Generating genomic resources for two crustacean species and their application to the study of White Spot Disease

Submitted by

Bas Verbruggen

To the University of Exeter as a thesis for the degree of

Doctor of Philosophy in Biological Sciences, September 2016

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature: ……………………………………(Bas Verbruggen)

Abstract

Over the last decades the crustacean aquaculture sector has been steadily growing, in order to meet global demands for its products. A major hurdle for further growth of the industry is the prevalence of viral disease epidemics that are facilitated by the intense culture conditions. A devastating virus impacting on the sector is the White Spot Syndrome Virus (WSSV), responsible for over US $ 10 billion in losses in shrimp production and trade. The Pathogenicity of WSSV is high, reaching 100 % mortality within 3-10 days in penaeid shrimps. In contrast, the European shore crab *Carcinus maenas* has been shown to be relatively resistant to WSSV. Uncovering the basis of this resistance could help inform on the development of strategies to mitigate the WSSV threat. *C. maenas* has been used widely in studies on ecotoxicology and host-pathogen interactions. However, like most aquatic crustaceans, the genomic resources available for this species are limited, impairing experimentation. Therefore, to facilitate interpretations of the exposure studies, we first produced a *C. maenas* transcriptome and genome scaffold assembly. We also produced a transcriptome for the European lobster (*Homarus gammarus*), an ecologically and commercially important crustacean species in United Kingdom waters, for use in comparing WSSV responses in this, a susceptible species, and C. maenas.

For the *C. maenas* transcriptome assembly we isolated and pooled RNA from twelve different tissues and sequenced RNA on an Illumina HiSeq 2500 platform. After *de novo* assembly a transcriptome encompassing 212,427 transcripts was produced. Similar, *the H. gammarus* transcriptome was based on RNA from nine tissues and contained 106,498 transcripts. The transcripts were filtered and annotated using a variety of tools (including BLAST, MEGAN and RSEM) and databases (including GenBank, Gene Ontology and KEGG). The annotation rate for transcripts in both transcriptomes was around 20-25 % which appears to be common for aquatic crustacean species, as a result of the lack of well annotated gene sequences for this clade. Since it is likely that the host immune system would play an important role in WSSV infection we characterized the IMD, JAK/STAT, Toll-like receptor and other innate immune system pathways. We found a strong overlap between the immune system pathways in *C. maenas* and *H. gammarus*. In addition we investigated the sequence diversity of known WSSV interacting proteins amongst susceptible penaeid shrimp/lobster and the more resistant *C. maenas*. There were differences in viral receptor sequences, like Rab7, that correlate with a less efficient infection by WSSV.

To produce the genome scaffold assembly for *C. maenas* we isolated DNA from muscle tissue and produced both paired-end and mate pair libraries for processing on the Illumina HiSeq 2500 platform. A *de novo* draft genome assembly consisting of 338,980 scaffolds and covering 362 Mb (36 % of estimated genome size) was produced, using SOAP-denovo2 coupled with the BESST scaffolding system. The generated assembly was highly fragmented due to the presence of repetitive areas in the *C. maenas* genome. Using a combination of *ab initio* predictors, RNA-sequencing data from the transcriptome datasets and curated *C. maenas* sequences we produced a model encompassing 10,355 genes. The gene model for *C. maenas Dscam*, a gene potentially involved in (pan)crustacean immune memory, was investigated in greater detail as manual curation can improve on the results of ab initio predictors. The scaffold containing *C. maenas Dscam* was fragmented, thus only contained the latter exons of the gene. The assembled draft genome and transcriptomes for *C. maenas* and *H. gammarus* are valuable molecular resources for studies involving these and other aquatic crustacean species.

To uncover the basis of their resistance to WSSV, we infected *C. maenas* with WSSV and measured mRNA and miRNA expression for 7 time points spread over a period of 28 days, using RNA-Seq and miRNA-Seq. The resistance of *C. maenas* to WSSV infection was confirmed by the fact that no mortalities occurred. In these animals replicating WSSV was latent and detected only after 7 days, and this occurred in five of out 28 infected crabs only. Differential expression of transcripts and miRNAs were identified for each time point. In the first 12 hours post exposure we observed decreased expression of important regulators in endocytosis. Since it is established that WSSV enters the host cells through endocytosis and that interactions between the viral protein VP28 and Rab7 are important in successful infection, it is likely that changes in this process could impact WSSV infection success. Additionally we observed an increased expression of transcripts involved in RNA interference pathways across many time points, indicating a longer term response to initial viral exposure. miRNA sequencing showed several miRNAs that were differentially expressed. The most striking finding was a novel *C. maenas* miRNA that we found to be significantly downregulated in every WSSV infected individual, suggesting that it may play an important role in mediating the response of the host to the virus. *In silico* target prediction pointed to the involvement of this miRNA in endocytosis regulation. Taken together we hypothesize that *C. maenas* resistance to WSSV involves obstruction of viral entry by endocytosis, a process probably regulated through miRNAs, resulting in inefficient uptake of virions.

Acknowledgements

First of all I would like to thank my supervisors Dr. Ronny van Aerle, Dr. Eduarda Santos and Professor Charles Tyler for their continued support over the course of my PhD. Ronny did a great job at introducing me to the world of sequencing and bioinformatics. I was definitely not the only one that was sad to see him leave the University of Exeter. But even at his new job at Cefas he found time to help me out when I was stuck (or caused the server to break down). His relentless optimism makes working with him a pleasure. After Ronny's departure Eduarda and Charles really came through as supervisors and both in their own way have filled the gap that Ronny left. I am also grateful for the long hours they have put in reading and editing my writing.

Dr. Lisa Bickley was instrumental to the work presented in this thesis. I'm very grateful for the countless hours she spent in the lab making sure we had top notch sequencing libraries to work with. She was also a great companion on several conference trips. I would also like to give big thanks to the people at Cefas. Dr. Kelly Bateman did lots of experimental work for this project. Dr. Grant Stentiford's passion for aquatic crustacea and the world outside the University is highly contagious. Finally I'm grateful to David for helping out fixing problems on the server. I'd like to thank Konrad Paszkiewicz, Karen Moore and Audrey Farbos of the Exeter Sequencing Service for processing our samples and providing access to the zeus computer cluster.

I would like to thank all my colleagues from lab 201. In particular Aoife Parsons for managing to sit next to me for four years, being supportive and sharing the ups and downs of PhD life. Rhys for making my desk look tidy. Jenny for getting me away from my desk on coffee breaks and Lauren for providing entertainment on bioinformatics courses.

A shout out to all my friends. Damian Rumble and Lewis Elliot for introducing me to life in the UK and sharing the PhD experience with me. Martijn, Geert, Giovanni, Floris and Marloes for our holiday trips and Halve Liter Woensdagen whenever I was back in the Netherlands. In particular I would like to thank Lauren Hitchman for her support and wisdom. I could not have finished this PhD without her.

Finally I would like to thank my parents for their continued support and encouragement over the years.

Contents

Research Papers and Author's Declaration

Research paper 1: Verbruggen, B., Bickley, L. K., van Aerle, R., Bateman, K. S., Stentiford, G. D., Santos, E. M., & Tyler, C. R. (2016). Molecular Mechanisms of White Spot Syndrome Virus Infection and Perspectives on Treatments. *Viruses*, *8*(1), 23.

Research paper 2: Verbruggen, B., Bickley, L. K., Santos, E. M., Tyler, C. R., Stentiford, G. D., Bateman, K. S., & van Aerle, R. (2015). *De novo* assembly of the *Carcinus maenas* transcriptome and characterization of innate immune system pathways. *BMC Genomics*, *16*(1), 458.

Research paper 3: *De novo* Assembly of the European shore crab (*Carcinus maenas*) genome. *Manuscript in preparation*.

Research paper 4: Sequencing and *De novo* assembly of the *Homarus gammarus* transcriptome and characterization of its immune system. *Manuscript in preparation.*

Research paper 5: RNA sequencing of the shore crab (*Carcinus maenas*) after White Spot Syndrome Virus exposure points to endocytosis regulation as source of resistance to infection. *Manuscript in preparation*.

Statement: I, Bas Verbruggen, made the following contributions to the research papers presented in this thesis. I collected the information and wrote the manuscript for **paper 1**. For **papers 2, 3** and **4** Kelly Bateman and Lisa Bickley sourced the animals. Lisa Bickley prepared tissue samples, isolated RNA/DNA and prepared libraries for sequencing. Sequencing was performed by the Exeter Sequencing service. I analysed the sequencing data and wrote the manuscripts. For paper 5 the viral exposure experiment was performed by Lisa Bickley and Kelly Bateman at Cefas, Weymouth. Lisa Bickley isolated RNA and prepared libraries for sequencing. Sequencing was performed by the Exeter Sequencing service. I analysed the sequencing data and wrote the manuscript.

List of General Abbreviations

AGO	Argonaute
AHPND	Acute Hepatopancreatic Necrosis Disease
ALF	Anti-lipopolysaccharide factor
ATP	Adenosine Triphosphate
BLAST	Basic Local Alignment Search Tool
BLAT	Basic local alignment tool
bp	Basepair
BUSCO	Benchmarking Universal Single-Copy Orthologs
BWA	Burrows-Wheller Aligner
CA	California
cDNA	Complementary DNA
CEGMA	Core eukaryotic genes and a tool
con	Control
COPI	Coat protein
CV	Coefficient of variation
ddNTP	Dideoxynucleotides
DE	Differentially expressed
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide
Dscam	Down syndrome cell adhesion molecule
dsDNA	Double stranded DNA
EMS	Early Mortality Syndrome
ERC	Endocytic recycling compartments
EVE	Endogenous viral elements
FADD	Fas-Associated protein with Death Domain
FAO	Food and Agriculture Organization of the United Nations
FDR	False discovery rate
FN	Fibronectin
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
Gb	Gigabase
GLM	Generalized linear model
GNBP	Gram-negative binding protein
GO	Gene ontology
GSEA	Gene Set Enrichment Analysis
h	Hour
HMM	Hiddem Markov Model
hpi	Hours post injection
Hsc70	Heat shock protein 70 kDA
ie1	Immediate early 1
Ig	Immunoglobulin
IGV	Integrative genomics viewer
IHHNV	Infectious hypodermal and hematopoietic necrosis vir
IKK	IκB kinase

IMD	Immunodeficiency
iNOS	Inducible nitric oxide synthase
IPA	Ingenuity pathway analysis
JAK	Janus Kinase
KAAS	KEGG Automatic Annotation Server
KO	KEGG Orthologies
KOG	KEGG orthologous group
LINE	Long interspersed elements
LS	Library size
LSS	Loose Shell Syndrome
MAPK	Mitogen-activated protein kinase
miRNA	Micro RNA
miRNA-Seq	MiRNA sequencing
mp	Melanization protease
MQ	Median quartile
mRNA	Messenger RNA
NADH	Nicotinamide adenine dinucleotide
NB	Negative binomial
NCBI	National Center for Biotechnology Information
NF-κB	Nuclear Factor κB
NGS	Next generation sequencing
NO	Nitric oxide
nt	Nucleotide
ORF	Open reading frame
PAMP	Pathogen associated molecular patterns
PaV	*Panulirus argus* Virus
PCA	Principal component analysis
PCR	Polymerase chain reaction
PIAS	Protein inhibitors of activated STAT
PPAE	PPO activating enzyme
PPO	Prophenol oxidase
PPR	Pathogen recognition receptor
Rab	Ras-related protein
RAD	Restriction site Associated DNA
RC	Read count
RNA	Ribonucleic acid
RNAi	RNA interference
RNA-seq	RNA sequencing
rRNA	Ribosomal RNA
RSEM	RNA-Seq by Expectation Maximization
siRNA	Small interfering RNA
SMRT	Single molecule read-time
SND	Single nucleotide difference
SNP	Single nucleotide polymorphism

SOCS	Suppressors of cytokine signalling
sp	Serine protease
STAT	Signal transducer and activator of transcription
TECP	Thioester containing protein
TEM	Transmission electron microscopy
TLR	Toll like receptor
TM	Transmembrane
TMM	Trimmed mean of M values
TPM	Transcripts per million
TRBP	TAR RNA-binding protein
TSV	Taura Syndrome Virus
UK	United Kingdom
UQ	Upper quartile
USA	United States of America
USD	US dollar
VP	Viral Protein
ws	White spot
WSD	White Spot Disease
WSSV	White Spot Syndrome Virus
WSSV-CN	WSSV China isolate
WSSV-KR	WSSV Korea isolate
WSSV-TH	WSSV Thailand isolate
WSSV-TW	WSSV Taiwan isolate
Y2H	Yeast two-Hybrid
YHV	Yellow Head Virus

Chapter 1

General Introduction

Chapter 1: General introduction

1.1 Crustacean Aquaculture

Crustacean aquaculture is a significant part of global food production. Through commercial farming the sector helps satisfy the requirements of the human population for large quantities of shrimp, lobsters and other crustaceans. According to data from the Food and Agriculture Organization of the United Nations (FAO) the aquaculture industry has been growing steadily over the past three decades (Figure 1). The crustacean sector currently produces nearly 7 million metric tonnes of product, accounting for about 10 % of global aquaculture production. The market value for crustacea is considerable larger at nearly 24 % of total (Figure 2).

Within crustacean aquaculture shrimp are the most important species farmed. The production of shrimp accounts for more than half of total crustacean aquaculture production (~ 66 % of total production in 2013) [1]. Since 2006 between 3.0 and 4.0 Million metric tonnes of shrimp have been farmed by the global aquaculture industry [2]. Since 2011 production of shrimp in aquaculture has surpassed the yield from wild capture fisheries [2]. While many countries across the globe have an active crustacean aquaculture industry, the sector is mainly centred in (Southeast) Asia and India. The latter doubling its total production over the last five years [1]. While there are occasional drops in production, it can be expected that during the coming decades this multi-billion dollar market will continue to grow [2].

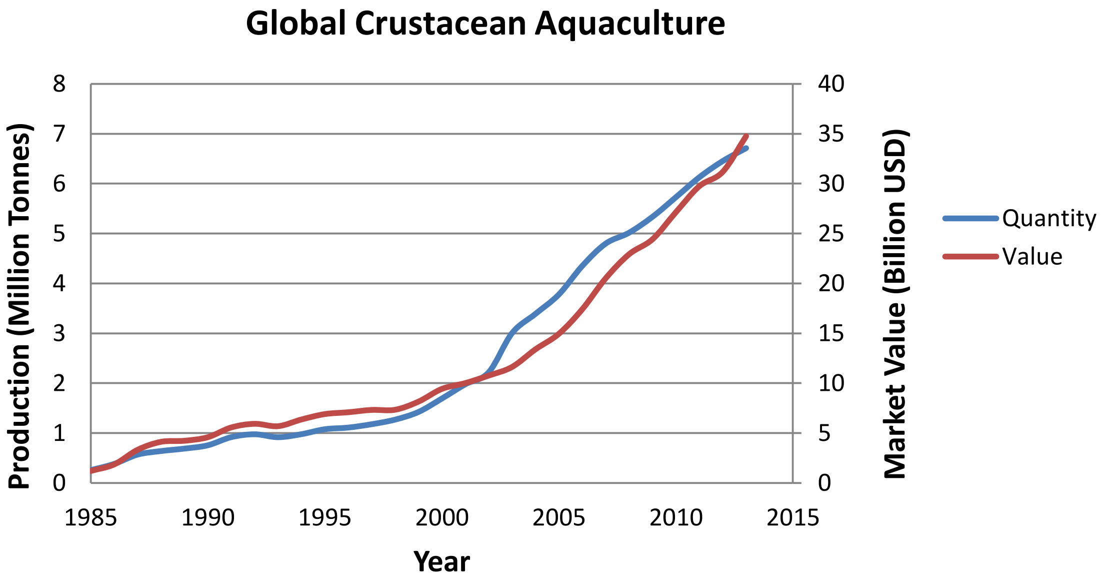


Figure 1 Growth of the crustacean aquaculture industry over time. Data from FAO Fishery Statistics Collections Database [1]

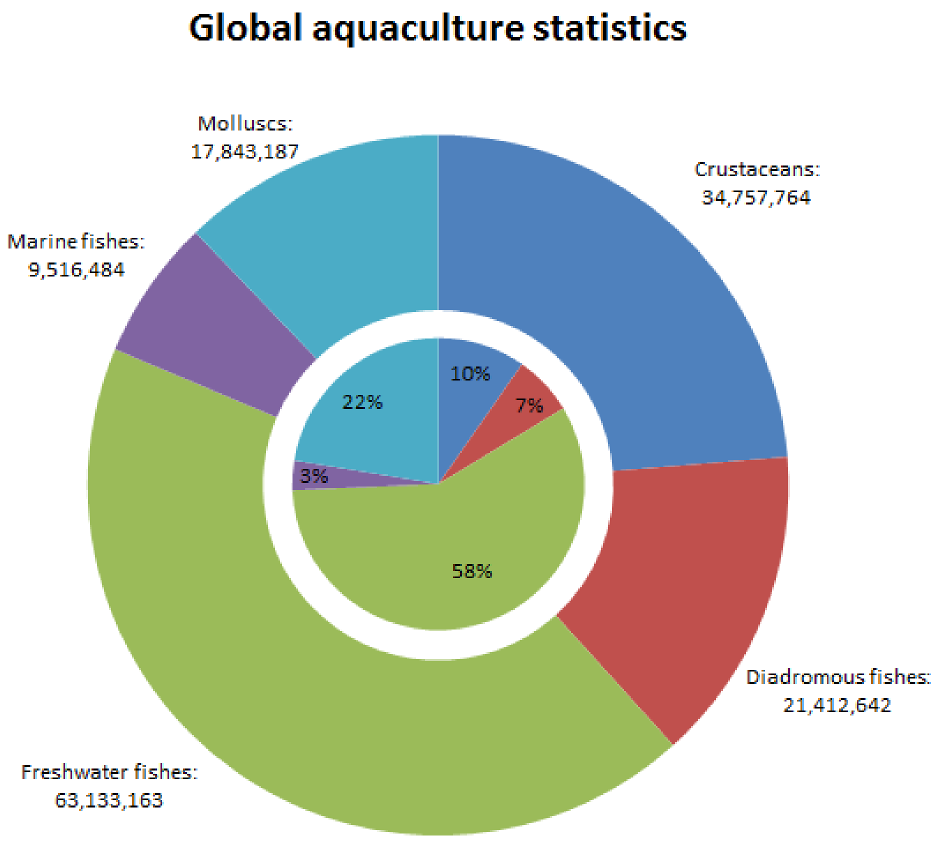


Figure 2 Global aquaculture statistics. Overview of global aquaculture by species type. *Inner pie chart*: Total production in percentages. *Outer circle*: Market value in USD 000. While crustaceans account for 10 % total production the market value is higher at nearly 25 %. Data from FAO aquaculture statistics 2013 [1].

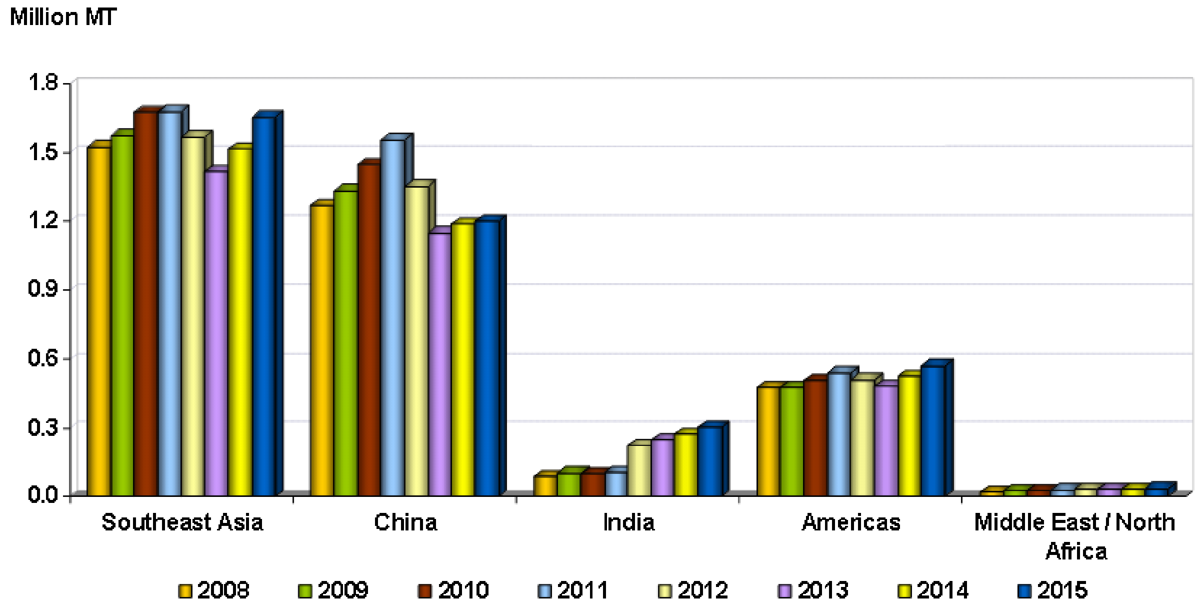


Figure 3 Shrimp aquaculture by major producing regions 2008-2015. Data from FAO (2013) and GOAL (2013), compiled by Anderson, J.L. [2].

However, while there is substantial growth in shrimp aquaculture, it is hampered by the emergence and spread of diseases. Due to the intense farming practices and the movement of live shrimp (broodstock) between farms, diseases spread quickly and thus can cause considerable damage. Over the past years surveys amongst the members of the shrimp aquaculture sector have listed diseases as the number one challenge the industry faces [2]. An example of a major disease affecting the industry is Early Mortality Syndrome (EMS) or Acute Hepatopancreatic Necrosis Disease (AHPND). In the Asian shrimp culture sector this disease has caused over USD 1 billion in annual losses since 2010 [3]. In countries that are struck by diseases (like EMS) the production can drop by 25 – 50 % [2]. For example, an assessment of the impacts of diseases on the Indian shrimp farming sector by Kalaimani *et al.* 2013 estimated the total losses due to disease was around 30 % of total production. These losses were mainly attributed to White Spot Disease (WSD) and Loose Shell Syndrome (LSS). While the shrimp aquaculture sectors continue to be hit by various pathogens, the viruses pose the greatest threat [4]. The most significant viral epidemics are: infectious hypodermal and hematopoietic necrosis virus (IHHNV), Yellow Head Virus (YHV), Taura Syndrome Virus (TSV), and White Spot Syndrome Virus (WSSV) [5].

The most devastating amongst these viruses is WSSV, the causative agent of WSD and which has spread across the full extent of the globe since its emergence in Asia during the 1990s. The global economic damage caused by WSSV to the shrimp aquaculture industry has been estimated at between 8-15 billion USD since its emergence [6], increasing by over 1 billion USD annually [4, 7]. The effect of a WSSV epidemic on individual countries can be very considerable. For example, in 1999-2000 the export of shrimp from Ecuador decreased by nearly 70 % due to WSSV, resulting in the loss of 500,000 jobs [8]. Globally losses are around 10 % of total annual production. Because of its large impact the FAO lists the research into WSSV and development of therapy and/or viral resistant shrimp lines as a top priority [7, 9]. Scientific effort has resulted in increased knowledge on how WSSV operates, but an effective treatment for the disease remains elusive.

Because of economic drivers, research into WSSV is strongly focused on important shrimp species including *Litopenaeus vannamei* and *Penaeus monodon*. However the host range of WSSV is large, and includes all aquatic decapod crustaceans (e.g. shrimps, lobsters and crabs) [10]. To date, ninety eight potential host species for WSD have been identified [10]. Although WSSV causes high levels of mortality in all cultured shrimps, it is not necessarily fatal to other hosts [11]. Variation exists in disease susceptibility across the Crustacea [12]. Bateman *et al.* 2012 classified several European crustacean species according to their response to WSSV exposure, see Table 1.

Table 1 White Spot Disease susceptibility categories and classification of European crustaceans [12].

Susceptibility type	Mortality	Pathology	Species
Type 1 – high	High mortality in both injected and fed exposures	Classic white spot pathology obvious in tissues from both fed and injected exposures	Penaeid shrimp, *Austropotamobius pallipes, Pacifastacus leniusculus, Eriocheir sinensis*
Type 2 – medium	High mortality in injected exposure, little or no mortality in fed exposure	Classic white spot pathology obvious in tissues from injected exposure. Little or no pathology evident in fed exposure	*Homarus gammarus, Nephrops norvegicus, Cancer pagurus*
Type 3 – low	Low level mortality in both injected and fed exposures	Little or no pathology evident in either injected or fed exposures	*Carcinus maenas*

The variation in WSSV susceptibly is most aptly demonstrated in the shore crab *Carcinus maenas*. Studies show that this particular crustacean, while confirmed as susceptible to infection, appears to be especially recalcitrant to development of disease, showing little pathology and low mortality rates [12]. Several hypotheses may explain why *C. maenas* appears to be relatively resistant to the virus, for example variation in the molecular receptors for the virus may make it harder to generate successful infections or *C. maenas* might have an immune system that can cope better with this particular virus (e.g. by expressing unique anti-viral proteins). Given the economic and environmental impact of WSSV understanding the molecular mechanisms underlying the observed resistance in *C. maenas* could be highly informative for understanding the disease process. Should a mechanism of resistance be identified in *C .maenas* there might be potential for translation of this mechanism to understanding the disease process in the more susceptible species. For example: if a miRNA was found to play a key role for resistance in *C. maenas*, then introducing this miRNA to penaeid shrimp species – through feed or genetically modified organisms – could result in similar resistance in those economically important species. The scenario of looking at disease resistance in *C. maenas* represents a novel angle from which to address the WSSV issue as opposed to the more conventional approach of tracking of WSSV infection in susceptible species and identifying molecular targets for potential therapeutics.

