>
<td>Complete and single-copy BUSCOs</td>
<td>631</td>
<td></td>
<td></td>
<td>74.85%</td>
</tr>
<tr>
<td>Complete and duplicated BUSCOs</td>
<td>41</td>
<td></td>
<td></td>
<td>4.86%</td>
</tr>
<tr>
<td>Fragmented BUSCOs</td>
<td>115</td>
<td></td>
<td></td>
<td>13.64%</td>
</tr>
<tr>
<td>Missing BUSCOs</td>
<td>56</td>
<td></td>
<td></td>
<td>6.64%</td>
</tr>
<tr>
<td>Total BUSCO groups searched</td>
<td>843</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mapped clean fragment reads</td>
<td></td>
<td></td>
<td></td>
<td>91.15%</td>
</tr>
<tr>
<td>Mapped RNA-Seq reads</td>
<td></td>
<td></td>
<td></td>
<td>68.88%</td>
</tr>
</tbody>
</table>
Supplementary Table 5.6. *S. glomerata* gene model annotation.

<table>
<thead>
<tr>
<th>Gene Models</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number protein-coding genes</td>
<td>29,738</td>
</tr>
<tr>
<td>Total gene length (Mb)</td>
<td>259.81</td>
</tr>
<tr>
<td>Total CDS length (Mb)</td>
<td>45.47</td>
</tr>
<tr>
<td>Longest gene (bp)</td>
<td>153,747</td>
</tr>
<tr>
<td>Longest CDS (bp)</td>
<td>60,294</td>
</tr>
<tr>
<td>Mean gene length (bp)</td>
<td>8,737</td>
</tr>
<tr>
<td>Mean CDS length (bp)</td>
<td>1,529</td>
</tr>
<tr>
<td>Mean intron length (bp)</td>
<td>987</td>
</tr>
<tr>
<td>Mean exon length (bp)</td>
<td>216</td>
</tr>
<tr>
<td>Mean number of exons per mRNA</td>
<td>8</td>
</tr>
<tr>
<td>% genome covered by genes</td>
<td>33.0</td>
</tr>
<tr>
<td>% genome covered by CDS</td>
<td>5.8</td>
</tr>
<tr>
<td>Any transcriptome support</td>
<td>26,158</td>
</tr>
<tr>
<td>BLASTp hits to NCBI nr database</td>
<td>29,556</td>
</tr>
<tr>
<td>Pfam annotation</td>
<td>21,116</td>
</tr>
<tr>
<td>KEGG annotation</td>
<td>7,932</td>
</tr>
<tr>
<td>KOG annotation</td>
<td>18,326</td>
</tr>
</tbody>
</table>

Supplementary Table 5.7. Completeness estimates of *Saccostrea* gene models and comparison with other published molluscan genomes. Performed using BUSCO tool searches against the metazoan library of 843 genes. Protein sequences were downloaded from ENSEMBL; GCA_000297895.1 for *C. gigas*, GCA_000327385.1 for *L. gigantea*, PRJNA270931 for *O. bimaculoides* and http://marinegenomics.oist.jp for *P. fucata* v2.0.

<table>
<thead>
<tr>
<th>BUSCOs</th>
<th><em>S. glomerata</em></th>
<th><em>C. gigas</em></th>
<th><em>P. fucata</em></th>
<th><em>L. gigantea</em></th>
<th><em>O. bimaculoides</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>695</td>
<td>656</td>
<td>525</td>
<td>721</td>
<td>669</td>
</tr>
<tr>
<td>Complete &amp; single copy</td>
<td>621</td>
<td>588</td>
<td>494</td>
<td>693</td>
<td>542</td>
</tr>
<tr>
<td>Complete &amp; duplicated</td>
<td>74</td>
<td>68</td>
<td>31</td>
<td>28</td>
<td>127</td>
</tr>
<tr>
<td>Fragmented</td>
<td>93</td>
<td>71</td>
<td>200</td>
<td>87</td>
<td>128</td>
</tr>
<tr>
<td>Missing</td>
<td>55</td>
<td>116</td>
<td>118</td>
<td>35</td>
<td>46</td>
</tr>
<tr>
<td>Total coding genes</td>
<td>29,738</td>
<td>26,101</td>
<td>31,477</td>
<td>23,340</td>
<td>38,556</td>
</tr>
</tbody>
</table>
S5.3 Supplementary Note 1. *S. glomerata* sequencing and *de novo* assembly

**S5.3.1 Oyster sample and sequencing**

The sequenced female oyster was obtained from the New South Wales Department of Primary Industries in Port Stephens, Australia. This oyster was from the 6th generation of a selective breeding program for fast growth. High molecular weight (HMW) DNA was extracted from mantle and gill tissue using either the MagAttract HMW DNA Kit (Qiagen) or the Gentra Puregene Tissue Kit (Qiagen). Oyster DNA was amplified and sequenced using a PCR assay for the cytochrome oxidase I (COI) gene as described by Liu et al.² and compared to published sequenced obtained from GenBank to confirm species identity. Short-insert libraries (210 bp) and mate-pair libraries (3, 6 and 9Kb) were prepared by the Ramaciotti Centre for Genomics, UNSW and sequenced over 4 lanes of the illumina HiSeq 2500. An additional short-insert library (450 bp) was prepared at the Australian Genome Research Facility and sequencing over 1 lane of the illumina HiSeq 2500 in rapid mode. All short-insert libraries were produced using TruSeq DNA PCR-free methods. Sequencing yield is presented in Supplementary Table 5.1.

**S5.3.2 Initial diploid genome assembly**

The raw sequence data was appraised for quality using FastQC [https://www.bioinformatics.babraham.ac.uk/projects/fastqc/](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Short-insert libraries were trimmed and filtered for quality using Trimmomatic (v0.36)³ and insert-size distributions calculated to confirm the accuracy of size selection in library preparation (Supplementary Figure 5.2). Mate-pair libraries were processed for quality and presence of junction adaptors using Skewer (0.2.2)⁴. The resulting clean read data is presented in Supplementary Table 5.2. *k*-mer frequencies among the short-insert reads were calculated for a *k*-mer size of 55 and shown to follow a bimodal distribution with a strikingly higher half-depth peak compared to the full-depth peak (Supplementary Figure 5.3a), indicating extremely high levels of heterozygosity in this genome. Genome size was estimated from these *k*-mers. The cumulative fraction of 55-mers at 1-1,000x depth is shown for the fragment libraries (Supplementary Figure 3b). The genome size was estimated using the script `kmerHistAnalyser.pl` from the Meraculous v2.2.2.5⁵ software and was determined to be 784.2 Mb using this method. Genome assembly was performed as outlined in Supplementary Figure 1. A number of different software using a range of different *k*-mer sizes were trialled and the best performing initial assembly was achieved using Meraculous2 with a *k*-mer size of 55. *k*-mers occurring at a frequency below 18 were considered erroneous and not included in the assembly. Due to the high levels of heterozygosity observed in the *k*-mer analysis, we opted to run Meraculous2 in diploid_mode 2. Using this parameter, the software preserves allelic variation encountered in the *k*-mer graph by outputting 2 haploid copies of the heterozygous regions. Traditional approaches of de Bruijn graph assembly software would attempt to incorrectly collapse heterozygous regions and previous attempts using these methods resulted in highly fragmented and incomplete assemblies. The Meraculous2 assembly and initial scaffolding at *k* = 55 resulted in a total scaffold length of 1,202 Mb, scaffold N50 of 107 Kb and scaffold number of 89,691.
S5.3.3 Construction of primary haploid assembly

As expected, the total scaffold length from the initial diploid assembly was far greater than the estimated genome size of 784 Mb as it incorporated 2 copies of the heterozygous regions of the genome. In order to construct a haploid reference assembly for the oyster, we used a set of scripts from the HaploMerger2⁶ software suite that was used successfully by Huang et al⁷ to separate haploid assemblies in the lancelet. HaploMerger2 attempts to separate homologous scaffolds within a repeat masked assembly to form both a reference haplotype assembly and an alternative haplotype assembly. Using the HaploMerger2 pipeline components B1-B5, we produced a haploid reference assembly with a total scaffold length of 796.5 Mb, closer to our estimated genome size. In order to confirm that the alternate haplotype content was successfully removed, a k-mer spectra plot was generated using the K-mer Analysis Tool (KAT)¹ that compared the k-mer content from the short-insert read data with the k-mer content from the draft assembly (Supplementary Figure 5.4). From this spectra plot, a 0x coverage heterozygous peak, representing k-mers not included in the assembly, is present at almost half that of the 1x heterozygous peak and only a very minimal amount of 2x coverage is present at this depth indicating that the majority of the alternative haplotype was successfully removed.