As stated above, there has been considerable effort in studying the infection mechanisms and pathogenesis of WSSV. Increasingly, most studies are aimed at clarifying the disease processes at a molecular level. Thus attempting to answer questions like: What are the interactions between host and viral proteins? Are immune system reactions like melanization activated in presence of the virus? Is there increased production of antiviral proteins? Are host transcription factors capable of expressing viral genes and vice versa? How are virions assembled and spread throughout the host organism?

Knowledge of the molecular basis of a disease is often a prerequisite for the design of potential treatments. When important gene or protein players in the disease process are identified a route is opened for experimentation to influence their action, e.g. through small molecular drugs, miRNA or antibodies. Elucidating the molecular basis of diseases require molecular techniques like polymerase chain reaction (PCR) and yeast two-Hybrid (Y2H). In addition, there has been an increase in the application of next generation sequencing to the study of host-pathogen interactions. Next generation sequencing forms a major part of this thesis and thus here the technology and analysis methods are described in detail.

1.2 Next generation sequencing technology and applications

The limited genomic resources for non-model species like *L.vannamei, P.monodon* and *C. maenas* can hinder scientific study on such species. Firstly, the availability of sequence information facilitates the design of new experiments that study the involvement of genes/proteins in an infection process. Popular techniques for gene expression measurements like PCR require sequences on which to base primer design and many molecular biology techniques also involve the action of enzymes that are sequence dependent, e.g. restriction enzymes. And secondly, sequence information can improve interpretation of results since genomic variations can have large impacts on the organism [13]. For instance, single nucleotide polymorphisms can change the amino acid sequence of a genes protein product, change expression patterns through altering promoter activity or change mRNA stability and localization [13]. In short, sequence information is valuable to biosciences and technologies like next generation sequencing that allow generation of sequence data have become of increasing importance in research and the understanding of biological processes.

Next generation sequencing is a technology that can quickly and relatively cheaply unravel genomic or transcriptomic sequences and has become increasingly popular in life sciences. Sequencing technology started with Sanger-sequencing in the 1970s [14]. This technique employed a DNA polymerase reaction combined with dNTPs and fluorescently labelled dideoxynucleotides (ddNTP), the latter terminating the polymerase reaction. From a single strand DNA template products of different lengths are generated, determined by when a ddNTP is incorporated. After size sorting the resulting fragments by gel electrophoresis, the fluorescence can be measured and the original sequence derived. Despite its low throughput and high costs Sanger-sequencing remained the main sequencing technology for many years.

With the advent of next generation sequencing (NGS) technologies the costs of sequencing dropped dramatically (Figure 4). There are currently various NGS technology platforms available e.g. Roche 454, Illumina, SOLiD and Ion Torrent [16-19]. These technologies have similarities in their sample preparation steps, but differ in the chemistry and sequence detection methods.

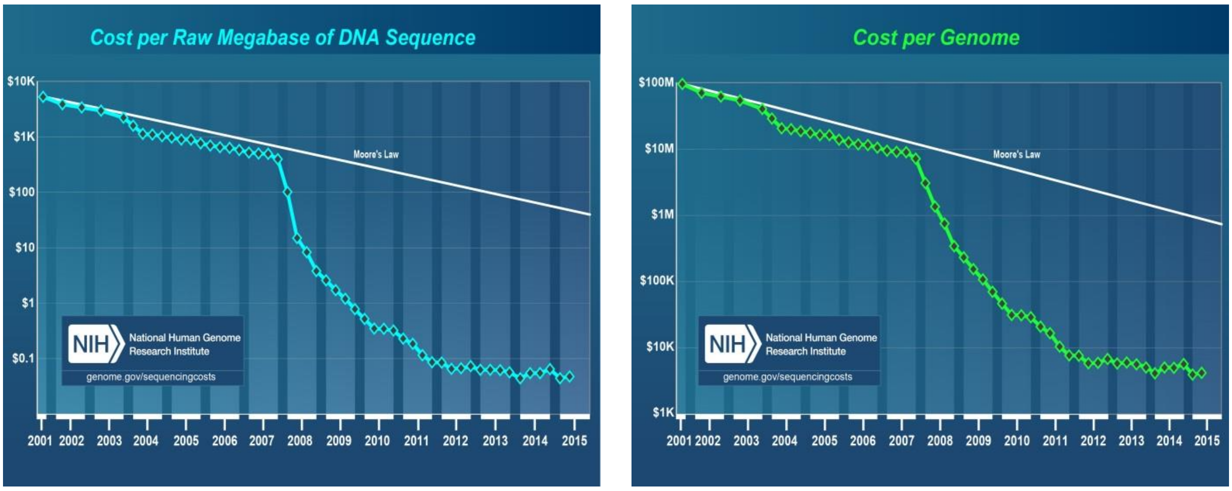
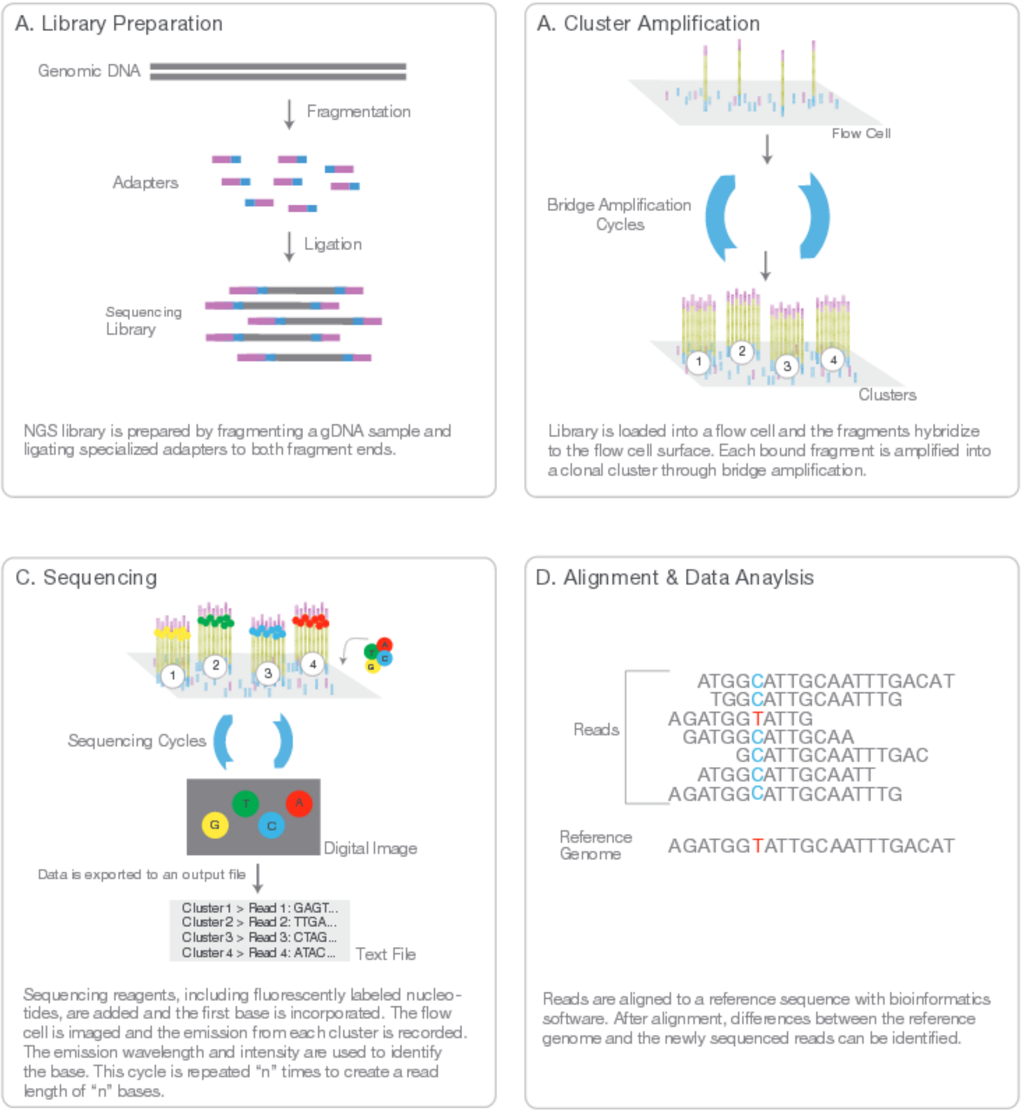


Figure 4 Declining cost of sequencing. The costs of DNA sequencing have fallen over the last decade according to NIH data [15].



NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

Library is loaded into a flow cell and the fragments hybridize to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

Cluster 1 > Read 1: GAGT...
Cluster 2 > Read 2: TTGA...
Cluster 3 > Read 3: CTAG...
Cluster 4 > Read 4: ATAC...

Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated "n" times to create a read length of "n" bases.

Reads are aligned to a reference sequence with bioinformatics software. After alignment, differences between the reference genome and the newly sequenced reads can be identified.

Figure 5 Illumina Next-Generation Sequencing chemistry overview. Overview of the steps involved in next generation sequencing with Illumina technology [20].

Taking genomic DNA sequencing as an example, generally the first step in NGS is to shear the DNA into smaller fragments. After shearing, adapters are added to both ends of the fragments and through the adapters the DNA fragments are attached to a surface (beads in 454 and SOLiD, a surface in Illumina). Next, each fragment is multiplied by PCR reaction based on the adapters. The final stage is to read the sequences through the aid of various methods (See Figure 5).

Most of these platforms employ DNA polymerase to incorporate new nucleotides to the PCR fragments and using various methods to detect this incorporation. In Roche 454 sequencing after a nucleotide is added pyrophosphate is released, converted to ATP and subsequently used by a Luciferase enzyme to emit light. Detection of this light signals when a nucleotide is incorporated and allows reading of the sequence. Illumina sequencing technology differs from Roche 454 in that it employs fluorescently labelled nucleotides, each with a distinct colour. By taking pictures of the surface and tracking the colour sequence the original nucleotide sequence can be derived. Finally Ion Torrent uses a semiconductor measuring H^+ that is released in each DNA polymerase reaction. The SOLiD technology stands apart since it uses DNA ligase instead of DNA polymerase. The substrates for DNA ligase are short fragments of random DNA sequence with a determined central nucleotide, which is labelled with a specific fluorescent dye for every nucleotide type. Again after every ligation reaction the colours are measured which is indicative of the original DNA sequence. One of the drawbacks on the methods described is that they depend partially on PCR. It is inherent that this step will develop biases in the data, in particular when AT/GC rich genomes are concerned. Furthermore the sequences that are obtained are often short in length, encumbering downstream analysis. While the emergence of the NGS technologies has already revolutionized the life sciences, innovation in this field is far from over.

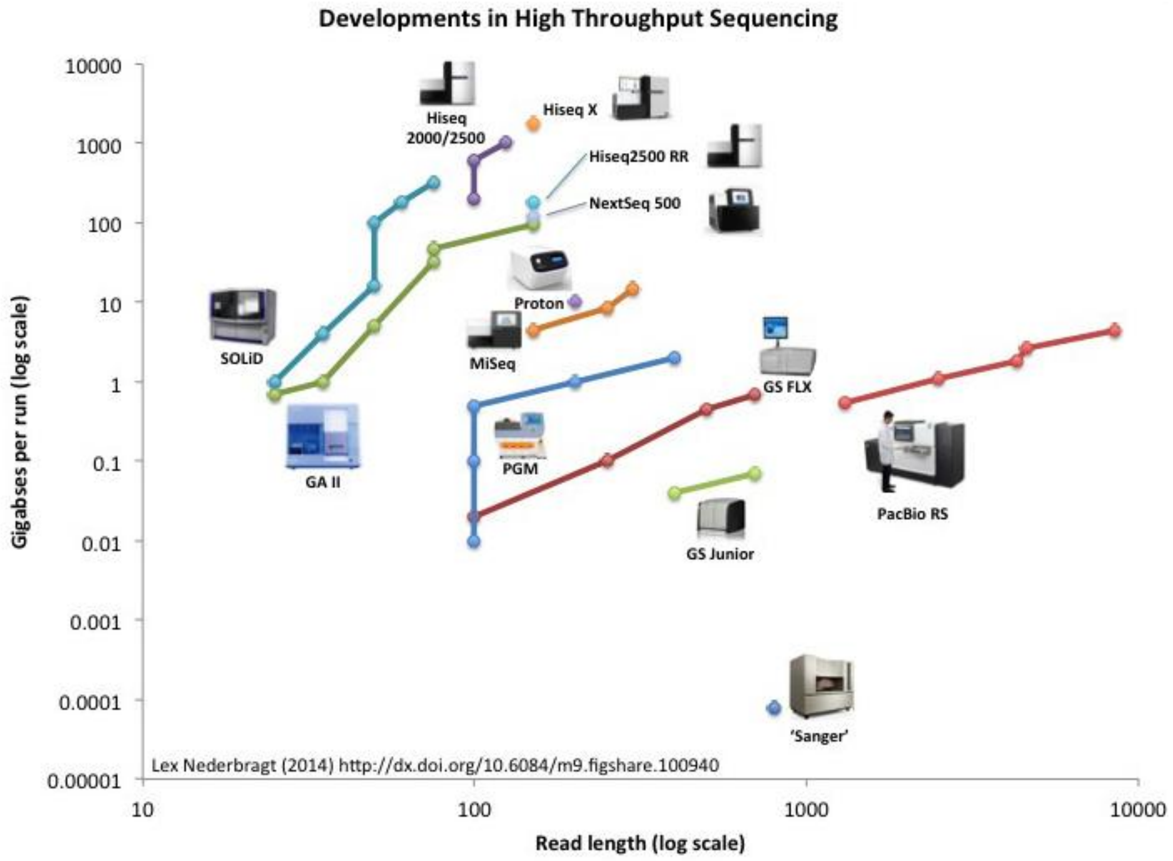


Figure 6 Developments in high throughput sequencing. Lines represent the development of different sequencing over the years, for most systems both length and quantity have increased.

The graph in Figure 6 shows that there now technologies available that exceed even the old Sanger sequencing technologies as far as read lengths are concerned. Compared with the technologies described earlier, sequencing by PacBio has notable differences [21]. PacBio sequencing is based on single molecule read-time (SMRT) technology. A SMRT chip is covered with very small wells that each contains a single DNA polymerase molecule. The DNA sample is loaded onto the chip such that each well contains a single DNA molecule. Then the DNA polymerase incorporates fluorescently labelled nucleotides, enabling the machine to read the sequence. Thus PacBio sequencing does not require a PCR step and theoretically large DNA molecules can be sequenced in a single run, resulting in very long sequencing reads. While these are strong advantages, particularly when considering genome sequencing, they do come at increased cost and reduced sequence yield. Thus currently there is not one technology that clearly excels over all the others and a choice has to be made taken specific project parameters in mind (number of samples, quantity and quality of DNA, monetary resources etc.)

To this point genome sequencing has been the focus, but this is by no means the only application of NGS technology. Several variations exist, each catered to answer a different research question. Table 2 provides an overview of some of the applications of NGS technology in life sciences. Some of these sequences require prior knowledge, e.g. exome sequencing requires an annotated genome and ChIP-seq requires antibodies targeting a protein of interest. But the beauty of NGS that many of the techniques described in Table 2 work without any prior knowledge. This makes NGS the technology of choice for experiments involving non-model species. Through genome/transcriptome sequencing and subsequent *de novo* assembly, a large amount of information can be gathered about said species in exchange for relatively little time and money.

Table 2 Several applications of NGS technology.

NGS technique		**Description**	**Possible applications**
Genome sequencing	DNA	Genomic DNA is isolated from the sample, fragmented and sequenced.	Variant detection, *de novo* genome assembly
Exome sequencing	DNA	Similar to genome sequencing. However fragmented DNA is enriched for exomes prior to sequencing. Enrichment can be achieved via various techniques, e.g. microarray hybridization.	Variant detection – enrichment of exomes accomplishes easy detection of impactful sequence variations.
Restriction site Associated DNA (RAD) sequencing	DNA	Prior to fragmentation the DNA is subjected to a restriction enzyme. After restriction fragments adapters are ligated and used to select only DNA fragments associated with the restriction sites.	Variant detection in populations (genotyping)
Metagenomics	DNA	DNA is not isolated from an individual but rather directly from the environment. The resulting DNA thus represents a snapshot of all the organisms in the sample.	Microbiological diversity analysis
ChIP sequencing	DNA	Proteins and DNA are crosslinked (e.g. by formaldehyde treatment) and sheared. Antibodies targeting proteins of interest are used to enrich the sample. The resulting DNA is sequenced.	Detect protein-DNA interactions
Methyl sequencing	DNA	DNA is pretreated with a bisulfite which converts cytosine to uracil, unless the cytosine is methylated. Subsequent sequencing thus results in information on methylation status of individual cytosines.	Epigenetics

Transcriptome sequencing	RNA	RNA is isolated from a sample. After isolation mRNA is enriched, e.g. by polyT-magnetic beads. A reverse transcriptase reaction generates cDNA from the mRNA which is subsequently sequenced.	Variant detection, gene expression analysis, *de novo* transcriptome assembly
miRNA sequencing	miRNA	RNA is isolated from a sample. After isolation RNA is size selected by running a polyacrylamide gel.	Novel miRNA detection, miRNA expression analysis

Sample preparation and sequencing are the primary stages of a next generation sequencing study. The sequencing run often results in the generation of enormous amounts of data in the form of millions of short sequencing reads. These have to be meticulously put together in order to make interpretation feasible. Because of its size and complexity the analysis of NGS datasets require bioinformatics approaches. In the following section an overview will be provided on the bioinformatics involved with analysis of NGS data, with focus on RNA sequencing analysis and its relevance to this thesis.

1.3 Bio-informatics of *de novo* transcriptome assembly

Broadly speaking there are two strategies for obtaining a transcriptome from RNA sequencing data, the choice depending on the availability of a reference genome. The reference based approach provides power to bioinformatic analysis of the data and improves the reliability because there is a framework in place onto which the reads can be assembled. Transcriptome assembly without a reference genome, *de novo* assembly, is more difficult but advances in bioinformatics algorithms make it possible. Whether the transcriptome assembly is reference based or *de novo*, the basic layout of the bioinformatic pipeline is similar in both situations. Differences mainly involve the choice of software at certain stages of the pipeline.

A general target of gene expression studies is the identification of differentially expressed genes between two sample groups. There is a plethora of possible next generation sequencing pipelines that are able to answer this question and the majority of steps within these can be placed in a small set of categories [22]. The initial step encompasses pre-processing of raw sequencing reads, removing low quality bases and remaining adapter sequences. After pre-processing the reads can either be mapped to a reference genome, or used as input for a *do novo* assembler to produce a transcriptome. With reads mapped against a transcriptome, expression levels for transcripts can be estimated from transcript coverage after normalization has been applied. The final step is to find the differentially expressed genes between the sample groups [22]. In this section an overview of the possibilities within each pipeline stage will be given.

1.3.1 Preprocessing of RNA-seq data

Before a start can be made on transcriptome assembly and/or mapping, reads have to be pre-processed. The main goal of this phase is to remove any low quality information contained in the reads to prevent it from introducing errors in contaminating the rest of the analysis. Most of the RNA-seq analysis pipelines contain similar features in the preprocessing stage. The initial step in most pipelines is the removal of adapter sequences from the reads [23, 24]. This artifact is caused by the chemistry involved in the sequencing process. The UniVec database at NCBI [25] contains sequences that may be of vector origin and can thus be used to identify adapter contamination, but usually the adapter sequences are known through the sample preparation steps. There are many software packages available with the capabilities to remove adapter sequences from the reads, including fastX [26], Trimmomatic [27] and cutadapt [28].

Sequencing machines assign a quality score for every base that is sequenced, usually given by a Phred score (Quality = -10 $\log_{10}$ P, where P corresponds to the probability that the base is called incorrectly) [29]. Lower quality scores are always found at the 3’ end of the reads as illustrated by Figure 7, a typical report of base quality scores according to read position.

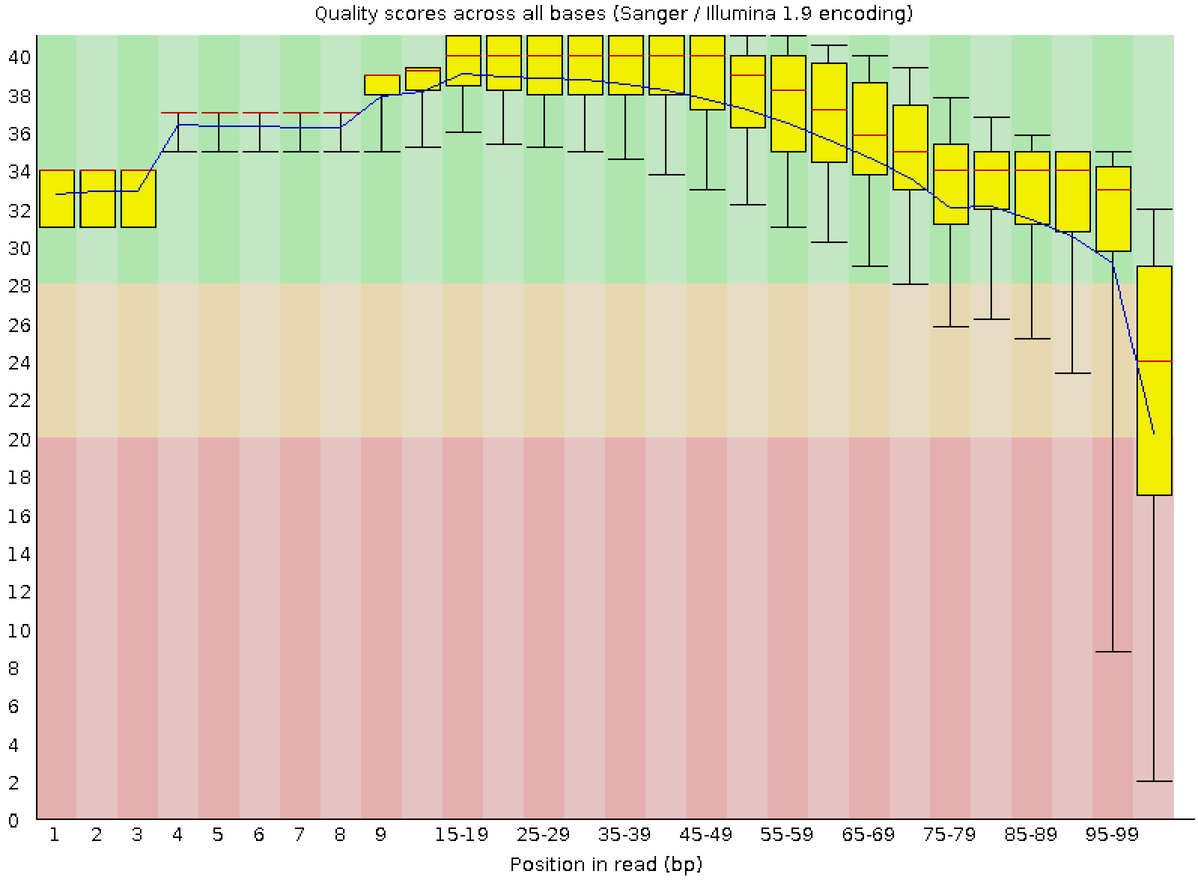


Figure 7 Quality scores across bases. Phred quality scores across all reads in a RNA-sequencing dataset, Illumina HiSeq 2500 sequencing of *Carcinus maenas* hepatopancreas cDNA. Graph generated by fastqc [30] and data from Verbruggen *et al*. 2015 [31].

Filtering of low quality bases reduces the chances of incorporating sequencing errors in the subsequent transcriptome assembly and analysis. In nearly all RNA-seq pipelines qualityscore directed trimming of reads is done, exceptions to this are sometimes made and explained by decisions at later stages in the pipeline [32]. There are multiple ways that quality trimming can be performed. The simplest method is to remove bases once the quality drops below a certain threshold (e.g. below 20 Phred Quality Score). There are more complicated methods available for quality e.g. trimming of bases with the aid of a sliding window; this has the advantage of retaining good quality sequence after a single base drop in quality. The quality score of reads can also be employed to determine whether or not entire reads should be removed from the RNA-seq dataset. Filters like these are easy to apply and can be found in the fastX toolkit [26] and Trimmomatic [27].

Besides the quality scores, certain reads can also be filtered out on the basis of the sequence itself. A large fraction of the RNA isolated from a sample will consist of ribosomal RNA (rRNA). Because RNA-seq experiments are often designed to answer questions regarding gene expression, some researchers decide to remove rRNA reads from the dataset at this stage to reduce data quantity [33]. However, most library construction kits contain a step that isolates only RNA molecules with a polyA-tail from the sample prior to sequencing. Thus the quantities of rRNA, tRNA, miRNA, ncRNA and others in the sample are reduced. Another sequence filter consideration might be to remove reads containing Ns. The reason to do this is that some software in the *de novo* assembly portion of the pipeline will replace N bases with another base in order to conserve memory space; Velvet is an example of this [34].