A small amount of 2x coverage is present at the homozygous depth peak suggesting a small amount of redundant sequence remaining in this assembly. This reference haploid assembly was comprised of 47,353 scaffolds with a scaffold N50 of 115 Kb, which was still unacceptably fragmented. To improve the contiguity of the reference assembly, additional scaffolding was undertaken using Chicago library reads by Dovetail Genomics, USA.

S5.3.4 Chicago library preparation and sequencing

A Chicago library was prepared as described previously⁸. Briefly, 500 ng of HMW gDNA (>50kb mean fragment size) was recovered from an additional oyster from the same breeding line, reconstituted into chromatin in vitro and fixed with formaldehyde. Fixed chromatin was then digested with DpnII, the 5’ overhangs were filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed and the DNA purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was sheared to ~350 bp mean fragment size, and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were then isolated using streptavidin beads before PCR enrichment of the library. The library was sequenced on an Illumina HiSeq 2500 in rapid run mode to produce 250 million 2x150bp read pairs for an estimated 39.1x (1-50kb pairs) physical coverage of the genome.

S5.3.5 Scaffolding input assemblies with HiRise

The single haplotype Meraculous2 genome assembly for S. glomerata and Chicago library read pairs in FASTQ format were used as input data for HiRise, a software pipeline purpose-built for using Chicago data to scaffold genomes⁹. Shotgun and Chicago library sequences were aligned to the draft input assembly using a modified
SNAP read mapper (http://snap.cs.berkeley.edu). The separations of Chicago read pairs mapped within draft scaffolds were analysed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify putative misjoins and score prospective joins. After scaffolding, shotgun sequences were used to close gaps between contigs. The resulting assembly contained 12,592 scaffolds with an N50 of 806 Kb. The largest scaffold increased from 1.02 Mb to 7.18 Mb in size.

**S5.3.6 Gap filling and redundancy removal**

As every join made during HiRise scaffolding creates a new gap in the assembly, the genome draft was processed using GapCloser v1.12 from the SOAPdenovo2 software suite with the high-quality reads generated from the 450 bp short-insert library. The result was a decrease in total gap length from 15.6% down to 5.3%. With the additional information recovered from gap closing, the scaffolds less than 10 Kb in size were aligned back to the major portion of the genome assembly and those that aligned with at least 90% identity over 80% of their length were removed as redundant. The statistics of the final scaffolds is presented in Table 5.1 and the length distribution of scaffolds is presented in Supplementary Table 5.3.

**S5.4 Supplementary Note 2. Gene model prediction and analysis**

**S5.4.1 Genome annotation**

A custom repeat library was produced from the genome assembly using RepeatModeler (http://www.repeatmasker.org/RepeatModeler/). Repeat sequences containing matches to Pfam domains not associated with transposable elements were removed. This custom library together with the Repbase library was used with RepeatMasker (http://www.repeatmasker.org/) to soft mask 45.03% of the assembly prior to annotation (Supplementary Table 5.4).

Stranded RNA-Seq data from mantle, muscle, digestive system, gonad, hemolymph and gill tissue previously generated by our group10 was aligned to the genome assembly using HiSat11 and annotated using StringTie12. The alignment features were passed to BRAKER13 as hints for training an AUGUSTUS model using Genemark. The stranded transcriptome assembly from the above RNA-Seq data, produced by Ertl et al.10, was used together with the published protein sequences from the genomes of 4 species of mollusc including *C. gigas, P. fucata, L. gigantea, O. bimaculoides* and also the protein sequences from the high quality genomes of *D. melanogaster* and *H. sapiens* as evidence for homology-based gene prediction using MAKER214, as outlined in Supplementary Figure 5.3. Transcriptome data is prone to misassembly and chimeras and was therefore only used for evidence and not for prediction. Proteins sequences identified from the CEGMA searches were passed to the MAKER2 pipeline for training SNAP along with the BRAKER1 trained AUGUSTUS model for *ab initio* predictions, using the HiSat RNA-Seq alignments as transcript-based evidence. Gene model predictions were retained if they were greater than 50 amino acids in length and had an annotation edit distance (AED) score of 0.5 or less. The AED score is a metric assigned to a gene model by MAKER2 that is a measure of the degree of fit that model has with the supporting evidence.
The initial gene model set produced using this method resulted in 38,545 predictions with the majority (over 69%) assigned an AED score of 0.25 or less (Supplementary Figure 5.5). This initial set of gene model predictions was filtered further by requiring genes to have either a BLASTp match (E-value <10^-10) to the training set of proteins from the 6 species or to contain a hit to the Pfam database. The final set of predictions consists of 29,738 protein-coding genes (Supplementary Table 5.6), 26,158 of which are supported by alignments to the RNA-Seq read data using HiSat. Using searches of the of gene models for BUSCOs from the metazoan libraries, we could identify 788 (93.5%) of the 843 genes in the metazoan library with 695 (82.4%) of these being complete matches (Supplementary Table 5.7), which is more complete than searching the entire genome assembly. Comparing searches of BUSCOs with the other published oyster gene models, the *S. glomerata* predictions were more complete, had fewer missing genes and were also comparable with other published mollusc species suggesting the production of a comprehensive set of gene models.