In short, the objective of the pre-processing stage is to remove read information that can have a negative impact on the rest of the analysis pipeline. Helpful software or scripts performing most of the pre-processing steps can be found in existing software packages like fastqc [30], fastx [26], RobiNA [35] and Trimmomatic [27]. If RNA-seq is being done on the Roche 454 platform, the Newbler [16] or LUCY [36] software packages also have quality control capabilities. Most of these tools can also be found on the Galaxy platform [37-39].

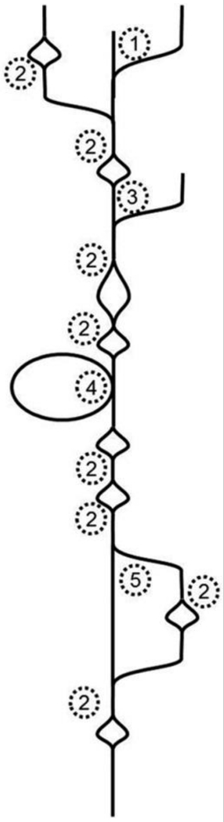
1.3.2 Transcriptome assembly and/or read mapping

After the preprocessing stage the information contained in the individual reads has to be combined and assembled into a valid transcriptome. Where the preprocessing stage is fairly similar in most transcriptomics research, the assembly and mapping stage is where many analysis pipelines diverge. The software selection for this stage in the pipeline depends on the sequencing platform and availability of a reference genome/transcriptome. When a genome sequence is available, reference based assembly is nearly always preferred.

1.3.2.1 de novo assembly

De novo assembly of a transcriptome from single or paired end reads is a computational problem that has been tackled by a large variety of assemblers. Assemblers can be split into two types: overlap-layout-consensus assemblers and de Bruijn graph based assemblers. Overlap-layout-consensus assemblers operate with a basic strategy: the assembly process starts off with a read and searches the remaining reads in the dataset for similarity. Similar read sequence will be added to the contig and so on until all the possible contigs are constructed. De Bruijn graph based assemblers first split up reads in to fragments of a predefined length, called k-mers, and constructs a graph from these k-mers. A node in this graph thus represents a sequence of k length and a vertex between two nodes represents a significant overlap (k-1) between the sequences of two nodes. After the graph has been constructed the assembly algorithms will search the graph for an Eulerian path that traverses it. Ideally the graph will be a series of straight lines however sequencing errors, insertions, deletions, alternative splicing events etc. will form bubbles and forks in the graph (Figure 8)

[24]. The assemblers' algorithms determine which of these bubbles and forks are formed by either genuine sequence variation or sequencing errors.



1. Alternative transcription start site or hybrid joining or DNA contamination

2. SND caused by a sequencing error or a SNP or mutation after gene duplication

3. Alternative transcription start site or DNA contamination

4. Alternative exon use

5. Alternative exon use or mutations after recent gene duplication

Figure 8 De Bruijn Graph with sequence variations. A De Bruijn Graph is seldom a set of linear lines. Sequence variations and sequencing errors cause bubbles to form which have to be resolved by the assemblers. Figure from Schliesky *et al.* 2012 [24].

Choosing between overlap-layout-consensus and de Bruijn graph based assemblers largely depends on the sequencing platform that can be used. The crux of the choice lies within the number and length of reads produced by the platform. Overlap-layout-consensus algorithms won't work well with large datasets due to computational demands. Therefore, it is more suited for sequencing platforms like 454 that produce a lower quantity of reads but of longer length (however the size of current 454 datasets still makes them unworkable). Due to the volume of data created by short read sequencers like Illumina technology, overlap-layout-consensus is not used because they are impractical and a De Bruijn graph approach is preferred.

There are numerous assemblers available for both *de novo* assembly protocols. Examples of overlap-layout-consensus assemblers are Newbler [16], Edena [40] and CAP3 [41]. Newbler is a very popular choice for long read RNA-seq datasets (Roche 454) that require *de novo* assembly [42-46] .CAP3 is sometimes employed to merge assemblies that have been created with other algorithms (e.g. [47]), a process that will be discussed later. Edena, although an overlap-layout-consensus assembler, was used to produce a transcriptome with short read Illumina data [47].

De Bruijn Graph Assemblers directed towards Illumina sequencing data have also been used in a wide selection of RNA-seq experiments. Examples of De Bruijn Graph genome assemblers are include Velvet, ABySS and SOAPdenovo. Because of an important difference between genomic sequencing and RNA-seq, adaptations to these genomics assemblers were made to enable transcriptome assembly. Genomics assemblers generally assume similar coverage of the entire genome, whereas in the transcriptome the coverage of a transcript depends on its expression level in the cellular system. Because of the large range in expression dynamics, the assembler cannot compare coverage between different transcripts. The read coverage of forks and bubbles in the De Bruijn Graph (Figure 8) is important to distinguish read sequence variations from sequencing errors. The adapted genomic assemblers, now *de novo* transcriptome assemblers are: Velvet/Oases, Trans-ABySS and SOAPdenovo-Trans [24]. Trinity is an assembler that has been developed with RNA-seq in mind and has grown to be the most popular *de novo* assembler. Table 3 list popular assemblers and the number of times they have been cited (data from January 2016).

Table 3 De Bruijn graph based assemblers

Genome assembler	**Reference**	**Web of Science citations**	**Transcriptome assembler**	**Reference**	**Web of Science citiations**
Velvet	[34]	2,701	Velvet / Oases	[48]	380
AbySS	[49]	867	Trans-ABySS	[50]	236
SOAP-denovo2	[51]	323	SOAP-denovo-trans	[52]	51
			Trinity	[53]	1,809

A study was done by Zhao et al 2011 [32] comparing the performance of a series of De Bruijn Graph assemblers on multiple Illumina RNA-seq datasets and provides some pointers on how to choose an assembler that fits the trade-off between computational demands and characteristics of the generated transcriptome. De Bruijn Graph based assemblers have also been tested on Roche/454 sequencing data [54]. The result suggests that Trinity has the best performance of the De Bruijn assemblers and is comparable to Newbler.

There have been explorations of using assemblers in successive rounds of assembly. The output generated by one assembler, a series of contigs and singletons, is used as input for another round of assembly. Coppe et al. 2012 [55] used a two stage assembly to produce a transcriptome for the striped venus clam (*Chamelea gallina*) from 454 sequencing data. At both stages MIRA [56] was the chosen assembler, and the second assembly run had more stringent criteria. The authors supported that by using two subsequent assemblies the redundancy in contig sequences can be reduced. Tao et al 2012 [47] used several assemblers, the output of three of these assemblers (Edena, SOAPdenovo and Velvet all at various settings) had been used as input for a secondary assembly run using the overlap-layout-consensus assembler CAP3, the final assembly outperforming Trinity on their data. Generally *de novo* studies are performed using a single assembler to put the data together.

The assemblers in Table 3 are/have been the most popular assemblers to date. Some of the publications the table are several years old, but innovation has not stopped in the *de novo* assembler field. Most existing assemblers receive updates and new assemblers are still being created. Some of the latter are generated for specific purposes, e.g. the EPGA2 assembler is memory efficient [57] and the IVA focuses on RNA virus genome assembly [58]. For transcriptomes the Bridger assembler combines algorithms from Trinity with those of reference based assembly algorithms [59].

1.3.2.2 Reference based assembly

For some species a genome sequence is available, a valuable source of information that can improve the quality of a transcriptome assembly. To use the framework of a fully sequenced genome in transcriptome assembly requires a different set of algorithms than the ones that have been discussed so far. The strategy of reference base transcriptome assemblers can be summarized as follows: first the individual reads are mapped to the reference genome and after that mapped reads are combined into transcripts. Mapping of reads to a reference genome can be done by several programs, including bowtie2 [60] and BWA [61], both able to work with short reads. TopHat [62] is an extension of bowtie that is able to map reads across splice junctions. After processing of raw reads by a BWA, bowtie or TopHat, the mapping information has to be interpreted to find out from which transcripts the reads originated. This task would be straightforward were it not for alternative splicing events, which makes it possible for reads mapping close to each other on the genome to have originated from different transcripts.

Cufflinks [63] is a popular software package that is able to deduce a minimal set of transcripts that is able to explain the read alignment produced by the read mapper. It does this by creating an overlap graph for each bundle of aligned reads. In the overlap graph nodes represent a sequenced fragment (Single or Paired end, although Cufflinks is designed with paired end data in mind) and a vertex represents compatibility between the two fragments. Compatibility means that it is possible for the two fragments to have originated from the same transcript. From the overlap graph the isoforms can be identified. Cufflinks achieves this by finding the largest group of mutual exclusive fragments and extends this information to find the minimal set of transcript isoforms able to explain the alignment (an implementation of a proof of Dilworth's Theorem). Cufflinks is also able to deduce the expression levels of transcripts; this will be discussed in a later section.

TopHat and Cufflinks, together part of the Tuxedo suite, is a very popular combination of tools and has been used excessively in reference-based RNA-sequencing experiments. The protocol paper by Trapnell et al. 2012 [64] has over 1000 citations on web of science. The suite has been extended further by incorporating cummRbund which can automatically draw publication style graphics for the data. The effectiveness of this suite was demonstrated by Ghosh et al. 2016 [65].

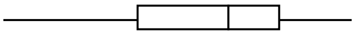
1.3.3 Clustering and filtering of transcriptomes

After assembly, *de novo* in particular, it is likely that not all the produced transcripts are useful for further analysis. Often redundant transcripts, small transcripts or even derived transcripts from contaminating species can be found in transcriptomes. In order to tackle these issues, many analysis pipelines apply clustering and filtering steps. Redundancy in transcripts can be reduced by grouping transcripts together when they fulfil certain similarity criteria and having a single representative for each group, CD-HIT-EST can do this procedure fast and efficiently for large datasets [66]. Transcript filtering can be done to varying degree of complexity. The most straightforward filter is to remove all contigs smaller than a certain threshold from the transcriptome. The size threshold can be set in multiple ways. Some groups decide on a hard threshold, e.g. 100 bp [67] or 200 bp [55]. It is also possible to let the threshold depend on the settings used in the assembly. In de Bruijn Graph assemblers the k-mer size is an input parameter that regulates how large the sequence pieces are that will eventually generate a contig. To filter contigs it can be a choice to let the size threshold depend on the k-mer size. For example: the length of a contig must at least be twice the k-mer size [23]. Removal of contaminating species can only be done after annotation because one requires an indication of the species of origin. Annotating transcripts through BLAST against databases like NCBI nt/nr provides information on the species of origin. Tools like MEGAN [68] provide an overview of transcriptome taxonomy and can be used to remove undesired transcripts. In case a mammalian transcriptome is assembled, a choice can be made to remove all transcripts that show a high degree of similarity to bacterial or viral sequences which can be considered contaminations [31].

1.3.4 Normalization and expression

While a produced transcriptome for a species can be a satisfactory result by itself, it is often part of a study investigating differentially expressed genes between two or more experimental groups. Because abundant transcripts produce more sequencing reads than rare transcripts, RNA-seq can be used to derive gene expression levels for each transcript in the transcriptome. This principle sounds intuitive, but there are subtleties that require attention. For example, every sequencing lane can differ in the total number of sequencing reads produced, therefore a comparison between samples cannot be done simply by mapping and counting reads, and normalization steps are required. Many types of normalization have been used and some have been analysed in comparative studies. The start of all the normalization methods is to map reads to the transcriptome. This can be done with any aligner such as bowtie [60], BWA [61] or SSAHA2 [69]. The read alignment produced by the mapping algorithm can be processed by several normalization methods. A short description on a diverse selection of normalization methods is provided in Table 4.

Table 4 RNA-seq normalization methods

Normalization method	**Explanation**
	Transcripts: i in 1 ... I Lanes/samples: s in 1 ... S Read Count: RC (RC' is normalized read count) Library Size: LS Median/upper Quartile: MQ UQ 1  max (RC) Fold change: $M_i = \log_2 (RC_{i,s} LS_s / RC_{i,s'} LS_{s'})$ Intensity: $A_i = \frac{1}{2} log_2 \left(\frac{RC_{i,s}}{LS_s} \bullet \frac{RC_{i,s'}}{LS_{s'}} \right)$ Probability distribution for distances (PE): P(d)
Total Count [22]	$$RC'_{i,s} = \frac{RC_{i,s}}{LS_s} mean(LS)$$ The gene read count in each sample is divided by the library size from which the sample originated and multiplied with the average library size across all lanes.
Upper Quartile [22]	$$RC'_{i,s} = \frac{RC_{i,s}}{UQ_s} mean(UQ)$$ The gene read count in each sample is divided by the Upper Quartile for that sample and multiplied with the average Upper Quartile across the samples. The upper quartile is the count that separates the top 25% counts and bottom 75% counts in the count distribution. Genes with a read count of 0 are neglected.
Median [22]	$$RC'_{i,s} = \frac{RC_{i,s}}{MQ_s} mean(MQ)$$ Similar to upper quartile normalization. The upper quartile is replaced by the median of the gene counts. Read counts of 0 are neglected.

DESeq [22, 70]	$RC'_{i,s} = \frac{RC_{i,s}}{s_s}$ $s_s = median_i \frac{RC_{i,s}}{\left(\prod_{s=1}^{s=J} RC_{i,s}\right)^{1/J}}$ A scaling factor for a lane is computed as the median of the ratio for each gene of its count over a pseudo-reference lane. The pseudo-reference lane is computed by the geometric mean across all the samples. All read counts for genes in a lane are divided by the scaling factor calculated for that lane to obtain the normalized gene count. *Assumption*: most genes are not differentially expressed
Trimmed mean of M values [22, 71, 72]	$log_2\left(TMM_s^{(r)}\right) = \frac{\sum_{i\in I} w_{i,s}^r M_{i,s}^r}{\sum_{i\in I} w_{i,s}^r}$ $M_{i,s}^r = \frac{log_2\left(\frac{RC_{i,s}}{LS_s}\right)}{log_2\left(\frac{RC_{i,s}}{LS_s}\right)}$ $w_{i,s}^r = \frac{LS_s - RC_{i,s}}{LS_s\, RC_{i,s}} + \frac{LS_r - RC_{i,r}}{LS_r\, RC_{i,r}}$ One lane is chosen as a reference and the others are test lanes. Before calculating a scaling factor, several genes are omitted from the calculation. The genes with the largest fold change between reference and test lane (M, default is top 30%), genes with intensity (A, default is top 5%) and genes with a count of 0 are trimmed. The TMM scaling factor is calculated as shown in the formula above as a weighted mean of log ratios between this test and the reference. The weights are the inverse of the approximate asymptotic variances, see [72] for details. Normalized read counts are calculated by dividing the counts by the scaling factor, as was done in the DESeq method.

RPKM (FPKM)	$$RC'_{i,s} = RPKM_{i,s} = \frac{RC_{i,s}}{L_i LS_s}$$ Reads per kilobase per million mapped reads (RPKM) or Fragments per kilobase per million mapped reads (FPKM) for paired data uses gene length and library size normalization. The normalized gene count equals the gene count divided by the length of the gene L_i in kb and the library size in millions.
NEUMA [73]	FPKM is replaced by FVKM (number of fragments per virtual kilobase per million fragments) in Normalization by expected uniquely mappable area (NEUMA). Below is the explanation for Paired End reads only because the single-end situation is more basic. The first step in NEUMA is to derive how many reads can be mapped uniquely to a gene. The information is found by collecting all the possible reads from a transcriptome of a certain size and distance and designating them as either uniquely mapping or not uniquely mapping. For every gene the number of uniquely mapping reads is counted and stored in a dxI matrix U. Next $EUMA_i$ is calculated by multiplying the probability of a certain distance d in a paired end read to the number of uniquely mapping reads for that distance for gene i. $$EUMA_i = \sum_d P_s(d)\, U_{d,i}$$ With $EUMA_i$ known, $FVKM_{i,s}$ can be calculated like RPKM with $GC_{i,s}$ replaced by $NIR_{i,s}$ (the number of uniquely mapping reads for gene i in the experimental data) $$RC'_{i,s} = FVKM_{i,s} = \frac{NIR_{i,s}}{EUMA_i\, LS_s}$$ Note: NEUMA can also be calculated at transcript level instead of gene level.

RSEM [74]	RNA-Seq by Expectation Maximization (RSEM) uses a maximum likelihood estimation to derive transcript expression levels in RNA-seq data. The statistical model can be applied to both single end and paired en reads and is described in Li et al. [74], an elaborated version of the model proposed in Li et al. [75]. This model encompasses read sequence (R), orientation (O), starting position (S) and isoform (G). The extended model also includes quality, paired end information and allows for variable read length. Given a dataset r with N reads and a transcriptome with M isoforms this results in the following data likelihood function: $$P(r \mid \theta) = \prod_{n=1}^{N}\sum_{i=0}^{M}\theta_i P(r_n \mid G_n = i)$$ Derivation of $P(r_n \mid G_n = i)$ is provided in Li and Bewey 2011 [74] and is too elaborate for this section. The parameters to be estimated describe the expression levels of the M isoforms in the transcriptome. $$\theta = [\theta_0, \theta_1 \dots, \theta_M]$$ Where θ_i represents the expression level of isoform i, and θ_0 the isoform used for unmappable reads. The final result is to convert the estimated expression levels (θ) to a *fraction of transcripts* (τ_i)*/ nucleotides*(ν_i): $$\nu_i = \frac{\theta_i}{1 - \theta_0}$$ $$\tau_i = \frac{\nu_i}{l_i}\left(\sum_j \frac{\nu_j}{l_j}\right)^{-1}$$ Estimation of θ with maximum likelihood is done with an EM algorithm.

Quantile [22, 76]	Data is arranged in an IxS matrix D, the columns representing lanes. From D a ranked matrix is made D_{sort}, sorting each column according to counts. The means of the rows in D_{sort} are calculated and assigned to $D*_{sort}$. Finally, the normalized data is formed in D' by rearranging the columns in $D*_{sort}$ by the ranking in D_{sort}.
Housekeeping genes	Certain genes are assumed to have a constant expression over different samples and can therefore be used to adjust for library size effects.

The selection of normalization methods described in Table 4 can be placed in groups according to their characteristics [22]. These groups provide an easier overview of the available methods and can facilitate the decision making process for researchers. The first group are the 'distribution concept' methods that assume a similar read count distribution as principle and include: Total Count, Upper Quantile, Median, Quantile and RPKM-like methods [22]. The read distributions between lanes can be normalized by a single quantile (total count, upper quantile or median) or multiple quantiles (Quantiles). The RPKM-like methods on the other hand apply both a total count and length normalization. The RPKM-like methods are RPKM, RSEM and NEUMA. NEUMA differs from the others by the definition of the length of a gene since only unique mappable areas are considered. Normalizing read count for a gene or transcript by its length enables a better comparison between the expression levels of different genes, even in the same sample. However when only pairwise comparisons are made, e.g. in differential expression studies, it is not necessary to perform length normalization. There are indications that normalization of read counts by transcript length can introduce biases affecting low abundance transcripts [22, 77]. The second group includes TMM and DESeq normalization. In these two methods the assumption that most genes are not differentially expressed leads to an estimation of the library size.

Dillies et al. 2012 [22] presents a comparison study of seven normalization methods, testing both on real and simulated data. The tested normalization methods included Total Count, Upper Quartile, Median, DESeq, Trimmed mean of M values, Quantile and RPKM. Performance was tested based on qualitative characteristics of the raw and normalized datasets and a differential expression analysis. Comparing the read count distributions before and after normalization showed that high-count genes and the number of 0 counts can affect the calculation of the scaling factors. Total Count and RPKM did not stabilise the read count distributions, probably due to the influence of the high-count genes. Calculating the coefficient of variation of normalized read counts for a set of housekeeping genes showed that DESeq and TMM methods had the lowest coefficients. Analysis on simulated data, with varying library size difference and presence of high count genes, showed that high count genes have a strong influence of the false positive rate for Total Count, Upper Quantile, Median, Quantile and RPKM methods. Only DESeq and TMM were able to control the false positive rate while retaining the power to detect differentially expressed genes. These two normalization methods are recommended by Dillies et al. over the popular RPKM and Total Count methods [22]. Methods like DESeq and TMM are included in software packages and can be implemented with only a few commands and have therefore become popular choices.

1.3.5 Differentially expressed genes

After properly normalized gene counts have been derived for every sample comparisons can be made between gene expressions in different conditions present in an experiment. Genes that are differentially expressed (DE) can provide insight on how an organism or cell line responds to environmental stimuli. Deriving the list of DE genes involves calculation of the statistical significance of the dissimilarity in gene expression amongst the conditions.

Methods to perform these types of calculations have already been developed for microarray data. However, these are not directly translatable to RNA-sequencing data because of inherent differences in the datatype. Microarray data is continuous whereas the gene counts derived from RNA-sequencing are discrete. This means that the distributions underlying significance calculations also have to be discrete. Examples of discrete probability distributions are: the Poisson distribution, the geometric distribution, the binomial distribution and the negative binomial distribution. Several DE analysis methods are based on these distributions. The fundamental statistics underlying the described methods will not be elaborated; rather a brief description on the basic principles is given.

Fang et al. 2012 [78] discusses several methods for the calculation of DE genes in RNA-seq data. All of the methods are parametric, meaning that assumptions are made on the underlying data distributions. The Poisson distribution can be used to represent the probability distribution for read counts. Using the Poisson distribution as a basis, DE assessment can be done in several ways: Fisher’s exact test [79] and through a likelihood ratio test [80]. Fisher’s exact test requires the construction of a 2x2 contingency table and can therefore only be used in experiments with two conditions. The contingency table for a gene contains the read counts for that gene and the remaining number of reads. Fisher’s exact test can derive a p-value for the association between the gene and the condition, based on the odds ratios found in the contingency table. Asides from Fisher’s exact test, a likelihood test is also a possible choice for the Poisson model. In a likelihood test the likelihood of the data given different models is used to derive DE transcripts.

The Poisson distribution implies that the variance is equal to the mean. However in RNA-seq experiments it has been shown that the observed variance is often larger than the variance as predicted by the model, this is called overdispersion. In such cases the merits of the Poisson distribution in DE analysis decrease and different models have to be applied in its stead. A possible method to deal with increased variance is to use a two-stage Poisson model, as described by Auer et al. 2011 [81]. In such a process, genes are split into two groups: genes likely to have an increased variance and those without overdispersion. Then genes without overdispersion are tested for differential expression with a standard likelihood test as described above. For genes with over-dispersion an estimate is made on the over-dispersion and used to calculate a quasi-likelihood statistic. In situation with likely over-dispersion it is also a possibility to resort to methods that are based on other distributions; several have been suggested and developed into R packages.

edgeR is a popular R-package that can be used for the detection of differentially expressed genes [71]. It is based on the negative binomial (NB) distribution. Because the NB distribution allows the variance and mean to differ, it can deal with over-dispersion in the dataset. In edgeR dispersion for a gene is estimated by a conditional maximum likelihood dependent on the total count for that gene [71]. If the dispersion approaches 0, the negative binomial distribution reduces to a Poisson distribution. Like the Poisson distribution based methods described above, the DE genes for a pairwise comparison are gathered from the data by using an exact test. edgeR employs a likelihood ratio test to derive the DE gene list when multiple comparisons are considered. There are other R-packages available that offer extended methods to the NB edgeR methods. DESeq is another well-used package and treats the relationship between mean and variance differently [70]. Instead of a conditional maximum likelihood estimation of the NB dispersion parameter, this estimation is performed in a hierarchical manner. Cufflinks, discussed above, is also able to generate a list of DE genes from a dataset and uses the same method to deal with overdispersion as DESeq [82]. Cuffdiff is the part of Cufflinks responsible for the calculations. Differential expression is determined through a Student’s T test, with overdispersion in the variance modelled by a beta negative binomial distribution. Cuffdiff is also able to provide changes in relative transcription between splice variants of the same gene.The baySeq R-package is another modification of edgeR that allows for multiple comparisons [83]. For every comparison a posterior probability for DE of a gene is calculated based on a Bayesian approach. The overall chance of DE of a gene will be the sum of all calculated posterior probabilities. EBSeq is also a Bayesian package that has improved power and performance to identify DE isoforms in particular, it is often coupled with RSEM in analysis pipelines [84].

A beta-binomial distribution has been used by Zhou et al. 2011 [85] to model overdispersion gene counts in the BBSeq R- package. Specifically, the probability that a read in a sample will map to a gene is modelled by a parameter that follows a beta-binomial distribution. The model for over dispersion can either be a free model, meaning that the dispersion can be calculated for every gene separately, or a constrained model wherein the mean-overdispersion relationship is assumed. BBSeq uses a design matrix, with phenotype indicators, and a logistic model to derive a Wald statistic for every gene, indicating whether or not it is differentially expressed [78].

Overall the edgeR, DEseq and EBSeq appear to be the most popular in *de novo* RNA-seq studies and Cuffdiff for reference based RNA sequencing.