### 5S.4.2 Pfam enrichment analysis

In order to identify expansions of gene families within the *Saccostrea* genome, the protein-coding sequences were searched for Pfam domains to assign gene function. We used HMMER (hmmscan) to search the Pfam A database (release 31) for 8,629 different domains for 23 different species of metazoa. Counts of each domain were collated for each species with domains that occurred multiple time in a protein sequence being counted only once. Fisher’s exact test was then iteratively conducted using R, comparing the number of counts in Pfam families found in an individual genome, normalised by the total gene count for that species, against the background which was defined as the average of the counts in the remaining species. Multiple testing corrections were done with the FDR method in R for all calculated p-values. A Pfam domain was considered expanded if it showed a significant corrected p-value (0.01). To present the results in the heatmap shown in Figure 5.2, counts were normalised within each gene family across species by the scale function in R.
S5.5 Supplementary References


Chapter 6: Conclusions

At inception, the scope of this doctoral study was to address the depauperate molecular resources available for aquaculture in Australia by leveraging the production and bioinformatic analysis of next-generation sequence data for two key Australian aquaculture species; the banana shrimp (Fenneropenaeus merguiensis) the Sydney rock oyster (Saccostrea glomerata).

The term “next-generation”, referring to sequence data produced on the high-throughput platforms that were developed after the Sanger sequencing instruments, has over the period of this study been more commonly replaced in the literature and elsewhere with the term “second-generation”. This is due to the development of subsequent, “third-generation” technologies to the pioneering 454FLX and Solexa (illumina) platforms that now offer expanded capabilities over these earlier technologies, notwithstanding further performance refinement of the enduring illumina systems. This change in vernacular expresses the realisation of the rapid pace at which sequencing technology is evolving and it is with this in mind that the following conclusions from this body of work are presented.

The four research chapters of this thesis each align with one of the four objectives set out in Chapter 1, developed specifically to target concerns important to the cultivation of shellfish.

Objective I. To uncover the key genes associated with traits important to reproduction, development and sex determination in F. merguiensis, the first dataset of transcriptomic sequences from tissues relevant to biological systems governing commercially important traits will be assembled, annotated and characterised.

The first research chapter (Chapter 2) reports a published article addressing the above objective. Using a combination of second-generation sequence technologies, a comprehensive reference transcriptome was developed for the banana shrimp and analysed using homology searchers against the available databases. This work provided a significant lift in the genomic resources available for this species and revealed a number of interesting findings with regard to gene content and expression. Working with RNA-Seq data from the now retired 454FLX platform in combination with more up-to-date illumina data was at the time somewhat novel, though, had the data been generated using current technologies, perhaps a more comprehensive view of gene expression across tissues could have been realised. Nonetheless, the decoding of thousands of gene sequences specific to the banana shrimp will remain a useful resource for the foreseeable future.

Objective II. To explore differences in the expression of genes between families of adult farmed F. merguiensis exhibiting extremes of variation in levels of viral load for the purpose of developing molecular markers for disease resistance and immune response to be used in the selective breeding of these animals.

Chapter 3 reports a published article in response to the second objective of this thesis. A more comprehensive hepatopancreas transcriptome was generated for this study that markedly improved upon the assembly from Chapter 2, realised by the production of an improved form of sequence data and by the use of more recent and advanced bioinformatics tools. The annotation of protein-coding sequences were also improved due
mainly to the expansion of the public databases. As additional sequences are submitted from more closely related species over time, significant matches will of course start to be found, extending homology based inference. As a large proportion of the genes identified as differentially expressed in this study were unable to be annotated, revisiting the database searches to identify these genes as more information becomes available, or with the release of a penaeid shrimp genome, could offer further insight into the molecular pathways involved in viral defence. This work employed a large number of biological replicates (n=24) taken from genetic families, which is not often reported for an aquaculture species, and offered some insight to the very relevant issue of viral infection in shrimp cultivation in Australia. Single nucleotide polymorphisms identified in immune related genes that were only found in disease resistant animals appear to be a promising starting point for developing molecular markers of disease resistance. However, significant investment is required to determine the efficacy of such markers in selective breeding programs; an effort that may be too demanding for the scale of the banana shrimp industry at this time.