1.3.6 Annotation and pathway analysis

Annotation and pathway analysis are methods by which context can be added to a transcriptome and/or list of DE transcripts. By adding context, the biological interpretability of the RNA-seq data can be increased. Annotation of the transcriptome itself could be considered as a quality statistic, increasing numbers of annotated transcripts can provide confidence in the chosen parameters of the transcriptome assembly. The presence of highly conserved genes/proteins in a transcriptome is a good indicator of assembly quality. Indeed, tools like cegma [86] have been developed that interrogate transcriptomes for the presence of a selection of such highly conserved sequences.

BLAST can be used in annotation by similarity of transcripts in a transcriptome to reference databases. When no species specific reference database is available, it can be considered to use a related species. For example, Wolf et al. 2010 [42] compared crow transcriptomes to zebra finch and chickens. Most pipelines, of data coming from a wide variety of species, include a BLAST search against the NCBI non-redundant database to annotate transcriptomes from a large variety of species. After identification of a significant sequence similarity, the annotation of the subject can be transferred to the query transcript. After a transcript has been linked to a gene, various online databases can be used to provide further annotations, e.g. function annotations. Gene Ontology (GO) is a functional annotation database with a hierarchical structure and a standardized vocabulary [87]. GO annotations are linked to proteins, and are available for multiple species. If the subject species is not available in the GO database, homology can be used to link annotations to transcripts. The tool of choice for this procedure is BLAST2GO which uses BLAST to search for significant sequence identity and subsequently links GO terms to the queries. Aside GO, another popular functional annotation is through KEGG [88]. KEGG is a database containing various pathways and can be queried to identify pathway components in a transcriptome, as in the example provided in Figure 9.

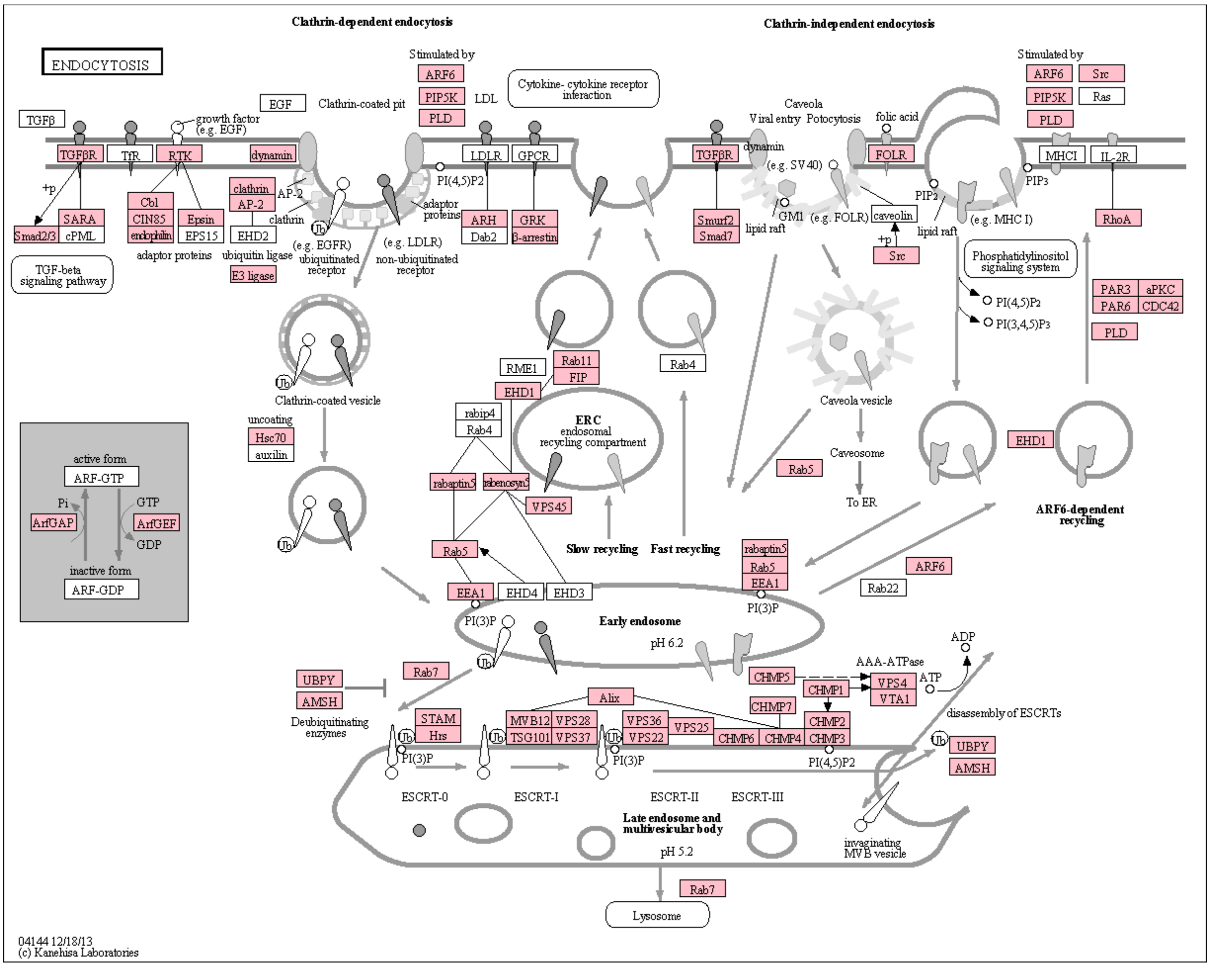


Figure 9 KEGG endocytosis pathway. Components of the KEGG pathway that have been identified in a shore crab transcriptome [31] (Chapter 3).

Lists of DE genes obtained from global expression studies can remain difficult to interpret. In such a situation functional enrichment analysis is frequently used to highlight certain pathways and processes. A standard technique for discovering the overrepresented annotations within a list of DE genes is to perform an enrichment calculation based on a hypergeometric distribution, which will return a p-value for every annotation. Such analyses have been widely applied on microarray studies coupled with GO annotations and can be transferred to RNA-seq data. GOseq and BLAST2GO are applications that have been developed to calculate GO enrichment specifically for RNA-seq data, compensating for possible biases [89, 90]. Enrichment calculations can be done for the annotation of choice, like KEGG pathway membership. Enrichment calculations on a custom selection of genes can be done with Gene Set Enrichment Analysis (GSEA), given that the gene set has been defined *a priori* [91]. Commercial software like IPA [92] and Pathway Studio[93] can perform similar analyses on their databases and couple the results with powerful visualizations.

1.4 Research undertaken in this thesis

The work described in this thesis is aimed at the study of WSSV infection in aquatic crustaceans. As mentioned above, most of the research into WSSV has been performed in shrimp species like *L. vannamei* and *P. monodon* and thus far this has not led to development of a successful treatment. The observation that *C. maenas* is relatively resistant to the virus opens up a new angle of study. Gathering information on this host organism and tracking its response to viral exposure opens the opportunity to identify the molecular basis for the apparent resistance of *C. maenas* to WSSV. Once the molecular basis is identified, a novel treatment against WSSV could be designed on this basis. For example: if *C. maenas* produced a miRNA with significant impact on WSSV infection then this miRNA may do the same for economically important species like *L. vannamei* and *P. monodon*.

Shrimp species are very important in respect to the global aquaculture market, but they are not as relevant from a European perspective since the European aquaculture sector is small. However Bateman *et al.* 2012 showed that WSSV can infect crustaceans in temperate regions, such as European waters, and thus there is potential for the virus to impact aquatic crustaceans in European waters [12]. Molecular knowledge of commercial European aquatic crustaceans will be useful for comparative studies and other future applications. The European lobster *Homarus gammarus* was selected for study in this thesis work in addition to the shore crab because it is a commercially fished species, particularly in the United Kingdom. It is furthermore, more susceptible to WSSV compared to the shore crab, providing a contrasting model for study of WSD.

C. maenas and *H. gammarus* do not have a large amount of available molecular information. Therefore next generation sequencing is particularly suited for viral exposure studies in these species. Through DNA and (small)RNA sequencing experiments a large amount of information can be gathered for these species and used to form draft genomes, transcriptomes and analyses on expressed miRNAs. This basic information then provides a background against which changes through exposure to WSSV can be tracked and potentially important targets in the disease infection/resistance processes identified.

1.4.1 Experimental species

Two aquatic crustacean species were used in the experiments described in this thesis, the shore crab *C. maenas* and the European lobster *H. gammarus*. The shore crab was selected primarily because of its unique resistance to WSSV exposure, and the lobster both because of its important to UK fisheries and the fact that it is more susceptible to WSD compared with the shore crab.

Carcinus maenas

The European shore crab (or green crab), *C. maenas*, is a keystone species in the European marine environment. It has a carapace length of up to 60 mm with a width of up to 90 mm. It is an invasive species which has spread from its native waters in Europe into Australia, South Africa and the United States. *C. maenas* is a molluscan predator [94] and thus it can threaten local fishing industries in areas it invades. For example: in New England the shoft-shell clam (*Mya arenaria*) fisheries were destroyed by *C. maenas* [95]. *C. maenas* is also an important study species for biomonitoring and ecotoxicology [96, 97]. The species has been used in monitoring for heavy metal contamination [98], metal toxicity studies [99], and more recently in exposures studies with nanomaterials [100] and microplastics [101]. Despite small exceptions, *C. maenas* is not commercially fished or cultured.

Homarus gammarus

The European Lobster, *Homarus gammarus*, is a highly valued seafood commodity. It has a maximum total body length of about 60 cm, with large specimens usually 23 to 50 cm long, and a weight of 5 or 6 kg. The flavour of its meat is held in high regard by consumers, enabling this species to yield high prices on the market [102]. It is commercially fished across Europe, yielding an average of 4972.5 tonnes between 2010 and 2013 [103]. Lobster fisheries are concentrated around the United Kingdom, which accounted for 65 % of total capture (2013) [103]. *H. gammarus* is grown in aquaculture, but not widespread due to issues regarding cannibalistic behaviour, operating costs and density [104].

1.4.2 Thesis roadmap

As stated the overall aim of the project is to identify the molecular mechanism of *C. maenas* apparent resistance to WSSV infection. The work is centred on using bioinformatic analyses to produce a transcriptome and draft genome for *C. maenas*. Additionally a transcriptome for *H. gammarus* is generated. Lastly, an experimental infection study with WSSV in *C. maenas* was undertaken and the molecular responses, both mRNA and smallRNA, were analysed.

The thesis contains 6 further chapters with the following remit and content:

Chapter 2 –Research paper 1 (Published)

Molecular Mechanisms of White Spot Syndrome Virus Infection and Perspectives on Treatments

Bas Verbruggen , Lisa K. Bickley, Ronny van Aerle, Kelly S. Bateman, Grant D. Stentiford, Eduarda M. Santos and Charles R. Tyler. *Viruses* **2016**, 8, 23

This review paper provides a critical analysis of WSSV*.* The review analyses the available knowledge on the WSSV infection process and includes information on the WSSV genome, current known molecular interactions between WSSV and host, and host signalling pathways important to successful infection. The review concludes with an overview on current treatment development avenues.

Chapter 3 –Research paper 2 (Published)

De novo assembly of the *Carcinus maenas* transcriptome and characterization of innate immune system pathways

Bas Verbruggen, Lisa K. Bickley, Eduarda M. Santos, Charles R. Tyler, Grant D. Stentiford, Kelly S. Bateman and Ronny van Aerle. BMC Genomics (2015) 16:458

This chapter describes sequencing of RNA from several *C. maenas* tissues and creating a transcriptome based on *de novo* assembly of the data. Transcripts are annotated based on sequence similarities. Functional annotation and taxonomic overviews based on significant similarities are provided. In relation to WSSV infection it was deemed important to have an initial overview of the immune system of *C. maenas* and therefore pathways that are likely relevant in relation to viral infection are investigated in detail.

Chapter 4

In this chapter a draft genome is assembled for *C. maenas.* Within the genome the location of genes and their introns/exons can be identified and in combination with the transcriptome the draft genome provides an excellent basis for studying WSSV in *C. maenas*. It also provides a significant resource for other scientists working with this species. Within the genome there are elements other than host genes that may be discovered that can have an impact on WSSV infection. For example fossilized remnants the WSSV genome could convey resistance. A genome sequence is furthermore vital for being able to identify miRNA precursors. The latter are an important class of regulatory molecules in viral infection.

Chapter 5

As for *C. maenas*, the genomic resources for *H. gammarus* are sparse. In this chapter RNA from specific tissues was isolated and sequenced for this species. The transcriptome was assembled and annotated in similar fashion as that for *C. maenas*. Differences in the virus relevant infection pathways were identified between *H. gammarus* and *C maenas*. In addition sequences of known WSSV receptors were compared across aquatic crustaceans.

Chapter 6

In this chapter the response of *C. maenas* to WSSV exposure was investigated. Individuals were injected with virus innoculum or saline solution. Over a series of timepoints, ranging from 6 hours to 28 days post injections, gill from the individuals were sampled. Isolated gill RNA was subjected to both RNA and miRNA sequencing. This sequencing data enabled identification of replicating WSSV in the samples for monitoring progression of WSSV infection. Through differential expression analysis of host transcripts and miRNAs, important elements in the infection process were identified. Pathway analysis provided additional biological context.

Chapter 7 - Discussion

The final discussion provides a critical appraisal on the key findings in this thesis and identifies both strengths and weakness in the work conducted. The findings are put into context with contemporary research into WSSV and it also provides some final conclusions on the work and suggestions for future research directions.

1.5 References

1. FAO: **Fishery Statistical Collections: Global Aquaculture Production** In.; 2015.
2. Anderson JL: **Shrimp production review**. *Conference GOAL 2013, Paris* 2013.
3. De Schryver P, Defoirdt T, Sorgeloos P: **Early Mortality Syndrome Outbreaks: A Microbial Management Issue in Shrimp Farming?** *PLoS pathogens* 2014, **10**(4):e1003919.
4. Flegel TW, Lightner DV, Owens L: **Shrimp disease control: past, present and future**. *Diseases in Asian Aquaculture* 2008, **VI**:355-378.
5. Lightner DV: **The Penaeid Shrimp Viruses TSV, IHHNV, WSSV, and YHV**. *Journal of Applied Aquaculture* 1999, **9**(2):27-52.
6. Lightner DV: **Global transboundry disease politics: the OIE perspective.** *Journal of Invertebrate Pathology* 2012, **110**:184-187.
7. Stentiford GD, Neil DM, Peeler EJ, Shields JD, Small HJ, Flegel TW, Vlak JM, Jones B, Morado F, Moss S *et al*: **Disease will limit future food supply from the global crustacean fishery and aquaculture sectors**. *Journal of Invertebrate Pathology* 2012, **110**(2):141-157.
8. Rosenberry B: **World shrimp farming 200.** *Shrimp News International* 2001, **13**:324.
9. Briggs M: **Cultured Aquatic Species Information Programme. *Penaeus vannamei*. Cultured Aquatic Species Information Programme.**© *FAO Fisheries and Aquaculture Department [online]* 2006.
10. Stentiford GD, Bonami JR, Alday-Sanz V: **A critical review of susceptibility of crustaceans to Taura syndrome, Yellowhead disease and White Spot Disease and implications of inclusion of these diseases in European legislation**. *Aquaculture* 2009, **291**(1-2):1-17.
11. Walker PJ, Mohan CV: **Viral disease emergence in shrimp aquaculture: origins, impact and the effectiveness of health management strategies**. *Reviews in Aquaculture* 2009, **1**(2):125-154.
12. Bateman KS, Tew I, French C, Hicks RJ, Martin P, Munro J, Stentiford GD: **Susceptibility to infection and pathogenicity of White Spot Disease (WSD) in non-model crustacean host taxa from temperate regions**. *Journal of Invertebrate Pathology* 2012, **110**(3):340-351.
13. Shastry B: **SNPs: Impact on Gene Function and Phenotype**. In: *Single Nucleotide Polymorphisms.* Edited by Komar AA, vol. 578: Humana Press; 2009: 3-22.

14. Sanger F, Coulson AR: **A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase**. *Journal of Molecular Biology* 1975, **94**(3):441-448.

15. **DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)Accessed [28/09/2015].** [Available at: www.genome.gov/sequencingcosts.]

16. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z *et al*: **Genome Sequencing in Open Microfabricated High Density Picoliter Reactors**. *Nature* 2005, **437**(7057):376-380.

17. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR *et al*: **Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry**. *Nature* 2008, **456**(7218):53-59.

18. Pandey V, Nutter RC, E P: **Applied Biosystems SOLiD™ System: Ligation-Based Sequencing**. *Wiley, pages 29-41* 2008.

19. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M *et al*: **An integrated semiconductor device enabling non-optical genome sequencing**. *Nature* 2011, **475**(7356):348-352.

20. Illumina: **An Introduction to Next-Generation Sequencing Technology**. 2016.

21. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B *et al*: **Real-Time DNA Sequencing from Single Polymerase Molecules**. *Science* 2009, **323**(5910):133-138.

22. Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J *et al*: **A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis**. *Briefings in Bioinformatics* 2013, **14**(6):671-683.

23. Kawahara-Miki R, Wada K, Azuma N, Chiba S: **Expression Profiling without Genome Sequence Information in a Non-Model Species, Pandalid Shrimp (*Pandalus latirostris*), by Next-Generation Sequencing**. *PloS one* 2011, **6**(10):e26043.

24. Schliesky S, Gowik U, Weber APM, Bräutigam A: **RNA-Seq Assembly – Are We There Yet?** *Frontiers in Plant Science* 2012, **3**:220.

25. **UniVec Database** [www.ncbi.nlm.nih.gov/tools/vecscreen/univec]

26. **fastx-toolkit github** [https://github.com/agordon/fastx_toolkit]

27. Bolger AM, Lohse M, Usadel B: **Trimmomatic: A flexible trimmer for Illumina Sequence Data**. *Bioinformatics* 2014.

28. Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads.** *EMBnetjournal* 2011, **17**:10-12.

29. Ewing B, Hillier L, Wendl MC, Green P: **Base-Calling of Automated Sequencer Traces Using *Phred*. I. Accuracy Assessment**. *Genome research* 1998, **8**:175-185.

30. **Fastqc. a quality control tool for high throughput sequence data** [http://www.bioinformatics.babraham.ac.uk/projects/fastqc]

31. Verbruggen B, Bickley LK, Santos EM, Tyler CR, Stentiford GD, Bateman KS, van Aerle R*:* ***De novo*** **assembly of the** ***Carcinus maenas*** **transcriptome and characterization of innate immune system pathways**. *BMC genomics* 2015, **16**:458.

32. Zhao QY, Wang Y, Kong YM, Luo D, Li X, Hao P: **Optimizing *de novo* transcriptome assembly from short-read RNA-Seq data: a comparative study**. *BMC bioinformatics* 2011, **12 Suppl 14**:S2.

33. Kawahara Y, Oono Y, Kanamori H, Matsumoto T, Itoh T, Minami E: **Simultaneous RNA-Seq Analysis of a Mixed Transcriptome of Rice and Blast Fungus Interaction**. *PloS one* 2012, **7**(11):e49423.

34. Zerbino DR, Birney E: **Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs**. *Genome research* 2008, **18**(5):821-829.

35. Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, Stitt M, Usadel B: **RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics**. *Nucleic acids research* 2012, **40**(Web Server issue):W622-627.

36. Chou H-H, Holmes MH: **DNA sequence quality trimming and vector removal**. *Bioinformatics* 2001, **17**(12):1093-1104.

37. Goecks J, Nekrutenko A, Taylor J, Galaxy T: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences**. *Genome biology* 2010, **11**(8):R86.

38. Blankenberg D, Kuster GV, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J: **Galaxy: A Web-Based Genome Analysis Tool for Experimentalists**. In: *Current Protocols in Molecular Biology.* John Wiley & Sons, Inc.; 2001.

39. Doerks T, Copley RR, Schultz J, Ponting CP, Bork P: **Systematic identification of novel protein domain families associated with nuclear functions**. *Genome research* 2002, **12**(1):47-56.

40. Hernandez D, Francois P, Farinelli L, Osteras M, Schrenzel J: ***De novo*** **bacterial genome sequencing: millions of very short reads assembled on a desktop computer**. *Genome research* 2008, **18**(5):802-809.

41. Huang X, Madan A: **CAP3: A DNA Sequence Assembly Program**. *Genome research* 1999, **9**(9):868-877.

42. Wolf JBW, Bayer T, Haubold B, Schilhabel M, Rosenstiel P, Tautz D: **Nucleotide divergence vs. gene expression differentiation: comparative transcriptome sequencing in natural isolates from the carrion crow and its hybrid zone with the hooded crow**. *Molecular ecology* 2010, **19**:162-175.

43. Zeng V, Villanueva KE, Ewen-Campen BS, Alwes F, Browne WE, Extavour CG: ***De novo* assembly and characterization of a maternal and developmental transcriptome for the emerging model crustacean *Parhyale hawaiensis***. *BMC genomics* 2011, **12**:581-581.

44. Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, Colbourne JK, Willis BL, Matz MV: **Sequencing and *de novo* analysis of a coral larval transcriptome using 454 GSFlx**. *BMC genomics* 2009, **10**:219-219.

45. Clark MS, Thorne MAS, Toullec J-Y, Meng Y, Guan LL, Peck LS, Moore S: **Antarctic Krill 454 Pyrosequencing Reveals Chaperone and Stress Transcriptome**. *PloS one* 2011, **6**(1):e15919.

46. Colgan TJ, Carolan JC, Bridgett SJ, Sumner S, Blaxter ML, Brown MJF: **Polyphenism in social insects: insights from a transcriptome-wide analysis of gene expression in the life stages of the key pollinator, *Bombus terrestris***. *BMC genomics* 2011, **12**:623-623.

47. Tao X, Gu Y-H, Wang H-Y, Zheng W, Li X, Zhao C-W, Zhang Y-Z: **Digital Gene Expression Analysis Based on Integrated *De Novo* Transcriptome Assembly of Sweet Potato [*Ipomoea batatas* (L.) Lam.]**. *PloS one* 2012, **7**(4):e36234.

48. Schulz MH, Zerbino DR, Vingron M, Birney E: **Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels**. *Bioinformatics* 2012, **28**(8):1086-1092.

49. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol İ: **ABySS: A parallel assembler for short read sequence data**. *Genome research* 2009, **19**(6):1117-1123.

50. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ *et al*: ***De novo* assembly and analysis of RNA-seq data**. *Nature Methods* 2010, **7**(11):909-912.

51. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y *et al*: **SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler**. *GigaScience* 2012, **1**:18-18.

52. Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S *et al*: **SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads**. *Bioinformatics* 2014, **30**(12):1660-1666.

53. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al*: **Full-length transcriptome assembly from RNA-Seq data without a reference genome**. *Nature biotechnology* 2011, **29**(7):644-652.

54. Ren X, Liu T, Dong J, Sun L, Yang J, Zhu Y, Jin Q: **Evaluating de Bruijn Graph Assemblers on 454 Transcriptomic Data**. *PloS one* 2012, **7**(12):e51188.

55. Coppe A, Bortoluzzi S, Murari G, Marino IAM, Zane L, Papetti C: **Sequencing and Characterization of Striped Venus Transcriptome Expand Resources for Clam Fishery Genetics**. *PloS one* 2012, **7**(9):e44185.

56. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WEG, Wetter T, Suhai S: **Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs**. *Genome research* 2004, **14**(6):1147-1159.

57. Luo J, Wang J, Li W, Zhang Z, Wu FX, Li M, Pan Y: **EPGA2: memory-efficient *de novo* assembler**. *Bioinformatics* 2015, **31**(24):3988-3990.

58. Hunt M, Gall A, Ong SH, Brener J, Ferns B, Goulder P, Nastouli E, Keane JA, Kellam P, Otto TD: **IVA: accurate *de novo* assembly of RNA virus genomes**. *Bioinformatics* 2015, **31**(14):2374-2376.

59. Chang Z, Li G, Liu J, Zhang Y, Ashby C, Liu D, Cramer CL, Huang X: **Bridger: a new framework for *de novo* transcriptome assembly using RNA-seq data**. *Genome biology* 2015, **16**(1):30.

60. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nature methods* 2012, **9**(4):357-359.

61. Li H, Durbin R: **Fast and accurate short read alignment with Burrows–Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.

62. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq**. *Bioinformatics* 2009, **25**(9):1105-1111.

63. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms**. *Nature biotechnology* 2010, **28**(5):511-515.

64. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks**. *Nature Protocols* 2012, **7**(3):562-578.

65. Ghosh S, Chan C-K: **Analysis of RNA-Seq Data Using TopHat and Cufflinks**. In: *Plant Bioinformatics.* Edited by Edwards D, vol. 1374: Springer New York; 2016: 339-361.

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

67. Zhang Y, Jiang R, Wu H, Liu P, Xie J, He Y, Pang H: **Next-generation sequencing-based transcriptome analysis of *Cryptolaemus montrouzieri* under insecticide stress reveals resistance-relevant genes in ladybirds**. *Genomics* 2012, **100**(1):35-41.

68. Huson DH, Mitra S, Ruscheweyh HJ, Weber N, Schuster SC: **Integrative analysis of environmental sequences using MEGAN4**. *Genome research* 2011, **21**(9):1552-1560.

69. Ning Z, Cox AJ, Mullikin JC: **SSAHA: A Fast Search Method for Large DNA Databases**. *Genome research* 2001, **11**(10):1725-1729.

70. Anders S, Huber W: **Differential expression analysis for sequence count data**. *Genome biology* 2010, **11**(10):R106-R106.

71. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics* 2010, **26**(1):139-140.

72. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data**. *Genome biology* 2010, **11**(3):R25-R25.

73. Lee S, Seo CH, Lim B, Yang JO, Oh J, Kim M, Lee S, Lee B, Kang C, Lee S: **Accurate quantification of transcriptome from RNA-Seq data by effective length normalization**. *Nucleic acids research* 2011, **39**(2):e9-e9.

74. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome**. *BMC bioinformatics* 2011, **12**:323.

75. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN: **RNA-Seq gene expression estimation with read mapping uncertainty**. *Bioinformatics* 2010, **26**(4):493-500.

76. Bolstad BM, Irizarry RA, Åstrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185-193.

77. Oshlack A, Wakefield MJ: **Transcript length bias in RNA-seq data confounds systems biology**. *Biology Direct* 2009, **4**:14-14.

78. Fang Z, Martin J, Wang Z: **Statistical methods for identifying differentially expressed genes in RNA-Seq experiments**. *Cell & Bioscience* 2012, **2**:26-26.

79. Bloom JS, Khan Z, Kruglyak L, Singh M, Caudy AA: **Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays**. *BMC genomics* 2009, **10**:221-221.

80. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays**. *Genome research* 2008, **18**(9):1509-1517.

81. Auer PL, Doerge RW: **A Two-Stage Poisson Model for Testing RNA-Seq Data**. *Statistical Applications in Genetics and Molecular Biology* 2011, **10**(1):26.

82. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: **Differential analysis of gene regulation at transcript resolution with RNA-seq**. *Nature Biotechnology* 2013, **31**(1):46-53.

83. Hardcastle TJ, Kelly KA: **baySeq: empirical Bayesian methods for identifying differential expression in sequence count data**. *BMC bioinformatics* 2010, **11**:422.

84. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, Haag JD, Gould MN, Stewart RM, Kendziorski C: **EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments**. *Bioinformatics* 2013, **29**(8):1035-1043.

85. Zhou Y-H, Xia K, Wright FA: **A powerful and flexible approach to the analysis of RNA sequence count data**. *Bioinformatics* 2011, **27**(19):2672-2678.

86. Parra G, Bradnam K, Korf I: **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes**. *Bioinformatics* 2007, **23**(9):1061-1067.

87. Ashburner ea: **Gene ontology: tool for the unification of biology**. *Nature Genetics* 2000, **25**(1):25-29.

88. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic acids research* 2000, **28**(1):27-30.

89. Young MD, Wakefield MJ, Smyth GK, Oshlack A: **Gene ontology analysis for RNA-seq: accounting for selection bias**. *Genome biology* 2010, **11**(2):R14.

90. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research**. *Bioinformatics* 2005, **21**(18):3674-3676.

91. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(43):15545-15550.

92. **Ingenuity Pathway Analysis** [www.qiagen.com/ingenuity]

93. Nikitin A, Egorov S, Daraselia N, Mazo I: **Pathway studio--the analysis and navigation of molecular networks**. *Bioinformatics* 2003, **19**(16):2155-2157.

94. DeGraaf JD, Tyrrell MC: **Comparison of the Feeding Rates of Two Introduced Crab Species, *Carcinus maenas* and *Hemigrapsus sanguineus*, on the BLue Mussel, *Mytilus edulis***. *Northeastern Naturalist* 2004, **11**(2):163-167.

95. Perry H: ***Carcinus maenas***. *USGS Nonindigenous Aquatic Species Database* 2014.

96. Hänfling B, Edwards F, Gherardi F: **Invasive alien Crustacea: dispersal, establishment, impact and control**. *BioControl* 2011, **56**(4):573-595.

97. Jebali J, Chicano-Galvez E, Fernandez-Cisnal R, Banni M, Chouba L, Boussetta H, Lopez-Barea J, Alhama J: **Proteomic analysis in caged Mediterranean crab (*Carcinus maenas*) and chemical contaminant exposure in Teboulba Harbour, Tunisia**. *Ecotoxicology and environmental safety* 2014, **100**:15-26.

98. Klassen L: **A biological synopsis of the European green crab, *Carcinus maenas***. *Canadian Manuscript Report of Fisheries and Aquatic Sciences* 2007, **2818**(vii+75pp.).

99. Ben-Khedher S, Jebali J, Houas Z, Naweli H, Jrad A, Banni M, Boussetta H: **Metals bioaccumulation and histopathological biomarkers in *Carcinus maenas* crab from Bizerta lagoon, Tunisia**. *Environmental science and pollution research international* 2013.

100. Windeatt KM, Handy RD: **Effect of nanomaterials on the compound action potential of the shore crab, *Carcinus maenas***. *Nanotoxicology* 2013, **7**(4):378-388.

101. Watts AJ, Lewis C, Goodhead RM, Beckett SJ, Moger J, Tyler CR, Galloway TS: **Uptake and retention of microplastics by the shore crab *Carcinus maenas***. *Environmental Science & Technology* 2014, **48**(15):8823-8830.

102. FAO: **Species Fact Sheets: *Homarus gammarus***. 2015.

103. FAO: **Fishery Statistical Collections: Global Capture Production**. 2015.

104. Drengstig A, Bergheim A: **Commercial land-based farming of European lobster (*Homarus gammarus* L.) in recirculating aquaculture system (RAS) using a single cage approach**. *Aquacultural Engineering* 2013, **53**:14-18.

Chapter 2

Molecular Mechanisms of White Spot Syndrome Virus Infection and Perspectives on Treatments.

Viruses (2016), *8*(1), 23.

Pages: 55 - 82

viruses

MDPI

Review

Molecular Mechanisms of White Spot Syndrome Virus Infection and Perspectives on Treatments

Bas Verbruggen [1,*], Lisa K. Bickley [1], Ronny van Aerle [2], Kelly S. Bateman [2], Grant D. Stentiford [2], Eduarda M. Santos [1,*,†] and Charles R. Tyler [1,*,†]

Received: 7 November 2015; Accepted: 6 January 2016; Published: 18 January 2016
Academic Editor: Alexander Ploss

[1] Biosciences, College of Life & Environmental Sciences, Geoffrey Pope Building, University of Exeter, Exeter, Devon EX4, UK; L.K.Bickley@exeter.ac.uk

[2] European Union Reference Laboratory for Crustacean Diseases, Centre for Environment, Fisheries and Aquaculture Science (Cefas), Weymouth Laboratory, Weymouth, Dorset DT4 8UB, UK; ronny.vanaerle@cefas.co.uk (R.V.A.); kelly.bateman@cefas.co.uk (K.S.B.); grant.stentiford@cefas.co.uk (G.D.S.)

* Correspondence: bv213@exeter.ac.uk (B.V.); e.santos@exeter.ac.uk (E.M.S.); c.r.tyler@exeter.ac.uk (C.R.T.); Tel.: +44-(0)-1392-724607 (B.V.); +44-(0)-1392-724607 (E.M.S.); +44-(0)-1392-724450 (C.R.T.)

† These authors contributed equally to this work.

Abstract: Since its emergence in the 1990s, White Spot Disease (WSD) has had major economic and societal impact in the crustacean aquaculture sector. Over the years shrimp farming alone has experienced billion dollar losses through WSD. The disease is caused by the White Spot Syndrome Virus (WSSV), a large dsDNA virus and the only member of the Nimaviridae family. Susceptibility to WSSV in a wide range of crustacean hosts makes it a major risk factor in the translocation of live animals and in commodity products. Currently there are no effective treatments for this disease. Understanding the molecular basis of disease processes has contributed significantly to the treatment of many human and animal pathogens, and with a similar aim considerable efforts have been directed towards understanding host–pathogen molecular interactions for WSD. Work on the molecular mechanisms of pathogenesis in aquatic crustaceans has been restricted by a lack of sequenced and annotated genomes for host species. Nevertheless, some of the key host–pathogen interactions have been established: between viral envelope proteins and host cell receptors at initiation of infection, involvement of various immune system pathways in response to WSSV, and the roles of various host and virus miRNAs in mitigation or progression of disease. Despite these advances, many fundamental knowledge gaps remain; for example, the roles of the majority of WSSV proteins are still unknown. In this review we assess current knowledge of how WSSV infects and replicates in its host, and critique strategies for WSD treatment.

Keywords: White Spot Syndrome Virus; host–pathogen interactions; viral infection pathway; endocytosis; stress responses; apoptosis; treatments; miRNA

1. Introduction

Since its emergence in the early 1990s, White Spot Disease (WSD) has become the greatest threat to global crustacean aquaculture industries [1]. The first case of WSD was reported in China in 1991 [2] and this was followed by spread to other major aquaculture regions of the world including East and Southeast Asia, the Americas, India, the Middle East [2–7], and even Europe [8]. The total economic damage caused by WSD to the shrimp aquaculture industry has been estimated at \$8–\$15 billion since its emergence [9], increasing by \$1 billion annually [10,11]. Annual losses have traditionally equated to approximately one tenth of global shrimp production [11]. A wide range of other crustacean hosts are susceptible to WSSV infection and disease. Natural populations of these hosts can also act as reservoirs for this pathogen [11,12]. Given the scale of WSD in captive crustaceans, it is not surprising that considerable scientific effort has been directed towards establishing the underlying mechanisms of the disease and identifying potential treatments for disease prevention or alleviation. However, despite a growing knowledge base and research interest, to date cost-effective vaccinations and/or treatments remain elusive.

WSD is caused by White spot syndrome virus (WSSV) [13], a double-stranded DNA virus and the only member of the genus *Whispovirus* and family *Nimaviridae* [14,15]. WSSV was originally classified as member of the *Baculoviridae* family, but it was later reclassified and named White Spot Syndrome Virus 1 by the International Committee on Taxonomy of Viruses ICTV [14–17]. The *Whispovirus* is a newly recognized family and its membership is likely to increase in the future as new taxa are discovered [17]. Vlak *et al.* [17] tentatively list B virus, B2 virus, τ (tau) virus, and Baculo-A and Baculo-B viruses as putative members of the *Whispovirus* genus and *Nimaviridae* family.

Animals suffering from WSD display various clinical signs including lethargy, reduced food consumption, reduced preening activities, a loosening of the cuticle, and a discoloration of the hepatopancreas [18,19]. White calcified spots appearing on the exoskeleton are diagnostic of WSD in some [19] but not all host species (e.g., the Indian prawn (*Penaeus indicus*) [20]). In farmed shrimp mortality is rapid and 3–10 days after infection cumulative mortality is generally between 90% and 100% [6,21].

WSSV infection occurs in all tissues of mesodermal and ectodermal origin (e.g., gills, lymphoid organ, cuticular epithelium, sub-cuticular connective tissues). Infected nuclei become hypertrophied with marginalized chromatin, and contain inclusion bodies that stain intensely eosinophillic in early-stage infection and basophilic in more advanced infection (Figure 1) [22]. Morphologically, WSSV virions are large and rod-shaped with dimensions in the range of 80–120 × 250–380 nm [13,23,24]. Structurally, they have a nucleocapsid surrounded by a trilaminar envelope [23] with a tail-like appendage, the function of which is unknown [13,23].

Economic drivers have focused research on WSD in farmed shrimp; however, the host range of WSSV includes many other decapod and non-decapod species including crabs, lobsters, prawns, crayfishes, and copepods [18,25]. To date 98 potential host species for WSD have been identified [2]. It is also the case that many organisms living in a WSSV-infected pond can function as a vector for WSSV [2]. Although WSD causes high levels of mortality in all cultured shrimps, it is not necessarily fatal to other hosts [26]. Variation in disease susceptibility occurs across the Crustacea and this is of particular interest as it may provide insights into disease resistance [3,27–36]. The shore crab (*Carcinus maenas*), for instance, while confirmed as susceptible to infection, appears to be especially recalcitrant to development of disease, showing little pathology and low mortality rates [36].

WSSV can be transmitted through the consumption of infected tissue, by cannibalism/predation, and via exposure to water containing WSSV virions [37,38]. Transmission through ingestion of contaminated prey tends to be the most effective path for infection of these two exposure routes [18,39]. A comparison between Asian tiger shrimp (*Penaeus monodon*) and Whiteleg shrimp (*Litopenaeus vannamei*) has shown that overall transmission rates are similar in these species but the relative contributions of direct and indirect transmission rates differ [40]. Transmission through physical contact with an infected animal is sufficient to initiate and progress WSD epidemics in ponds [40]. Vertical transmission (parents to progeny) of WSSV has been suggested and WSSV particles have been identified in oocytes. However, WSSV particles appear to be absent in mature eggs, suggesting that oocytes that contain the virus do not develop to mature eggs [18]. Vertical transmission might therefore occur through ingestion of viral particles that have been shed by adults during spawning into larval stages [18,37,41]. Virus pathogenicity can also be affected as it passes between species. As an example, WSSV, when passed through *Macrobrachium rosenbergii*, had a reduced pathogenicity (reduced mortality rates) in *P. monodon* when compared to non-transmitted virus. How this change in pathogenicity is mediated has not been established but was shown to be accompanied by variations in tandem repeat regions in the WSSV genome [42]. When considering transmission of WSSV between different geographical locations there is good evidence that this is facilitated by the transport of live and frozen uncooked shrimp [43,44] and the import of brood stock [11].

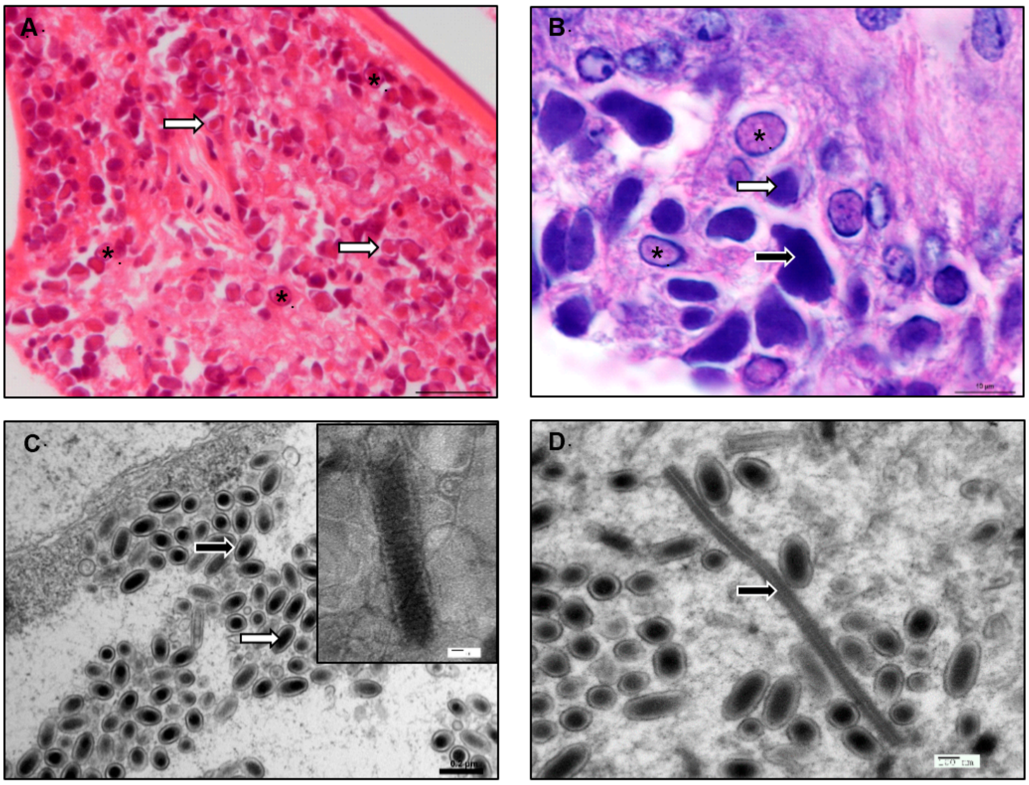


Figure 1. White Spot Syndrome Virus (WSSV) infection of *Litopenaeus vannamei*. The infection progresses through different stages that can be seen in the nucleus via histology. (**A**) Early-stage infected cells display enlarged nuclei with marginalized chromatin and a homogenous eosinophilic central region. These then develop an intranuclear eosinophilic Cowdry A-type inclusion (*); this can be surrounded by a clear halo beneath the nuclear membrane (**white arrow**). Scale bar = 25 µm; (**B**) The eosinophilic inclusion usually expands to fill the nucleus (*). This inclusion becomes basophilic in staining and denser in color as the infection progresses (**white arrow**). Nuclei then disintegrate so that the content fuses with the cytoplasm (**black arrow**). Scale bar = 10 µm. H & E stain; (**C**) WSSV virions appear ovoid in shape and contain an electron-dense nucleocapsid (**white arrow**) within a trilaminar envelope (**black arrow**). Scale bar = 0.2 µm. Inset. Negatively stained WSSV nucleocapsid, showing the presence of cross-hatched or striated material that is structured as a series of stacked rings of subunits and is a key diagnostic feature of WSSV. Scale bar = 20 nm; (**D**) Presumptive nucleocapsid material within the nucleus prior to envelopment. This material is cross-hatched or striated in appearance and linear prior to its incorporation in the formation of mature WSSV particles. This linear nucleocapsid material is observed sporadically in the manufacture of the WSSV particles. Scale bar = 100 nm. Transmission electron microscopy images.

Prevention or treatment strategies for WSD disease could be advanced through an understanding of how this virus infects organisms and/or how relatively resistant animals process WSSV during the infection process. This requires understanding of the (molecular) interactions between WSSV and its potential hosts. In the infection process, WSSV invades host cells and initiates replication of its components. This is followed by assembly and release of new virions, resulting in host cell death and disease. To prevent disease, hosts must recognize the invading pathogen and elicit appropriate defense strategies or create a cellular environment that is not appropriate for production of new virions. A number of review articles have been published detailing the interactions between viruses and the host innate immune system (Li *et al.* [45], Shekhar and Ponniah [46], Sanchez-Paz [47], and Sritunyalucksana *et al.* [48]) but the interactions between WSSV and the host intracellular environment have received less attention. This is fundamental for advancing our understanding of the WSD infection process and exploring potential opportunities for disease treatment and prevention. In this review, we analyze the current knowledge on the WSSV genome, with a focus on the molecular mechanisms that enable WSSV to interact with host machinery and maintain a cellular environment favorable for the production of new virions. We then investigate the current treatment options that have been explored and consider possible future directions for advancing disease treatment and mitigation.

2. The WSSV Genome and miRNAS

2.1. WSSV Genome

WSSV contains a circular dsDNA genome of approximately 300 kb in size. Genome sequences for four WSSV isolates are available (a Chinese isolate (WSSV-CN; GenBank Accession AF332093) [49], an isolate from Thailand (WSSV-TH; GenBank Accession AF369029) [16], a Taiwanese isolate (WSSV-TW; GenBank Accession AF440570), and a Korean isolate (WSSV-KR; GenBank Accession JX515788) [50]). They differ in size, indicating some degree of genomic instability: (293 kb (Thailand), 296 kb (Korea), 305 kb (China), and 307 kb (Taiwan)). Overall, the sequence identity between isolates ranges between 97% and 99%, and the GC content in all isolates is 41% [50]. The WSSV genome contains nine homologous regions (*hr 1–9*) consisting of several repeats of 250–300 bp fragments, encompassing direct repeats, atypical inverted repeat sequences, and imperfect palindromic sequences [16]. Such *hr* regions in the genomes of baculoviruses have been hypothesized to play a part in DNA replication [51].

Several important genetic variations have been identified between the genomes of the Thai, Taiwanese, and Chinese isolates [51]. The largest involves a deletion of approximately 13 kb (WSSV-TH) and 1 kb (WSSV-CN), in the same genomic region, relative to WSSV-TW [51]. A region of approximately 750 bp shows variation between WSSV-TH and WSSV-TW, but the sequence of this region has no homology to sequences in publically available databases. Another variation includes a 1337 bp insert in WSSV-TH that shows 100% homology with a known transposable element [51]. Variable number tandem repeats (VNTR) are present amongst four *hrs*: *hr1*, *hr8*, *hr3* and *hr9*. VNTRs occur in open reading frames (ORF), e.g., *ORF75*, *ORF94* and *ORF125* in WSSV-TH, and have been suggested for use in genotyping WSSV [51,52]. Sequence variations have been shown to be stable over at least six passages through three different penaeid shrimp species [52]. Smaller variations like single nucleotide polymorphisms and short insertions/deletions are found throughout the WSSV genome [47]. The effects of genetic variation on WSSV virulence and fitness have been investigated in a number of studies, which have demonstrated higher virulence and competitive fitness in isolates with smaller genomes [53].

Putative ORFs in the WSSV genome have been determined for every isolate. These ORFs are present on both strands (~54% forward and ~46% reverse [16,50]), and the total number of estimated ORFs in the different isolates are 532 for WSSV-TW, 515 for WSSV-KR, 531 for WSSV-CN, and 684 for WSSV-TH. The nomenclature of ORFs differs in GenBank; for reference the conventions are as follows: WSSV-TW = WSSVxxx, WSSV-KR = wssv_xxxxx, WSSV-CN = wsvxxx and WSSV-TH = ORFxxx. The number of expressed ORFs is difficult to determine since many do not show homology with known protein sequences. Estimates are based on the potential to code for proteins of at least 50–60 amino acids. Functional predictions based on sequence similarity and protein motifs indicate that around 180 ORFs are likely to be expressed [16,49]. Microarrays constructed based on 184 putative ORFs of WSSV-TH confirmed expression of 79% of these ORFs in the gills of *P. monodon* infected with WSSV [54]. An analysis of the codons of ORFs (>100 codons) in WSSV-TW, WSSV-CN, and WSSV-TH indicated that codon usage bias and base composition are determined by compositional limitations and mutational pressure [55]. Interesting features of WSSV ORFs include one exceptionally long ORF—the longest among all viruses, specifically wssv_03600—which is 18,221 bp in the Korean isolate.

Its protein product, VP664, is a major structural protein that forms the stacked rings of the WSSV nucleocapsid [16,50,56–58]. Furthermore, only around 30% of the ORFs in the WSSV genome contain a polyadenylation signal. ORFs without polyadenylation signals are usually part of a cluster of ORFs with small intergenic regions and identical transcriptional orientation that can produce polycistronic mRNAs (e.g., the *vp60b/wssv478/wssv479/vp28* cluster) [59–62]. Translation from these polycistronic mRNAs is likely to be facilitated by internal ribosome entry sites (IRES) [62,63]. Several IRES have been identified thus far including in the 5′ UTR of *wssv480* (*vp28*) [64], in the coding regions of *wssv396* (vp31) and *wssv395* (*vp39b*) [63], and in the 5′ UTR of *icp35* [62]. Translation initiation through IRES enables viral protein production even under unfavorable conditions such as during host response to viral infection, therefore making the virus more robust to host interference [63,65].

2.2. miRNAs, WSSV Infection, and Pathogenesis

MiRNAs (small non-coding RNA molecules) have been documented to be widely expressed in both animal and plant species, and have key regulatory roles in a number of cellular pathways, including early development, cell differentiation and proliferation, apoptosis, signal transduction, and immunity [66].

It is now well established that miRNAs are involved in many host–pathogen interactions, as reviewed in Asgari 2011 and Skalsky *et al.* 2010 [67,68]. The primary function of miRNAs involves regulation of gene expression at the post-transcriptional level. Typically, miRNAs bind to complementary sequences (either partial or complete) in the mRNA of target genes, regulating gene expression by repressing translation or directing sequence-specific degradation of the mRNA. However, evidence has also emerged that interaction of miRNAs with target genes may also function to induce gene expression [69–71]. MiRNAs encoded in a virus genome may function to regulate either viral or host genes to manipulate immune responses and cellular functions for the benefit of the virus. Additionally, viruses may regulate and use host miRNAs to facilitate their own replication or to regulate the virus life cycle [72,73]. Alternatively, host miRNAs can act to limit viral replication or alter cellular processes to the disadvantage of the virus [68,74].

In WSSV-challenged shrimp (*M. japonicus*) 63 host miRNAs have been identified through small RNA sequencing, of which 48 could be mapped to other known arthropod miRNAs in the miRBase database [75]. Thirty-one of the miRNAs show differential expression in response to WSSV infection. Using target gene prediction algorithms, many of the miRNAs were predicted to target genes involved in host immunity, including the small GTPase-mediated signaling transduction pathway, autophagy, phagocytosis, apoptosis, the Toll-like receptor signal pathway, antimicrobial humoral response, endocytosis, RNAi, and regulation of the innate immune response. Huang and Zhang [76] showed a more direct interaction between host miRNA and WSSV: a miRNA found to be upregulated in WSSV-infected *M. japonicus*, miR-7, was predicted to target the 3′-untranslated region of *wsv477*. In insect High Five cells, the expression of enhanced green fluorescent protein (EGFP) was dramatically reduced when coupled to the *wsv477* 3′UTR compared to controls [76]. *Wsv477* is an early gene that is involved in viral gene replication. Its inhibition would thus have negative effects on WSSV replication. Indeed, injection of miR-7 was shown to reduce WSSV copies 1000-fold compared with WSSV only at 72 and 96 h post-infection [76].