**Objective III.** To reveal the sex-specific transcript expression in the digestive gland of female and male *F. merguiensis* bred and grown together in captivity in order to improve our understanding of the molecular basis for the female-superior sexual dimorphism observed in this species.

The work reported in Chapter 4 provides some perspective into Objective III and is ready to be submitted for publication. This study exploited the utility of the transcriptome assembly constructed in the previous chapter to test a larger number of samples (n=28) for sex-specific differences. This chapter highlights the utility of using a large number of biological replicates in gene expression studies by furthering the power of differential expression analysis and enabling insight into unannotated genes through the use of weighted gene co-expression network analysis. With the 14 female samples included only just reaching the minimum recommended for this type of analysis, increasing the sample size towards 20 females may have resulted in improved network construction. RNA-Seq experiments are already becoming more affordable and studies will soon benefit from the addition of larger sample sizes. This work proposes a partial explanation for the differences in size observed between male and female shrimp that would be interesting to examine further with physiological studies.

**Objective IV.** To investigate the genomic complexity of *S. glomerata*, the first draft genome for this species will be constructed using a combination of sequencing, assembly and annotation methods with the goal of providing information on genomic features and to perform comparisons with the genomes of other species.

Chapter 5 presents the completed draft genome of *S. glomerata* and a major component of the genome paper soon to be submitted for publication. The genome sequencing strategy for the Sydney rock oyster referred to in the final research chapter of this thesis was developed and modified over a three year period. During this time, what was available and affordable at the initial planning stages was quite limited compared with the technologies released toward the later stages of the project. It has become clear that genome projects should be completed as quickly as practicable after the first sequence data has been generated in order to capitalise
on the impact of the assembly outcome achieved using the technologies available at the time. Assembly statistics are quickly improving with ever increasing contig and scaffold N50 scores such that chromosome-scale assemblies are already being realised in many species livestock and crops. Nevertheless, the assembly methods reported in this chapter take advantage of cutting-edge scaffolding techniques that were able to achieve a more contiguous assembly than any bivalve published previously. Like the reference transcriptomes produced in the earlier stages of this project, this draft assembly is a resource that will be of benefit to the shellfish aquaculture industry into the future, and has already revealed the expansion of immune genes in the *S. glomerata*, possibly explaining its higher resistance to harsh environmental conditions, especially in comparison with the Pacific oyster.

6.1 Future directions

This research has produced a rich set of molecular resources for species of Australian shrimp and oysters intended to support and further the capacity for studies involving these animals. It is anticipated that a broad array of future research will stem from the availability of these data which could be discussed at length. However, there are two particularly exciting research opportunities I wish to mention.

RNA-Seq has been used to discover large numbers of SNPs in humans, animals and plants and is thought to be more likely to detect functional variants given the capture of transcribed regions (Davey et al. 2011). Further study of the homozygous SNPs discovered among immune-related genes in disease resistant shrimp (Chapter 3), especially those found in the transcripts of putative receptors, presents an exciting opportunity for elucidating potential markers of disease resistance. Measuring the prevalence of viruses in animals selected for these variants would help to ascertain the importance of these genes to the rate of persistent infection. As there is a moderate rate of heritability estimated for HPV infection in shrimp (Knibb et al. 2015), functional marker-based selection may be applied to rapidly address viral infection in these animals.

Some of the obvious questions arising from comparative genome analysis with other molluscs is currently being addressed in view of the forthcoming *S. glomerata* genome paper, such as syntenic relationships, divergence estimates and rates of gene evolution. Though, additional work that may be outside the scope of this paper is in the development of genetic linkage maps to further anchor scaffolds and improve contiguity. Linkage maps are developed using thousands of short molecular tags that can be aligned to the genome sequence and have been used successfully to order scaffolds in line with the chromosomal arrangement (Chen et al. 2014; Braasch et al. 2016; Shao et al. 2016). The development of linkage maps for the Sydney rock oyster has the potential to lift the draft assembly towards chromosome scale scaffold lengths and will enable a clearer examination of gene family distributions across the genome. Given there is a breeding program currently in operation, samples would be easily obtained to generate these data. And last, but not least, a good draft genome is one of the few scientific outputs which is expected to be useful for a growing number of applications, as we continue with our quest to understand genomes.
6.2 References