The specific viral miRNAs encoded by WSSV have been investigated in *M. japonicus* [77,78]. In the first study on these miRNAs in *M. japonicus*, WSSV was shown to have the capacity to encode 40 distinct miRNAs, which is a miRNA density 360 times greater than in humans [77]. The authors of that work suggested that this high miRNA content may contribute to the ability of viruses to respond rapidly to selective pressures placed on them from the host organism. In this study miRNAs were first predicted through bioinformatic analyses and subsequently their presence was confirmed through expression (microarray) studies and Northern blots. Subsequent work applying small RNA sequencing identified additional WSSV miRNAs [78], bringing the total number of WSSV miRNAs collectively to 89 [78]. It was observed that the majority of the miRNAs were expressed during the early stages of infection and that host genes like *Drosha*, *Dicer1*, and *Ago1* are necessary for successful miRNA biogenesis [77,78]. Interestingly, several miRNAs showed differential expression across tissue types, indicating that viral regulatory strategies could be regulated to fit the infected tissue type [78]. One example of this regulatory effect was the potential of WSSV-miR-N24 to inhibit apoptosis through downregulation of caspase 8 expression [78]. WSSV miRNAs can also regulate the balance between promotion of viral infection and latency. He *et al.* [79] identified two WSSV miRNAs, WSSV-miR-66 and WSSV-miR-68, that promote WSSV infection through regulating expression of WSSV genes. It is hypothesized that the targets of these miRNAs, *wsv094* and *wsv177* (WSSV-miR-66), and *wsv248* and *wsv309* (WSSV-miR-68), are related to latency. These studies provide key evidence that WSSV miRNAs function as regulatory factors involved in the virus life cycle as well as the host immune response.

3. WSSV Infection

The life cycle of viruses is well studied, and can be broadly divided into three phases: entry into the host cell (either directly or through host mechanisms such as endocytosis), uncoating of the genome followed by replication, and, finally, particle assembly and release. A model for the WSSV life cycle and morphogenesis was suggested by Escobedo-Bonilla and colleagues [80]. Throughout these phases a wide range of molecular interactions occur between the WSSV and its host (see Figure 2). These molecular interactions can be key factors in determining host susceptibility and pathogenicity, and also provide opportunities for treatment interventions.

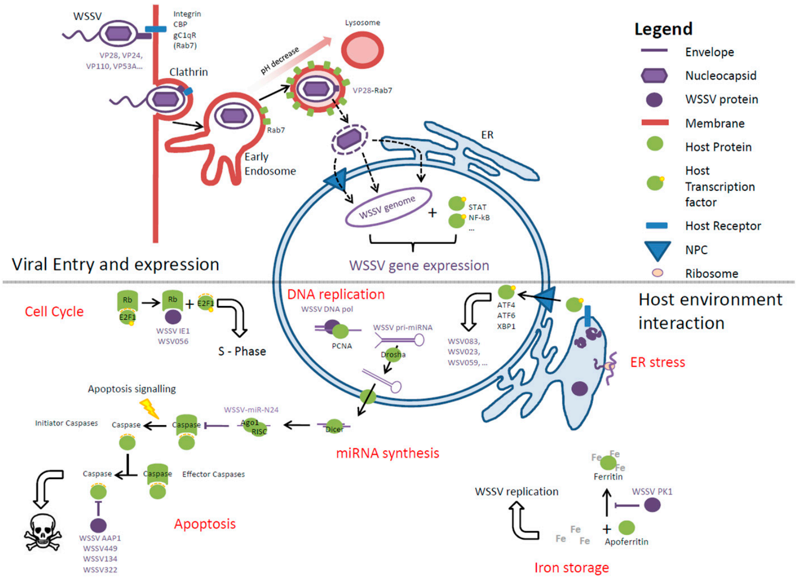


Figure 2. Overview of WSSV entry and environment interactions. Top panel: Viral entry into the host cell. WSSV proteins interact with host receptors, which leads to induction of Clathrin-mediated endocytosis. WSSV then travels through endosomes. During maturation the pH decreases, a cue for viruses to exit the endosomes. This stage probably involves an interaction between VP28 and Rab7. How WSSV passes through the nuclear envelope is unknown. Once in the nucleus, host transcription factors bind the WSSV genome (e.g., *ie1*) and initiate expression of viral genes. Bottom panel: Intracellular interactions between WSSV and the host cell. WSSV DNA replication requires host machinery (e.g., processivity factors) and to make these available WSSV can act to halt the cell cycle in the S-phase through E2F1. A high level of viral protein production can lead to ER stress, e.g., activation of unfolded protein response (UPR) pathways. Transcription factors of the UPR can activate expression of viral genes, which in turn may inhibit translation through eIF2. WSSV replication requires essential nutrients including iron. To prevent the host from withholding iron, WSSV can inhibit the binding of iron to Ferritin. WSSV can influence apoptosis signaling either through miRNA-mediated inhibition of initiator caspases or through viral proteins that inhibit effector caspase activity.

3.1. Viral Receptors, Interactions, and Entry

The primary obstacle for entry into host cells is the host's cell membrane. Viruses have evolved several methods to overcome this barrier including via lipid fusion and membrane perforation, but they enter the host cell predominantly by endocytosis [81]. In the latter case viruses bind with host cell surface proteins, carbohydrates, and lipids [81], which then trigger one of the various endocytic pathways including Clathrin-mediated endocytosis, caveolar endocytosis, and macropinocytosis [81–84]. WSSV enters the host cell through activating the endocytotic process with protein interactions. Known interacting elements for WSSV and its hosts are presented in Table 1. It can be seen that VP28, the major envelope protein, is a major player in host–virus protein interactions [85–88]. Some interactions are beneficial for the virus; others have negative impacts. There is also a series of interacting proteins for which the functions are not well established.

Table 1. Established WSSV–host protein interactions.

Viral Protein	Host Protein	Species	Reference
VP24, VP32, VP39B, VP41A, VP51B, VP53A, VP53B, VP60A, VP110, VP124, VP337	Chitin-binding protein (PmCBP)	*Penaeus monodon*	[89,90]
VP53A	Glut1	*P. monodon*	[91]
VP15, VP26, VP28	gC1qR (PlgC1qR)	*Pacifastacus leniusculus*	[92]
VP95, VP28, VP26, VP24, VP19, VP14	C-type lectin (LvCTL1)	*Litopenaeus vannamei*	[93]
VP28	C-type lectin (FcLec3)	*Fenneropenaeus chinensis*	[94]
VP26, VP28	C-type lectins (MjLecA, MjLecB, MjLecC)	*Marsupenaeus japonicus*	[95]
VP28	C-type lectins (MjsvCL)	*M. japonicus*	[96]
VP28	C-type lectins (LdlrLec1, LdlrLec2)	*M. japonicus*	[97]
VP187	β-Integrin	*P. japonicus* *P. clarkii*	[98]
VP26, VP31, VP37, VP90, VP136	β-Integrin	*L. vannamei*	[99]
WSSV-CLP	α-integrin, β-integrin, Syndecan	*F. chinensis*	[100] (Bioinformatic prediction)
VP15, VP28	Calreticulin (PlCRT)	*P. leniusculus*	[101]
VP466	Rab (PjRab)	*P. japonicus*	[102]
VP28	Rab7 (PmRab7)	*P. monodon*	[103]
ORF514	PCNA (lvPCNA)	*L. vannamei*	[104] (Bioinformatic prediction)
WSSV PK1	Ferritin (lvFerritin)	*L. vannamei*	[105]
Wsv083	FAK (MjFAK)	*M. japonicus*	[106]
AAP1 (WSSV449)	Caspase (PmCaspase)	*P. monodon*	[107]
WSSV134, WSSV332	Caspase (PmCasp)	*P. monodon*	[108]
WSSV249	Ubc (PvUbc)	*L. vannamei*	[109]
ICP11	Histones	*P. monodon*	[110]

Table 1. *Cont.*

Viral Protein	Host Protein	Species	Reference
VP9	RACK1 (PmRACK1)	*P. monodon*	[111]
VP15	FKBP46 (PmFKBP46)	*P. monodon*	[112]
VP15	CRT (PlgCRT)	*P. leniusculus*	[101]
WSSV-miRNA	Dorsha, Dicer, Ago1	–	[77]
VP14	Arginine kinase (LvAK)	*L. vannamei*	[113]
VP26	Actin	*Procambarus clarkii*	[114]
ORF427	PPs	*L. vannamei*	[115]
WSSV IE1	TATA box-binding protein (PmTBP)	*P. monodon*	[116]
WSSV IE1, WSV056	Retinoblastoma protein (Lv-RBL)	*L. vannamei*	[117]

Integrin receptors on host cell surfaces have been shown to be important targets for WSSV. Principally involved in the binding of cells to the extracellular matrix (or cell–cell adhesion), integrins are heterodimeric surface receptors that recognize Arg-Gly-Asp (RGD) motifs in target proteins [118]. Several WSSV proteins can bind to α- or β-integrin homologues. Immunoprecipitation experiments have shown binding of β-integrin by VP187 (wsv209), a viral protein that contains an RGD motif [98]. This body of work also found that WSSV infection could be blocked by soluble integrin, integrin-specific antibody, an RGD-containing peptide, and silencing of β-integrin, indicating an important role for integrins as WSSV receptors [98]. The viral proteins VP26, VP31, VP37, VP90 and VP136 also interact with the integrin receptors of *L. vannamei*, mainly through interactions with RGD-, YGL-, and LDV peptide motifs present in these viral proteins [99]. VP26 is a tegument protein that associates with a protein complex involving VP24, VP28, VP38A, VP51A, and WSV010 in the viral envelope [57,119]. Since VP26 is not present on the cell exterior it is unlikely to play a role in binding host cells. Finally, bioinformatics analyses have predicted that WSSV collagen-like protein (WSSV-CLP) interacts with integrins on the basis of sequence similarity to known interacting protein pairs [100]. Together, these studies show that multiple proteins are involved in the recognition and binding of host integrins [99]. In *P. monodon*, the cell surface Chitin-binding protein (PmCBP) has been shown to bind to 11 WSSV proteins that likely form a complex on the surface of WSSV (VP24, VP110, VP53A, VP53B, VP337, VP32, VP124, VP41A, VP51B, VP60A, and VP39B) [90]. Facilitating viral entry through protein complex interactions also occurs for other enveloped DNA viruses (e.g., Herpesviridae [120]). Interactions of the viral protein complex with PmCBP and with VP53A appear to be facilitated by glucose transporter1 (Glut1) [91].

The innate immune system of the host employs groups of proteins that are able to recognize and bind pathogen-associated molecular patterns (PAMP) molecules. One family of proteins that recognizes non-self molecules are the lectins. Calcium-dependent lectins (C-type lectins) have been demonstrated to interact with WSSV proteins. Several interactions between host C-type lectins and WSSV proteins have been identified (see Table 1), including a C-type lectin in *L. vannamei* (LvCTL1) that can interact with VP95, VP28, VP26, VP24, VP19, and VP14 [93]. Treating WSSV with recombinant LvCTL1 prior to a shrimp WSSV exposure was shown to result in higher survival rates, indicating a protective effect [93]. Lectins from other shrimp species have shown interactions with VP26 and VP28. For example, in the Chinese white shrimp (*F. chinensis*) a lectin (FcLec3) has been identified that can interact with VP28 [94]. Using VP26, VP28, and VP281 to screen a phage display library of *M. japonicus*, three lectins (MjLecA, MjLecB, and MjLecC) were identified to interact with the viral proteins [95]. Of these, MjLecA and MjLecB were shown to reduce viral infection rate *in vitro* [95]. Interactions with other lectins found in *M. japonicus* indicate a contradictory relationship between WSSV and lectins.

For example, a C-type lectin isolated from the stomach of *M. japonicus* (MjsvCL) has been shown to interact with VP28 [96] but, in contrast to other lectins, MjsvCL appears to facilitate WSSV infection. MjsvCL expression is induced by viral infection, and inhibition of its expression by RNAi results in lower virus replication, whereas exogenous MjsvCL enhances replication [96]. MjsvCL contributes to viral entry by binding of VP28 and the calreticulin receptor on the cell surface. MjsvCL thus works as a bridge between the virus and the calreticulin receptor [96]. The dual roles of lectins are indicative of the arms race between the virus and the host immune system, as is also observed for various viral pathogens in humans [121]. Members of the complement system have also been shown to interact with WSSV proteins. In the signal crayfish *Pacifastacus leniusculus* the surface receptor for C1q (a component of the complement system) can interact with WSSV proteins VP15, VP26, and VP28 [92] and this is upregulated upon WSSV infection, producing a protective effect.

After binding to receptors on the cell surface, the enveloped virus can either penetrate the membrane directly or undergo uptake through endocytosis. Since there are several routes of endocytosis it is useful to identify which of these may be adopted by WSSV, if any. Several experiments have investigated WSSV endocytosis and the results depend, at least in part, on cell type. Huang *et al.* [101] showed that WSSV can enter hemocytes of *L. vannamei* in primary culture through caveolar endocytosis. The evidence for this included inhibition of WSSV uptake under methyl-β-cyclodextrin (MβCD, an inhibitor of caveolar endocytosis) treatment and a lack of effect of chlorpromazine (CPZ, an inhibitor of Clathrin-mediated endocytosis) [122]. There were similar findings in the crayfish, *Cherax quadricarinatus* [123]. However, in crayfish hematopoietic tissue (HPT) it appears to be the Clathrin-mediated endocytosis pathway that is responsible for viral uptake [124]. For this tissue, CPZ significantly inhibited WSSV internalization. Furthermore, WSSV particles co-localized with Clathrin and there was a dependence on membrane cholesterol and dynamin supported uptake, indicating Clathrin-mediated endocytosis. It is difficult to reconcile this apparent contradiction in uptake pathways; however, Huang *et al.* [124] point to data suggesting that MβCD also affects Clathrin-mediated endocytosis and further suggesting that WSSV could employ more than one endocytosis pathway for host entry.

3.2. *Escaping from Endosomes*

After penetrating the cellular membrane, WSSV particles become localized within early endosomes. Cellular cargo carried by endocytic vesicles can have a variety of different destinations in the cell depending on the sorting that occurs in the early endosome [125]. Some proteins or lipids are recycled to the plasma membranes, while others are degraded in lysosomes. Viruses need to avoid this fate. The sorting process occurs within minutes of vesicular entry into the cell. Due to their size, viruses are sorted by the host to the degradation pathway [126]. In completing their primary roles in sorting the endocytotic cargo, early endosomes "mature" in a process that includes several changes to the organelles including lumen acidification, their movement toward the perinuclear region, and changes in membrane lipid/protein composition. It is this maturation process that provides the cue for viruses to initiate their escape from the endosomes, which involves conformational changes of viral proteins in response to a lower pH. Escape from the endosomes can occur either through membrane fusion or lysis/leakage of the endosomal compartments [127]. After release from the endosomes the virus directs itself to the nucleus and penetrates the nuclear envelope. The pathways adopted to accomplish this by various virus families are reviewed by Kobiler *et al.* [128], the main difference being whether the nuclear pore complex (NPC) is involved or not.

An important class of regulatory proteins in endocytosis are the small Rab GTPases. Rab GTPases are key regulators of endosome maturation. Early endosomes are characterized by the presence of Rab5 on their membranes [129]. During the endosome maturation process Rab5 is exchanged for Rab7, the Rab GTPase associated with late endosomes, and this operates in a positive feedback loop mechanism [130]. Rab7 can subsequently associate with RAB effector proteins like RILP that bind Dynein, ensuring movement of the vesicle to the perinuclear area [129].

The Rab GTPases have been implicated in the WSSV infection process. In *P. monodon* Rab7 (PmRab7) has been shown to interact with WSSV protein VP28 [103]. Moreover it has been shown that in shrimp injected with WSSV and PmRab7, or PmRab7 antibody, there is a decreased rate in mortality to 15% and 5%, respectively, compared with a mortality rate of 95% in animals injected with WSSV alone [103]. In *L. vannamei* brood stock, dsRNA-mediated silencing of *lvRab7* has also been shown to reduce mortality rates, albeit a mild reduction [131]. Similarly, in *P. monodon* silencing of both viral ribonucleotide reductase small subunit (*rr2*) and host *Rab7* resulted in a 95% survival of infected animals compared with 100% mortality in animals treated with WSSV only [132,133]. It is unclear how the interaction between VP28 and Rab7 is mediated as the proteins are separated by the endosomal membrane. It is possible that this interaction occurs after WSSV has been released from the endosomes, but this can occur only if its envelope containing VP28 remains attached to the nucleocapsid. Alternatively, the Rab proteins could be present on the cell surface [103]. Applying subtractive hybridization, the amplification of only differentially expressed genes has shown that a Rab GTPase gene was upregulated in the hepatopancreas of WSSV-resistant *P. japonicus* [134]. This Rab GTPase gene showed homology with Rab6A, a protein involved in transport between the endosomes, Golgi, and endoplasmic reticulum [135]. It is possible that a higher expression of Rab6A leads to an increase in recycling of endosomes, removing the potential for WSSV to interact with Rab7, which in turn inhibits infection and thereby leads to resistance.

Overall, the mechanisms through which WSSV reaches the nucleus after its initial uptake by caveolar-mediated endocytosis remain poorly understood. Since the journey from the endosomes to nucleus is a limiting step in the infection process for many viruses, a greater knowledge of this pathway for WSSV is likely to help identify targets for drug development for disease prevention [128].

3.3. Viral Replication—The Molecular Processes

Once inside the host nucleus the virus has to express its own genes to allow for its own replication. Since WSSV does not carry its own transcriptional machinery, it relies initially on the host to supply these. Host transcription factors can bind to viral promoters and activate transcription. Genes expressed this way are typically named immediate early genes. These genes encode transcription factors and other regulators that enable transcription of viral genes. Genes dependent on the expression of the immediate early genes are classified as “early genes”. So-called “late genes” are expressed after initiation of viral DNA synthesis and typically include structural proteins [47]. A list of known interactions between host proteins and viral genes is provided in Table 2.

Table 2. Established WSSV–host gene expression interactions.

Transcription Factor (Host, Virus)	Target (Host, Virus)	Species	Reference
STAT (PmSTAT)	*ie1*	*Penaeus monodon*	[136]
PHB2 (Sf-PHB2)	*ie1*	*Spodoptera frugiperda*	[137]
Nf-κB (LvRelish, LvDorsal)	*ie1*, *WSSV303*, *WSSV371*	*Litopenaeus Vannamei*	[138–140]
c-JUN	*ie1*	*L. vannamei*	[141] (Bioinformatic prediction)
XBP1 (LvXBP1)	*wsv083*	*L. vannamei*	[142]
ATF4 (LvATF4)	*wsv023*	*L. vannamei*	[142]
KLF (PmKLF)	*WSSV108*	*P. monodon*	[143]
ATFβ (LvATFβ)	*wsv059*, *wsv166*	*L. vannamei*	[144]
VP38, VP41B	Caspase (PjCaspase)	*M. japonicus*	[145]
WSSV-miR-N24	Caspase 8	*M. japonicus*	[78]

Viral protein/gene/miRNA is underlined.

Immediate early genes expressed upon WSSV infection have been identified in infected crayfish hemocytes via the use of a protein synthesis inhibitor, cycloheximide (CHX) [146]. Inhibiting protein synthesis disables the ability of WSSV to create its own transcription factors and thus relies solely on host factors. Using this approach a total of 16 ORFs have been identified including WSV069 (IE1), *WSV051*, *WSV100*, *WSV079* with transactivation activity, *WSV083* with Ser/Thr kinase domain, and *WSV249*, believed to function as an ubiquitin E3 ligase [146]. In the Taiwan isolate of WSSV three WSSV ORFs (*WSSV126*, *WSSV242*, and *WSSV418*) have been identified that were insensitive to CHX treatment and were subsequently named *immediate early 1* (*ie1*), *ie2*, and *ie3*, respectively [147]. Further experiments established that the promoter of *ie1* could express EGFP in the fall army worm, *Spodoptera frugiperda*, Sf9 cells indicating that *ie1* can be activated even by non-decapod host transcription factors. Other work has shown that the *ie1* promoter of WSSV is one of the most inducible promoters in insect cells [148] and can regulate the expression of genes in mammalian cells also [149]. It has been suggested that the broad expression capabilities of *ie1* are a reason for the wide host range of WSSV [147]. Characterization of this promoter has revealed that *ie1* contains an initiator element, TATA-box, and a binding site for the transcription factor Sp1 [137]. Activation studies of the *ie1* promoter in natural hosts of WSSV have shown that it can be activated by various host factors.

In *P. monodon*, STAT (PmSTAT) can increase the activity of the *ie1* promoter through a STAT-binding motif [136]. Interestingly, STAT is part of the JAK-STAT anti-viral signaling pathway, indicating that WSSV can hijack the host immune response in order to promote expression of its own genes. Knockdown experiments of a cytokine receptor that activates the JAK-STAT pathway in *L. vannamei* (*LvDOME*) resulted in lower cumulative mortality and fewer WSSV copies, providing further evidence for the interaction between JAK-STAT and WSSV [150].

In addition to STAT, WSSV can hijack other immune-related pathways, notably Nuclear Factor-κ-B (NF-κB) signaling and MAP kinase signaling. NF-κB is a key regulator of the immune response and important for cell survival [151]. The *ie1* promoter contains a binding site for the NF-κB family of proteins [139]. In *L. vannamei* a homologue of NF-κB, LvRelish, can bind to the putative NF-κB binding site in the *ie1* promoter [139]. Other *in vivo* experiments confirm that NF-κB homologues lvDorsal and LvRelish can stimulate the expression of WSSV genes (e.g., *WSSV069*, *WSSV303*, and *WSSV371*) through interaction with *ie1* [140]. Furthermore, the expression of these transcription factors is upregulated upon WSSV infection [138,140]. The WSSV genome also encodes a protein (WSSV449) that shows similar functionality to Tube, a component of the NF-κB pathway. WSSV449 activates the host NF-κB pathway and through that system promotes expression of viral genes. WSSV, as for many other viruses (e.g., HIV-1 and hepatitis B), thus employs the NF-κB pathway for its own benefit [138]. The MAP kinase c-Jun N-terminal kinase (JNK) also has an ability to bind the *ie1* promoter, implicating involvement of MAP kinase signaling in WSSV gene expression. In *L. vannamei*, silencing of *LvJNK* with dsRNA resulted in decreased viral proliferation, and specific MAP kinase inhibitors delay viral gene expression [141]. A potential binding site for c-JUN has been identified in the *ie1* promoter via sequence similarity to c-JUN binding sites in TRANSFAC, a database of eukaryotic transcription factors and their binding sites. However, transcription factor binding has yet to be validated experimentally [141,152]. After expression of *ie1*, WSSV IE1 protein (a viral transcription factor) is then able to facilitate expression of the viral early genes. WSSV IE1 has been shown to accomplish this through cooperation with *P. monodon* TATA box-binding proteins (PmTBP) in transcription initiation [116].

WSSV108 (a probable transcription factor/activator and/or regulator through SUMOylation) is another WSSV immediate early gene [153]. Liu *et al.* [143] used the transcription factor binding site databases TRANSFAC and JASPAR to identify regulatory elements upstream of *WSSV108*, and found elements for Sp1/KLF, GATA-1, C/EBP, c-Myc, and AP-1. Furthermore, it was confirmed that recombinant Krüppel-like factor from *P. monodon* (rPmKLF) could bind to the KLF element and that a deletion in this element had the largest impact on expression through the *WSSV108* promoter [143]. The action of these transcription factors result in a higher expression level of the *WSSV108* promoter than *ie1* [143,146,153].

Expression of genes in the different phases of the infection process for WSSV has been reviewed previously by Sánchez-Paz [47]. Briefly, the genes can be grouped according to their function and they comprise of at least three classes: transcription factors, kinases, and ubiquitin E3 ligases [47]. Early phase genes target replication of the viral genome and include DNA polymerases, DNA helicases, and genes involved in nucleic acid metabolism [47]. After genome replication the functional characteristics of expressed genes shifts to structural proteins and to particle assembly during the late phase [47].

Different viruses exploit different phases of the host cell cycle for viral genome replication. Some arrest cells in G0, G1, or at the G1/S boundary so host DNA replication is prevented, whereas others tend to favor the S-phase where host DNA replication machinery is widely available [117]. Those that arrest host cells in G0/G1/G1/S, which include the herpes simplex virus and Epstein–Barr virus, often carry DNA replication machinery in their genomes instead of relying on the host [117,154,155]. WSSV appears to operate in the S-phase and depend on host factors for replication. Viruses often control the host cell cycle through interactions with proteins of the retinoblastoma (Rb) family, which are central regulators of the cell cycle, operating through interaction with E2F transcription factors [117]. WSSV IE1 and WSV056 (two paralogue genes) have been shown to be able to bind to an Rb-homologue in *L. vannamei* (lv-RBL), thereby possibly activating E2F1 and leading to S-phase entry. Indeed, overexpression of these two viral proteins in *Drosophila* S2 cells resulted in an increased portion of S-phase cells and a correlated decrease in G0/G1-phase cells [117]. The authors deduced that although WSSV carries some DNA replication machinery, it still relies on host factors; they therefore propose that WSSV utilizes the DNA replication machinery present in host cells during S-phase and promotes S-phase arrest to support successful replication of the viral genome.

Examples of host factors required for efficient replication are processivity factors. Processivity is the ability to catalyze consecutive reactions without releasing the substrate. While the DNA polymerase encoded in the WSSV genome (*WSV514*, [156]) shows polymerase activity [157], the average number of nucleotides added per DNA association event has been shown to be low [157]. Usually, processivity of DNA polymerases is improved by association with processivity factors, e.g., DNA clamps. WSSV does not encode its own processivity factors, which indicates that WSSV DNA polymerase might employ host processivity factors. Interaction of DNA polymerases with processivity factors occurs through a PIP-box motif. This motif is present on WSSV DNA polymerase [157]. To investigate the possibility of interaction with a host processivity factor in *L. vannamei*, the 3D structure of Proliferating Cell Nuclear Antigen (LvPCNA) was elucidated [104,158]. Likely PIP-box interaction models between LvPCNA and WSSV DNA polymerase were established but have yet to be confirmed experimentally.

3.4. Maintaining the Host Cell Environment

The presence of a virus within a host cell places demands on that cell, e.g., through draws on energy for anabolic reactions, demand for essential nutrients, and accumulation of non-host proteins. Such impacts on the cell lead to a deterioration of the cellular environment, making it less conducive for viral replication. Host cells also have evolved mechanisms to reduce the ability of viruses to replicate, for example through withholding nutrients in cases of infection or inhibiting translation. The cell might enter apoptosis, preventing further viral replication by undergoing self-destruction. To counteract these adverse cellular responses, viruses have evolved ways to interact with host metabolism, stress response systems, and apoptosis signaling for the purpose of retaining an appropriate environment for replication. Some of these interactions have been investigated for WSSV infections, but the limited understanding of the function of many WSSV proteins limits interpretation for some of the effects identified (Figure 2).

3.4.1. Metabolism

The replication of the viral genome and synthesis of its structural components require a large amount of energy. To supply the cell with such large quantities of energy, host metabolism can be directed to induce aerobic glycolysis [159]. Often observed in cells during rapid proliferation, a high rate of glycolysis provides both energy and glycolytic intermediates that can be used in a variety of anabolic reactions [159]. This metabolic shift was originally described in cancer cells and is known as the Warburg effect [160]. The Warburg effect also encompasses enhancement of the pentose phosphate pathway, amino acid metabolism, and lipid homeostasis [161]. Induction of a Warburg-like effect by WSSV infection has been observed in *L. vannamei* hemocytes [162]. In a study by Su *et al.* [139], it was established that the PI3K-Akt-mTOR pathway was the mechanism through which metabolic changes where induced by WSSV. PI3K-Akt-mTOR is also employed by cancer cells and human papillomavirus to induce the Warburg effect [161,163,164]. However, the viral factors that interact with the host PI3K-Akt-mTOR pathway have yet to be identified.

3.4.2. Iron

In addition to energy, WSSV and other invading pathogens also require essential nutrients including iron. In response to this, hosts have evolved mechanisms through which they can withhold iron from pathogens, thereby inhibiting their proliferation. The protein responsible for this mechanism is ferritin [105]. The ferritin defense mechanism has been demonstrated in *L. vannamei*, where its injection reduced susceptibility to WSSV infection. Higher amounts of ferritin resulted in a greater binding of iron and thus less iron available for WSSV proliferation. Conversely, dsRNA-mediated knockdown of ferritin resulted in a three-fold increase in viral copy number [165]. Through the use of a yeast two-hybrid experiment, it has been shown that WSSV protein kinase 1 (WSSV PK1) can interact with host ferritin and influence the availability of iron [105]. Binding of WSSV PK1 does not release iron from ferritin but rather prevents iron from binding to apoferritin (ferritin without bound iron). Injection of dsRNA specific to *WSSV PK1* decreases cumulative mortality, showing that disrupting host iron withholding mechanisms is important to successful WSSV infection [166].

3.4.3. Endoplasmic Reticulum Stress Responses

The presence of a virus within the cell can induce the unfolded protein response (UPR), a cellular stress response activated because of accumulation of misfolded protein at the ER. This leads to phosphorylation of transmembrane kinases and activation of transcription factors. Outputs of the UPR include global translation shutdown, arrest of cell cycle, increase in expression of chaperone genes to aid in protein folding, and potentially apoptosis [167]. While increased folding capabilities can aid the virus, translation attenuation and apoptosis have adverse effects on viral replication [168]. Therefore many viruses have evolved mechanisms through which the host UPR response can be manipulated [169].

There are three pathways that can initiate the UPR response, namely via Inositol-requiring enzyme-1 to X-box binding protein 1 (IRE1-XBP1), via double-stranded RNA-activated Protein kinase-like ER kinase and Activating Transcription Factor 4 (PERK-ATF4), and via Activating Transcription Factor 6 (ATF6). Interactions between viruses and these UPR pathways are complex, as some viruses inhibit responses whereas others induce them. For example, the Epstein–Barr virus (EBV) induces UPR to aid in lytic replication [170], Rotavirus induces and controls UPR to prevent ER stress-related cell death [171], and Herpes simplex virus-1 inhibits UPR to maintain an environment that permits expression of viral genes [172,173].

Interactions between WSSV and UPR in shrimp have been demonstrated through studies on the expression of chaperone proteins [174–178]. Induction of the IRE1-XBP1 UPR pathway upon WSSV infection in *L. vannamei* has been shown through enhanced expression of *LvXBP1.* Furthermore, WSSV appears to benefit from its induction, as demonstrated by lower cumulative mortality following dsRNA-mediated knockdown of *LvXBP1* [173]. A potential mechanism for this effect is LvXBP1-mediated upregulation of expression of the viral gene *wsv083*, a predicted protein kinase 2, and *wsv023*, of unknown function [49,142]. Wsv083 has been shown to inhibit focal adhesion kinase, a regulator of innate immune system signaling [106,142]. The transcription factors of the PREK-ATF4 pathway can also induce expression of WSSV genes, depending on the presence of an

ATF/CRE in the promoter (15 in total: *wsv023*, *wsv049*, *wsv064*, *wsv069*, *wsv138*, *wsv242*, *wsv256*, *wsv282*, *wsv303*, *wsv306*, *wsv313*, *wsv321*, *wsv343*, *wsv406*, and *wsv453*). In *L. vannamei* the UPR transcription factor LvATF4 (as for LvXBP1) also upregulates expression of *wsv023* [142].

The shutdown of translation following UPR activation is achieved through phosphorylation of the translation initiation factor subunit eIF2α [179]. Phosphorylated eIF2α inhibits activation of eIF2 by preventing the exchange of GDP for GTP by its eIF2β subunit. During WSSV infection in *L. vannamei* the level of LveIF2α decreases and the phosphorylation ratio increases, suggesting that the virus indeed initiates eIF2α-mediated translation inhibition [180]. Furthermore, adding an inhibitor of eIF2α phosphatases decreases viral loads, indicating that WSSV requires active eIF2 for successful replication. Xu *et al.* [180] suggest that WSSV could code for proteins able to prevent phosphorylation of eIF2α, thereby halting UPR initiated translation inhibition, a strategy that is also observed in the Chikungunya virus. In summary, data to date suggest that the presence of WSSV induces the UPR and is capable of interacting with downstream effectors and transcription factors activated through the UPR. These processes aid in the transcription of viral genes. However, limited knowledge of viral protein function limits our current understanding on the roles of these viral genes.

3.4.4. Apoptosis

Cellular responses to stress can result in the induction of apoptosis of the host cell. Through apoptosis organisms can remove cells that are potentially harmful to their own health. Apoptosis is an important defense mechanism against abnormalities in cell programming and disease infection, including viruses. The relationship between virus and apoptosis is complex and viruses can influence the host apoptosis system by either inhibiting or inducing it. Inhibition is necessary to keep a cellular environment conducive to the production of new virions [181]. In later stages of viral infection, however, some viruses might induce apoptosis as a means of leaving cells and spreading further throughout the host [182]. Apoptosis has been documented to occur in *L. vannamei*, *P. monodon*, and *M. japonicus* cells infected with WSSV [183–186]. Leu *et al.* [187] proposed a detailed model for the apoptotic interaction between WSSV and shrimp, in which invasion of WSSV leads to activation of signaling pathways that increase expression of pro-apoptosis proteins (e.g., Caspases and voltage-dependent anion channels), membrane permeabilization of mitochondria, and increased oxidative stress. These molecular pathways lead to the initiation of the apoptosis program. In parallel, WSSV anti-apoptotic proteins attempt to block apoptosis and thereby keep the cell viable for replication. The balance between the pro- and anti-apoptosis activation processes will determine the fate of the WSSV-infected cell [187].

Caspases are key regulators of apoptosis and therefore important targets for viruses. The family can be generally grouped into initiator caspases and effector caspases. The first are activated through autocatalytic processes, whereas the latter are present as zymogens activated through cleavage by the initiator caspases [188]. Activation of effector caspases leads to an accelerated feedback loop of effector caspase activation and eventually permits the controlled destruction of cellular components [189]. WSSV proteins show interactions with shrimp caspases. Direct inhibition of the activity of a *P. monodon* effector caspase (PmCaspase) by the WSSV protein anti-apoptosis protein 1 (AAP1 or WSSV449) has been shown in Sf9 cells [107]. AAP1 binds to, and is cleaved by, PmCaspase at two possible sites, with only one resulting in PmCaspase inhibition [107]. In a similar cellular system another effector caspase in *P. monodon*, PmCasp, can be bound by viral proteins WSSV134 and WSSV322, resulting in anti-apoptotic activity [108]. However, inhibition of PmCasp by AAP1 and WSSV449 does not occur, illustrating the diversity of effector caspases and their activity [108]. A different method of host caspase regulation has been revealed in *M. japonicus*. Here, the caspase gene *PjCaspase* was identified and shown to be upregulated in survivors of WSD [190]. Silencing of this gene resulted in inhibition of apoptosis, and the subsequent increase in viral copy number showed that induction of apoptosis in infected cells is beneficial to the host. Zuo *et al.* [145] have also shown that WSSV can regulate the expression of caspase. Viral proteins VP38 (WSV259) and VP41B (WSV242) were capable of binding the *PjCaspase* promoter, the first acting as a repressor and the latter acting as an activator. Interestingly, both of these WSSV proteins are envelope proteins: VP41B having a potential transmembrane domain and VP38 through association with envelope protein VP24 [191,192]. The capability of WSSV to both repress and promote caspase genes follows the observation that some viruses induce apoptosis to facilitate departure from the host cell.

Yet another route through which WSSV can induce apoptosis is through ubiquitination, a process by which attachment of ubiquitin to a protein can flag it for degradation in proteasomes. This mechanism plays an important role in regulation of apoptosis as many substrates of ubiquitination are regulatory proteins for apoptosis [193]. Ubiquitin is activated by an ubiquitin-activating enzyme (E1), subsequently transferred to a ubiquitin-conjugating enzyme (E2) and, through interaction with an E3-ligase, transferred to the target protein [194]. WSSV249 has been shown to be involved in ubiquitination by acting as an E3-ligase in cooperation with the conjugating enzyme PvUbc in *L. vannamei* [109]. This WSSV249 interaction is accomplished through its RING-H2 domain, a domain associated with E3-ligases [46,195]. RING domains are present in four WSSV proteins, namely WSSV199, WSSV222, WSSV249, and WSSV403 [109,196]. The RING-H2 domain of WSSV222 enables it to perform as an E3-ligase in the ubiquitination of tumor suppressor-like protein, thereby inhibiting apoptosis [197]. Whether any of the other RING domain containing proteins have apoptosis inhibiting function remains to be determined. Together with caspase inhibition, the success of ubiquitination of pro-apoptotic proteins will determine the fate of the host cell [187].

Wang *et al.* [110,198] have suggested a third virus–host interaction that could influence the apoptotic state of the host cell, involving the highly expressed WSSV protein ICP11. The crystal structure of dimers of ICP11 indicates that it could act as a DNA mimic [110]. The electrostatic surface of ICP11 shows patches of negatively charged amino acids that are arranged in two rows and at similar distances as dsDNA phosphate groups [110]. Furthermore, it was shown that ICP11 could interfere with the binding of host DNA to histones (H3) in HeLa cells [110]. This binding can lead to disruption of the host nucleosome assembly and even apoptosis [110]. An alternative crystal structure of the dimer of VP9 (ICP11) has been proposed that does not show such rows of negative charges [199]. Instead, VP9 shows structural folds that bear resemblance to E2, a transcription/replication factor of the human papillomavirus [111,200]. In *P. monodon* it has been shown that VP9 can interact with a receptor for activated protein kinase C1 (PmRACK1), using the yeast two-hybrid and GST pulldown assay, [111]. Mammalian RACK1 receptors are involved in a large variety of functions including cell signaling pathways, cell development, and the immune response, and can interact with a large number of viral proteins [111]. The interaction between these two proteins may be involved in intracellular VP9 functions, for example by transporting VP9 to the nucleus [111].

3.4.5. Particle Assembly and Release

After replication of the viral genome and production of structural proteins, these components come together to be assembled into new virions. One of the most challenging aspects of viral particle assembly is packaging the WSSV 300 kb genome into the nucleocapsid. In eukaryotes, DNA is compacted in nucleosomes but this does not occur in viruses. In dsDNA viruses, DNA packaging is often accomplished through interactions between DNA and nucleocapsid proteins [201]. A small (6.7 kDa) viral protein, VP15, has been associated with WSSV DNA packaging. VP15 is a basic protein that shows homology to putative baculovirus DNA-binding proteins [202]. Studies have shown that VP15 can form homomultimers and is able to bind to (preferably supercoiled) DNA [202]. Application of Atomic Force Microscopy has shown that VP15 is able to condense DNA and that VP15-induced DNA condensates resemble packaged viral DNA [201]. In the packaging process VP15 can interact with host proteins. In *P. monodon* binding has been shown between VP15 and PmFKBP46, an immunophilins-like protein [112]. The role of PmFKBP46 in *P. monodon* is not known but human and yeast homologues are involved in histone deacetylation and act as a histone chaperone, respectively [112,203,204]. A more recent study uncovered an interaction between VP15 and calreticulin

(CRT), which could also act as a histone chaperone [101,205]. DsRNA-mediated knockdown of CRT resulted in a significant decrease in viral DNA duplication and viral gene transcription, indicating that the interaction between CRT and VP15 is necessary for viral replication. It has also been suggested that the interaction between VP15 and CRT can play a role in the export of viral RNA from the nucleus, a role that CRT has been shown to fulfill for glucocorticoid receptors [101,206].

Morphogenesis of the WSSV envelope requires long chain fatty acids (LCFA). During the early stages of infection fatty acids in the host cell are depleted in order to generate the energy required for viral replication, but during the late stage of infection LCFAs are upregulated, thus replenishing the LCFA that have been used [207]. Applying an inhibitor of fatty acid synthase (FAS) that blocked the replenishment of LCFA was shown to result in impaired virion formation [207]. There is little information available on how assembled WSSV virions are subsequently released from the host cell.

4. Current Treatment Options for WSD

Despite the large body of work, our understanding of WSSV infection and pathogenesis are far from complete. Unresolved questions include those focused on mechanisms of endosomal escape, virion assembly, exit from the host cell, and the role of miRNA regulation. Nevertheless, because of the urgent need for disease mitigation (and prevention) avenues have been explored for potential therapy. Here we discuss some of the more promising results that include strategies spanning immune system activation, vaccinations, RNAi, and the application of herbal extracts.

Early studies investigated whether WSD could be prevented by enhancing immune competence of the host. This can be achieved by feeding with agents containing pathogen-associated molecular patterns (PAMP) that are known to activate the host innate immune system. It was found that feeding *P. japonicus* with peptidoglycans over a period of time increased the phagocytic activity of host granulocytes, resulting in a significant decrease in mortality upon WSSV exposure [208]. Similarly, injection of β-glucan prior to WSSV infection has been shown to result in activation of the prophenoloxidase system and subsequently a reduction of mortality (25%–50% as compared to 100% in controls). However, repeated dosages of β-glucan cause high mortality rates in the host, which is probably due to excess generation of reactive oxygen species [209].

Invertebrates do not possess an adaptive immune system and they thus rely on innate immune defenses. As a consequence, it is not possible to develop vaccines in the traditional way as is done in mammals and other vertebrates. However, recent developments have provided evidence that certain forms of pathogen-specific "immune priming" are possible in some invertebrates [210]. For example, short-term protection against WSSV can be achieved through exposing shrimp to inactivated viral particles [211]. Musthaq and Kwang reviewed the possibilities of such vaccinations for WSD and the types of treatments (vaccines) that can provide temporal protection against WSD [210]. These include inactivated virus, recombinant viral proteins, viral DNA, and double-stranded RNA, and often involve WSSV envelope proteins like VP28 because of their importance to the infection process. However, all of these treatments provide temporal protection that does not usually last beyond 14 days. New generation technologies are attempting to provide more efficient delivery systems for these treatments. For example, baculovirus and *Bacillus subtilis* spores, both modified to express VP28, could convey protection via oral vaccination [212,213]. Elucidation of the mechanisms underlying invertebrate immune priming could lead to further gains in vaccine efficacy and provide an optimized solution.

RNAi can potentially play a significant part in host–pathogen interactions. Hosts can express small RNA molecules that target and inhibit the expression of viral proteins. It has been shown that *M. japonicus* can generate small interfering RNA (siRNA) that targets *vp28* (*vp28-siRNA*) in response to infection by WSSV. Blocking siRNA synthesis resulted in increased viral copy numbers, indicating that RNAi had a protective effect for the host [214]. In other work *vp28-siRNA*s encapsulated with β-1,3-D-glucan and injected along with WSSV in *M. japonicus* [215] was shown to inhibit WSSV replication, illustrating a potential avenue for treatment development.

Another direction of research focused on combating WSD has been the utilization of plant extracts with potential pharmaceutical activity. Large screens with extracts from plants have been carried out in an attempt to identify chemicals with anti-WSSV properties and there have been a number of successes. Examples include extracts of *Cynodon dactylon* and *Ceriops tagal* [216,217] that have shown protective effects against WSSV in *P. monodon*. Extracts from the seaweed *Sagrassum weighti* have been shown to have a significant effect in reducing WSSV infection in both the Indian prawn (*P. indicus*) and freshwater crab (*Paratelphusa hydrodomous*) [218,219]. Oral administration of a plant extract from *Momordica charantia* has been shown to result in an 86% survivorship in *L. vannamei* infected with WSSV [219]. Furthermore, several diets based on extracts from *Agathi grandiflora* have also been shown to decrease mortality rates substantially in WSSV-infected *Fenneropenaeus indicus* [220]. The nature of the pharmaceutically active components of these extracts, and how they interact with host and virus, is unknown for the treatments described. Filling these knowledge gaps could result in a better understanding of WSSV infection and aid in optimizing efficacy in the use of these materials in preventing WSD.

5. Future Perspectives

In recent years WSD has received significant scientific attention, driven by the commercial impacts of this disease on shrimp. Global research efforts into understanding the biology of WSSV and infection process for WSD have contributed to the formulation of strategies for disease treatment and prevention. Molecular studies have identified a substantial number of WSSV envelope and cell surface proteins involved in the first stage of virus infection, established a large number of host transcription factors replication of WSSV, and established some of the mechanisms by which the virus maintains a favorable host cell environment including through arresting the cell cycle, changing metabolism, and preventing apoptosis. Major gaps, however, remain in our understanding of how the virus, upon entering the host cell via endosomes, subsequently delivers its genome to the host nucleus, the virion assembly process, and in fact the function of the majority of WSSV proteins.

Research on WSD infection has been complicated by the lack of well-annotated genomic resources for host species. There has been a reliance placed on sequence similarities with other (sequenced) species to provide required annotation. Sequence information for aquatic crustaceans, however, is extremely limited. Application of modern sequencing technologies now allows for relatively rapid generation of the required sequence information to support both *de novo* genome assemblies and for transcriptome analyses, and this work needs to be encouraged to provide the required molecular resources for commercially important species as a minimum. Current research on WSSV infection has focused overwhelmingly on susceptible shrimp species, which is not surprising given their economic importance, but opportunities could lie in studies on more resistant (and perhaps less popular) host organisms. Understanding how those organisms resist WSD could equally lead to effective disease treatments for shrimp.

Understanding the WSSV infection process completely may not be necessary in order to produce effective therapeutics. Current knowledge provides a plethora of potential pathways/genes/miRNAs that could serve as potential treatment targets and indeed there has been some success in applying targeted molecular approaches in the treatment of WSD. To date, however, these treatments applied in a laboratory context have not been successfully applied in the field. As an example, RNAi has shown great potential as therapeutic technology and indeed VP28-siRNA has been shown to be effective against WSSV. However, it requires delivery through injections, which is not viable for commercial application. The development of an effective delivery method for siRNA that can be used in shrimp farms would be a significant step forward in the prevention of WSD and would also have great potential for use in the treatment of other viruses that impact aquatic Crustacea. Other prospects include production of genetically modified shrimp for resistance to WSD, for example by enhancing anti-WSSV RNAi.

Another potential avenue for exploration in combating WSD relates to resistance that has developed in certain hosts through integration of viral DNA into the host genome, conveying resistance to the inserted virus (and sometimes also closely related pathogens). If such a mechanism of resistance could be transferred from resistant to susceptible shrimp species, this could result in the production of shrimp lines that are resistant to WSSV.

Bringing this review to a close, we would emphasize that while understanding the molecular basis of WSD is likely to lead to possible intervention strategies for combating WSD, the practical implementation of this knowledge, given the nature of the shrimp farming industry, will require the resulting treatments to be cheap, easy to use, and easily distributed. It should also be recognized that there are many other factors that need due consideration in our attempts to develop improved treatments for, and to combat, WSD. For example, it is the case that outbreaks of WSD in farming will inevitably depend on the health status of host organisms and the environments in which they live and this is not necessarily determined by individual pathogens alone but by a combination of local abiotic and biotic factors, pathogen assemblages, and pathogen loads in host tissues. Almost nothing is yet known in this regard for WSD. Indeed, WSSV may be endemic in aquaculture and may only occur under certain conditions. Identifying those conditions and the biological indicators associated with health status and disease outbreaks in shrimp aquaculture ponds could help in predicting and pre-empting WSD outbreaks, allowing for intervention strategies, and uncoupling the ability of WSSV to interact with its host. More effective and intelligent surveillance systems for preventing the spread of WSD outbreaks are also a major research need.

Acknowledgments: This work was funded by the Open Innovation Platform at the University of Exeter (Open Innovation Fund Initiative PHSW029) and by the Centre for Environment, Fisheries and Aquaculture Science (Cefas) (under seedcorn project DP318 to GDS) under the Strategic Alliance partnership between the University of Exeter and Cefas.

Author Contributions: Bas Verbruggen wrote the paper. Kelly S. Bateman produced the histology images. Lisa K. Bickley, Ronny van Aerle, Grant D. Stentiford, Eduarda M. Santos, and Charles R. Tyler had significant contributions through writing, editing, and intellectual input.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Stentiford, G.D.; Oidtmann, B.; Scott, A.; Peeler, E.J. Crustacean diseases in European legislation: Implications for importing and exporting nations. *Aquaculture* **2010**, *306*, 27–34. [CrossRef]
2. Stentiford, G.D.; Bonami, J.R.; Alday-Sanz, V. A critical review of susceptibility of crustaceans to taura syndrome, yellowhead disease and white spot disease and implications of inclusion of these diseases in European legislation. *Aquaculture* **2009**, *291*, 1–17. [CrossRef]
3. Nakano, H.; Koube, H.; Umezawa, S.; Momoyama, K.; Hiraoka, M.; Inouye, K.; Oseko, N. Mass mortalities of cultured kuruma shrimp, *Penaeus japonicus*, in Japan in 1993: Epizootiological survey and infection trails. *Fish Pathol.* **1994**, *29*, 135–139. [CrossRef]
4. Flegel, T.W. Special topic review: Major viral diseases of the black tiger prawn (*Penaeus monodon*) in thailand. *World J. Microbiol. Biotechnol.* **1997**, *13*, 433–442. [CrossRef]
5. Mohan, C.V.; Shankar, K.M.; Kulkarni, S.; Sudha, P.M. Histopathology of cultured shrimp showing gross signs of yellow head syndrome and white spot syndrome during 1994 Indian epizootics. *Dis. Aquat. Org.* **1998**, *34*, 9–12. [CrossRef] [PubMed]
6. Zhan, W.B.; Wang, Y.H.; Fryer, J.L.; Yu, K.K.; Fukuda, H.; Meng, Q.X. White spot syndrome virus infection of cultured shrimp in China. *J. Aquat. Anim. Health* **1998**, *10*, 405–410. [CrossRef]
7. Wang, Q.; Poulos, B.T.; Lightner, D.V. Protein analysis of geographic isolates of shrimp white spot syndrome virus. *Arch. Virol.* **2000**, *145*, 263–274. [CrossRef] [PubMed]
8. Stentiford, G.D.; Lightner, D.V. Cases of white spot disease (WSD) in European shrimp farms. *Aquaculture* **2011**, *319*, 302–306. [CrossRef]
9. Lightner, D.V. Global transboundary disease politics: The OIE perspective. *J. Invertebr. Pathol.* **2012**, *110*, 184–187. [CrossRef] [PubMed]

10. Flegel, T.W.; Lightner, D.V.; Owens, L. Shrimp disease control: Past, present and future. *Dis. Asian Aquacult.* **2008**, *6*, 355–378.
11. Stentiford, G.D.; Neil, D.M.; Peeler, E.J.; Shields, J.D.; Small, H.J.; Flegel, T.W.; Vlak, J.M.; Jones, B.; Morado, F.; Moss, S.; *et al*. Disease will limit future food supply from the global crustacean fishery and aquaculture sectors. *J. Invertebr. Pathol.* **2012**, *110*, 141–157. [CrossRef] [PubMed]
12. Shields, J.D. Research priorities for diseases of the blue crab *Callinectes sapidus*. *Bull. Mar. Sci.* **2003**, *72*, 505–517.
13. Wang, C.H.; Lo, C.F.; Leu, J.H.; Chou, C.M.; Yeh, P.Y.; Chou, H.Y.; Tung, M.C.; Chang, C.F.; Su, M.S.; Kou, G.H. Purification and genomic analysis of baculovirus associated with white spot syndrome (WSBV) of *Penaeus monodon*. *Dis. Aquat. Org.* **1995**, *23*, 239–242. [CrossRef]
14. Mayo, M. A summary of taxonomic changes recently approved by ictv. *Arch. Virol.* **2002**, *147*. [CrossRef] [PubMed]
15. Mayo, M. Virus taxonomy—Houston 2002. *Arch. Virol.* **2002**, *147*. [CrossRef]
16. Van Hulten, M.C.; Witteveldt, J.; Peters, S.; Kloosterboer, N.; Tarchini, R.; Fiers, M.; Sandbrink, H.; Lankhorst, R.K.; Vlak, J.M. The white spot syndrome virus DNA genome sequence. *Virology* **2001**, *286*, 7–22. [CrossRef] [PubMed]
17. Vlak, J.M.; Bonami, J.R.; Flegel, T.W.; Kou, G.H.; Lightner, D.V.; Lo, C.F.; Loh, P.C.; Walker, P.W. Nimaviridae. In *Eighth Report of the International Committee on Taxonomy of Viruses*; Elsevier/Academic Press: Cambridge, MA, USA, 2005; pp. 187–192.
18. Pradeep, B.; Rai, P.; Mohan, S.A.; Shekhar, M.S.; Karunasagar, I. Biology, host range, pathogenesis and diagnosis of white spot syndrome virus. *Indian J. Virol.* **2012**, *23*, 161–174. [CrossRef] [PubMed]
19. Chou, H.Y.; Huang, C.Y.; Wang, C.H.; Chiang, H.C.; Lo, C.F. Pathogenicity of a baculovirus infection causing white spot syndrome in cultured penaeid shrimp in Taiwan. *Dis. Aquat. Org.* **1995**, *23*, 165–173. [CrossRef]
20. Rajan, P.R.; Ramasamy, P.; Purushothaman, V.; Brennan, G.P. White spot baculovirus syndrome in the indian shrimp *Penaeus monodon* and *P. Indicus*. *Aquaculture* **2000**, *184*, 31–44. [CrossRef]
21. Wang, Y.G.; Hassan, M.D.; Shariff, M.; Zamri, S.M.; Chen, X. Histopathology and cytopathology of white spot syndrome virus (WSSV) in cultured *Penaeus monodon* from peninsular malaysia with emphasis on pathogenesis and the mechanism of white spot formation. *Dis. Aquat. Org.* **1999**, *39*, 1–11. [CrossRef] [PubMed]
22. Lightner, D.V. *A Handbook of Shrimp Pathology and Diagnostic Procedures for Diseases of Cultured Penaeid Shrimp*; World Aquaculture Society: Baton Roughe, LA, USA, 1996; p. 304.
23. Durand, S.; Lightner, D.V.; Redman, R.M.; Bonami, J.R. Ultrastructure and morphogenesis of white spot syndrome baculovirus (WSSV). *Dis. Aquat. Org.* **1997**, *29*, 205–211. [CrossRef]
24. Nadala, E.C.B.; Loh, P.C. A comparative study of three different isolates of white spot virus. *Dis. Aquat. Org.* **1998**, *33*, 231–234. [CrossRef] [PubMed]
25. Oidtmann, B.; Stentiford, G.D. White spot syndrome virus (WSSV) concentrations in crustacean tissues: A review of data relevant to assess the risk associated with commodity trade. *Transbound. Emerg. Dis.* **2011**, *58*, 469–482. [CrossRef] [PubMed]
26. Walker, P.J.; Mohan, C.V. Viral disease emergence in shrimp aquaculture: Origins, impact and the effectiveness of health management strategies. *Rev. Aquacult.* **2009**, *1*, 125–154. [CrossRef]
27. Momoyama, K.; Hiraoka, M.; Nakano, H.; Koube, H.; Inouye, K.; Oseko, N. Mass mortalities of cultured kuruma shrimp, *Penaeus japonicus*, in japan in 1993: Histopathological study. *Fish Pathol.* **1994**, *29*, 141–148. [CrossRef]
28. Takahashi, Y.; Itami, T.; Kondom, M.; Maeda, M.; Fuji, R.; Tomonaga, S.; Supamattaya, K.; Boonyaratpalin, S. Electron microscopic evidence of baciliform virus infection in kuruma shrimp (*Penaeus japonicus*). *Fish Pathol.* **1994**, *29*, 121–125. [CrossRef]
29. Cai, S.; Huang, J.; Wang, C.; Song, X.; Sun, X.; Yu, J.; Zhang, Y.; Yang, C. Epidemiological studies on the explosive epidemic disease of prawn in 1993–1994. *J. Fishery Sci. China* **1995**, *19*, 112–117.
30. Chen, L.L.; Lo, C.F.; Chiu, Y.L.; Chang, C.F.; Kou, G.H. Natural and experimental infection of white spot syndrome virus (WSSV) in benthic larvae of mud crab *Scylla setrata*. *Dis. Aquat. Org.* **2000**, *40*, 157–161. [CrossRef] [PubMed]
31. Sahul Hameed, A.S.; Charles, M.X.; Anilkumar, M. Tolerance of *Macrobrachium rosenbergii* to white spot syndrome virus. *Aquaculture* **2000**, *183*, 207–213. [CrossRef]

32. Hossain, M.S.; Chakraborty, A.; Joseph, B.; Otta, S.K.; Karunasagar, I. Detection of new hosts for white spot syndrome virus of shrimp using nested polymerase chain reaction. *Aquaculture* **2001**, *198*, 1–11. [CrossRef]
33. Jiravanichpaisal, P.; Bangyeekhun, E.; Soderhall, K.; Soderhall, I. Experimental infection of white spot syndrome virus in freshwater crayfish *Pacifastacus leniusculus*. *Dis. Aquat. Org.* **2001**, *47*, 151–157. [CrossRef] [PubMed]
34. Rodriguez, J.; Bayot, B.; Amano, Y.; Panchana, F.; de Blas, I.; Alday, V.; Calderon, J. White spot syndrome virus infection in cultured *Penaeus vannamei* (boone) in ecuador with emphasis on histopathology and ultrastructure. *J. Fish Dis.* **2003**, *26*, 439–450. [CrossRef] [PubMed]
35. Yoganandhan, K.; Thirupathi, S.; Sahul Hameed, A.S. Biochemical, physiological and hematological changes in white spot syndrome virus-infected shrimp, *Penaeus indicus*. *Aquaculture* **2003**, *221*, 1–11. [CrossRef]
36. Bateman, K.S.; Tew, I.; French, C.; Hicks, R.J.; Martin, P.; Munro, J.; Stentiford, G.D. Susceptibility to infection and pathogenicity of white spot disease (WSD) in non-model crustacean host taxa from temperate regions. *J. Invertebr. Pathol.* **2012**, *110*, 340–351. [CrossRef] [PubMed]
37. Lo, C.F.; Ho, C.H.; Chen, C.H.; Liu, K.F.; Chiu, Y.L.; Yeh, P.Y.; Peng, S.E.; Hsu, H.C.; Liu, H.C.; Chang, C.F.; *et al*. Detection and tissue tropism of white spot syndrome baculovirus (WSBV) in captured brooders of *Penaeus monodon* with a special emphasis on reproductive organs. *Dis. Aquat. Org.* **1997**, *30*, 53–72. [CrossRef]
38. Chou, H.Y.; Huang, C.Y.; Lo, C.F.; Kou, G.H. Studies on transmission of white spot syndrome associated baculovirus (WSBV) in *Penaeus monodon* and *P. japonicus* via waterborne contact and oral ingestion. *Aquaculture* **1998**, *164*, 263–276. [CrossRef]
39. Lotz, J.M.; Soto, M.A. Model of white spot syndrome virus (WSSV) epidemics in *Litopenaeus vannamei*. *Dis. Aquat. Org.* **2002**, *50*, 199–209. [CrossRef] [PubMed]
40. Tuyen, N.X.; Verreth, J.; Vlak, J.M.; de Jong, M.C. Horizontal transmission dynamics of white spot syndrome virus by cohabitation trials in juvenile *Penaeus monodon* and *P. vannamei*. *Prev. Vet. Med.* **2014**, *117*, 286–294. [CrossRef] [PubMed]
41. Lo, C.F.; Ho, C.H.; Peng, S.E.; Chen, C.H.; Hsu, H.C.; Chiu, Y.L.; Chang, C.F.; Liu, K.F.; Su, M.S.; Wang, C.H.; *et al*. White spot syndrome baculovirus (WSBV) detected in cultured and captured shrimp, crabs and other arthropods. *Dis. Aquat. Org.* **1996**, *27*, 215–225. [CrossRef]
42. Waikhom, G.; John, K.R.; George, M.R.; Jeyaseelan, M.J.P. Differential host passaging alters pathogenicity and induces genomic variation in white spot syndrome virus. *Aquaculture* **2006**, *261*, 54–63. [CrossRef]
43. Nunan, L.M.; Poulos, B.T.; Lightner, D.V. The detection of white spot syndrome virus (WSSV) and yellow head virus (YHV) in imported commodity shrimp. *Aquaculture* **1998**, *160*, 19–30. [CrossRef]
44. Durand, S.V.; Tang, K.F.J.; Lightner, D.V. Frozen commodity shrimp: Potential avenue for introduction of white spot syndrome virus and yellow head virus. *J. Aquat. Anim. Health* **2000**, *12*, 128–135. [CrossRef]
45. Li, F.; Xiang, J. Signaling pathways regulating innate immune responses in shrimp. *Fish Shellfish Immunol.* **2013**, *34*, 973–980. [CrossRef] [PubMed]
46. Shekhar, M.S.; Ponniah, A.G. Recent insights into host-pathogen interaction in white spot syndrome virus infected penaeid shrimp. *J. Fish Dis.* **2014**, *38*, 599–612. [CrossRef] [PubMed]
47. Sanchez-Paz, A. White spot syndrome virus: An overview on an emergent concern. *Vet. Res.* **2010**, *41*, 43. [CrossRef] [PubMed]
48. Sritunyalucksana, K.; Utairungsee, T.; Sirikharin, R.; Srisala, J. Virus-binding proteins and their roles in shrimp innate immunity. *Fish Shellfish Immunol.* **2012**, *33*, 1269–1275. [CrossRef] [PubMed]
49. Yang, F.; He, J.; Lin, X.; Li, Q.; Pan, D.; Zhang, X.; Xu, X. Complete genome sequence of the shrimp white spot bacilliform virus. *J. Virol.* **2001**, *75*, 11811–11820. [CrossRef] [PubMed]
50. Chai, C.Y.; Yoon, J.; Lee, Y.S.; Kim, Y.B.; Choi, T.J. Analysis of the complete nucleotide sequence of a white spot syndrome virus isolated from Pacific white shrimp. *J. Microbiol.* **2013**, *51*, 695–699. [CrossRef] [PubMed]
51. Marks, H.; Goldbach, R.W.; Vlak, J.M.; van Hulten, M.C. Genetic variation among isolates of white spot syndrome virus. *Arch. Virol.* **2004**, *149*, 673–697. [CrossRef] [PubMed]
52. Sindhupriya, M.; Saravanan, M.; Otta, S.K.; Bala Amarnath, C.; Arulraj, R.; Bhuvaneswari, T.; Ezhil Praveena, P.; Jithendran, K.P.; Ponniah, A.G. White spot syndrome virus (WSSV) genome stability maintained over six passages through three different penaeid shrimp species. *Dis. Aquat. Org.* **2014**, *111*, 23–29. [CrossRef] [PubMed]

53. Shekar, M.; Pradeep, B.; Karunasagar, I. White spot syndrome virus: Genotypes, epidemiology and evolutionary studies. *Indian J. Virol.* **2012**, *23*, 175–183. [CrossRef] [PubMed]
54. Marks, H.; Vorst, O.; van Houwelingen, A.M.; van Hulten, M.C.; Vlak, J.M. Gene-expression profiling of white spot syndrome virus *in vivo*. *J. Gen. Virol.* **2005**, *86*, 2081–2100. [CrossRef] [PubMed]
55. Sablok, G.; Sanchez-Paz, A.; Wu, X.; Ranjan, J.; Kuo, J.; Bulla, I. Genome dynamics in three different geographical isolates of white spot syndrome virus (WSSV). *Arch. Virol.* **2012**, *157*, 2357–2362. [CrossRef] [PubMed]
56. Tsai, J.M.; Wang, H.C.; Leu, J.H.; Hsiao, H.H.; Wang, A.H.; Kou, G.H.; Lo, C.F. Genomic and proteomic analysis of thirty-nine structural proteins of shrimp white spot syndrome virus. *J. Virol.* **2004**, *78*, 11360–11370. [CrossRef] [PubMed]
57. Tsai, J.M.; Wang, H.C.; Leu, J.H.; Wang, A.H.; Zhuang, Y.; Walker, P.J.; Kou, G.H.; Lo, C.F. Identification of the nucleocapsid, tegument, and envelope proteins of the shrimp white spot syndrome virus virion. *J. Virol.* **2006**, *80*, 3021–3029. [CrossRef] [PubMed]
58. Leu, J.-H.; Tsai, J.-M.; Wang, H.-C.; Wang, A.H.J.; Wang, C.-H.; Kou, G.-H.; Lo, C.-F. The unique stacked rings in the nucleocapsid of the white spot syndrome virus virion are formed by the major structural protein VP664, the largest viral structural protein ever found. *J. Virol.* **2005**, *79*, 140–149. [CrossRef] [PubMed]
59. Liu, W.J.; Yu, H.T.; Peng, S.E.; Chang, Y.S.; Pien, H.W.; Lin, C.J.; Huang, C.J.; Tsai, M.F.; Huang, C.J.; Wang, C.H.; *et al*. Cloning, characterization, and phylogenetic analysis of a shrimp white spot syndrome virus gene that encodes a protein kinase. *Virology* **2001**, *289*, 362–377. [CrossRef] [PubMed]
60. Tsai, M.F.; Lo, C.F.; van Hulten, M.C.; Tzeng, H.F.; Chou, C.M.; Huang, C.J.; Wang, C.H.; Lin, J.Y.; Vlak, J.M.; Kou, G.H. Transcriptional analysis of the ribonucleotide reductase genes of shrimp white spot syndrome virus. *Virology* **2000**, *277*, 92–99. [CrossRef] [PubMed]
61. Chen, L.L.; Leu, J.H.; Huang, C.J.; Chou, C.M.; Chen, S.M.; Wang, C.H.; Lo, C.F.; Kou, G.H. Identification of a nucleocapsid protein (*vp35*) gene of shrimp white spot syndrome virus and characterization of the motif important for targeting vp35 to the nuclei of transfected insect cells. *Virology* **2002**, *293*, 44–53. [CrossRef] [PubMed]
62. Kang, S.T.; Wang, H.C.; Yang, Y.T.; Kou, G.H.; Lo, C.F. The DNA virus white spot syndrome virus uses an internal ribosome entry site for translation of the highly expressed nonstructural protein ICP35. *J. Virol.* **2013**, *87*, 13263–13278. [CrossRef] [PubMed]
63. Kang, S.T.; Leu, J.H.; Wang, H.C.; Chen, L.L.; Kou, G.H.; Lo, C.F. Polycistronic mrnas and internal ribosome entry site elements (IRES) are widely used by white spot syndrome virus (WSSV) structural protein genes. *Virology* **2009**, *387*, 353–363. [CrossRef] [PubMed]
64. Han, F.; Zhang, X. Internal initiation of mrna translation in insect cell mediated by an internal ribosome entry site (IRES) from shrimp white spot syndrome virus (WSSV). *Biochem. Biophys. Res. Commun.* **2006**, *344*, 893–899. [CrossRef] [PubMed]
65. Gale, M.; Tan, S.L.; Katze, M.G. Translational control of viral gene expression in eukaryotes. *Microbiol. Mol. Biol. Rev.* **2000**, *64*, 239–280. [CrossRef] [PubMed]
66. Bartel, D.P.; Chen, C.-Z. Micromanagers of gene expression: The potentially widespread influence of metazoan microRNAs. *Nat. Rev. Genet.* **2004**, *5*, 396–400. [CrossRef] [PubMed]
67. Asgari, S. Role of microRNAs in insect host–microorganism interactions. *Front. Physiol.* **2011**, *2*, 48. [CrossRef] [PubMed]
68. Skalsky, R.L.; Cullen, B.R. Viruses, microRNAs, and host interactions. *Annu. Rev. Microbiol.* **2010**, *64*, 123–141. [CrossRef] [PubMed]
69. Ørom, U.A.; Nielsen, F.C.; Lund, A.H. MicroRNA-10a binds the 5′UTR of ribosomal protein mRNAs and enhances their translation. *Mol. Cell* **2008**, *30*, 460–471. [CrossRef] [PubMed]
70. Ma, F.; Liu, X.; Li, D.; Wang, P.; Li, N.; Lu, L.; Cao, X. MicroRNA-466l upregulates IL-10 expression in TLR-triggered macrophages by antagonizing RNA-binding protein tristetraprolin-mediated IL-10 mRNA degradation. *J. Immunol.* **2010**, *184*, 6053–6059. [CrossRef] [PubMed]
71. Hussain, M.; Frentiu, F.D.; Moreira, L.A.; O'Neill, S.L.; Asgari, S. Wolbachia uses host microRNAs to manipulate host gene expression and facilitate colonization of the dengue vector *Aedes aegypti*. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 9250–9255. [CrossRef] [PubMed]
72. Hussain, M.; Asgari, S. Functional analysis of a cellular microRNA in insect host-ascovirus interaction. *J. Virol.* **2010**, *84*, 612–620. [CrossRef] [PubMed]

73. Wang, F.Z.; Weber, F.; Croce, C.; Liu, C.G.; Liao, X.; Pellett, P.E. Human cytomegalovirus infection alters the expression of cellular microRNA species that affect its replication. *J. Virol.* **2008**, *82*, 9065–9074. [CrossRef] [PubMed]
74. Pedersen, I.M.; Cheng, G.; Wieland, S.; Volinia, S.; Croce, C.M.; Chisari, F.V.; David, M. Interferon modulation of cellular microRNAs as an antiviral mechanism. *Nature* **2007**, *449*, 919–922. [CrossRef] [PubMed]
75. Huang, T.; Xu, D.; Zhang, X. Characterization of host microRNAs that respond to DNA virus infection in a crustacean. *BMC Genet.* **2012**, *13*, 159. [CrossRef] [PubMed]
76. Huang, T.; Zhang, X. Functional analysis of a crustacean microRNA in host-virus interactions. *J. Virol.* **2012**, *86*, 12997–13004. [CrossRef] [PubMed]
77. He, Y.; Zhang, X. Comprehensive characterization of viral miRNAs involved in white spot syndrome virus (WSSV) infection. *RNA Biol.* **2012**, *9*, 1019–1029. [CrossRef] [PubMed]
78. Huang, T.; Cui, Y.; Zhang, X. Involvement of viral microRNA in the regulation of antiviral apoptosis in shrimp. *J. Virol.* **2014**, *88*, 2544–2554. [CrossRef] [PubMed]
79. He, Y.; Yang, K.; Zhang, X. Viral microRNAs targeting virus genes promote virus infection in shrimp *in vivo*. *J. Virol.* **2014**, *88*, 1104–1112. [CrossRef] [PubMed]
80. Escobedo-Bonilla, C.M.; Alday-Sanz, V.; Wille, M.; Sorgeloos, P.; Pensaert, M.B.; Nauwynck, H.J. A review on the morphology, molecular characterization, morphogenesis and pathogenesis of white spot syndrome virus. *J. Fish Dis.* **2008**, *31*, 1–18. [CrossRef] [PubMed]
81. Mercer, J.; Schelhaas, M.; Helenius, A. Virus entry by endocytosis. *Annu. Rev. Biochem.* **2010**, *79*, 803–833. [CrossRef] [PubMed]
82. Sodeik, B. Mechanisms of viral transport in the cytoplasm. *Trends Microbiol.* **2000**, *8*, 465–472. [CrossRef]
83. Marsh, M.; Bron, R. Sfv infection in CHO cells: Cell-type specific restrictions to productive virus entry at the cell surface. *J. Cell Sci.* **1997**, *110*, 95–103. [PubMed]
84. Kalia, M.; Jameel, S. Virus entry paradigms. *Amino Acids* **2011**, *41*, 1147–1157. [PubMed]
85. Van Hulten, M.C.W.; Witteveldt, J.; Snippe, M.; Vlak, J.M. White spot syndrome virus envelope protein VP28 is involved in the systemic infection of shrimp. *Virology* **2001**, *285*, 228–233. [CrossRef] [PubMed]
86. Li, L.J.; Yuan, J.F.; Cai, C.A.; Gu, W.G.; Shi, Z.L. Multiple envelope proteins are involved in white spot syndrome virus (WSSV) infection in crayfish. *Arch. Virol.* **2006**, *151*, 1309–1317. [CrossRef] [PubMed]
87. Yi, G.; Wang, Z.; Qi, Y.; Yao, L.; Qian, J.; Hu, L. Vp28 of shrimp white spot syndrome virus is involved in the attachment and penetration into shrimp cells. *J. Biochem. Mol. Biol.* **2004**, *37*, 726–734. [CrossRef] [PubMed]
88. Wan, Q.; Xu, L.; Yang, F. Vp26 of white spot syndrome virus functions as a linker protein between the envelope and nucleocapsid of virions by binding with VP51. *J. Virol.* **2008**, *82*, 12598–12601. [CrossRef] [PubMed]
89. Chen, L.L.; Lu, L.C.; Wu, W.J.; Lo, C.F.; Huang, W.P. White spot syndrome virus envelope protein VP53A interacts with *Penaeus monodon* chitin-binding protein (PmCBP). *Dis. Aquat. Org.* **2007**, *74*, 171–178. [CrossRef] [PubMed]
90. Chen, K.Y.; Hsu, T.C.; Huang, P.Y.; Kang, S.T.; Lo, C.F.; Huang, W.P.; Chen, L.L. *Penaeus monodon* chitinbinding protein (PMCBP) is involved in white spot syndrome virus (WSSV) infection. *Fish Shellfish Immunol.* **2009**, *27*, 460–465. [CrossRef] [PubMed]
91. Huang, H.-T.; Leu, J.-H.; Huang, P.-Y.; Chen, L.-L. A putative cell surface receptor for white spot syndrome virus is a member of a transporter superfamily. *PLoS ONE* **2012**, *7*, e33216. [CrossRef] [PubMed]
92. Watthanasurorot, A.; Jiravanichpaisal, P.; Soderhall, I.; Soderhall, K. A gC1qR prevents white spot syndrome virus replication in the freshwater crayfish *Pacifastacus leniusculus*. *J. Virol.* **2010**, *84*, 10844–10851. [CrossRef] [PubMed]
93. Zhao, Z.Y.; Yin, Z.X.; Xu, X.P.; Weng, S.P.; Rao, X.Y.; Dai, Z.X.; Luo, Y.; Yang, G.; Li, Z.H.; Guan, H.J.; *et al.* A novel C-type lectin from the shrimp *Litopenaeus vannamei* possesses anti-white spot syndrome virus activity. *J. Virol.* **2009**, *83*, 347–356. [CrossRef] [PubMed]
94. Wang, X.W.; Xu, W.T.; Zhang, X.W.; Zhao, X.F.; Yu, X.Q.; Wang, J.X. A c-type lectin is involved in the innate immune response of Chinese white shrimp. *Fish Shellfish Immunol.* **2009**, *27*, 556–562. [CrossRef] [PubMed]
95. Song, K.K.; Li, D.F.; Zhang, M.C.; Yang, H.J.; Ruan, L.W.; Xu, X. Cloning and characterization of three novel WSSV recognizing lectins from shrimp *Marsupenaeus japonicus*. *Fish Shellfish Immunol.* **2010**, *28*, 596–603. [CrossRef] [PubMed]

96. Wang, X.W.; Xu, Y.H.; Xu, J.D.; Zhao, X.F.; Wang, J.X. Collaboration between a soluble C-type lectin and calreticulin facilitates white spot syndrome virus infection in shrimp. *J. Immunol.* **2014**, *193*, 2106–2117. [CrossRef] [PubMed]
97. Xu, Y.H.; Bi, W.J.; Wang, X.W.; Zhao, Y.R.; Zhao, X.F.; Wang, J.X. Two novel C-type lectins with a low-density lipoprotein receptor class a domain have antiviral function in the shrimp *Marsupenaeus japonicus*. *Dev. Comp. Immunol.* **2014**, *42*, 323–332. [CrossRef] [PubMed]
98. Li, D.F.; Zhang, M.C.; Yang, H.J.; Zhu, Y.B.; Xu, X. B-integrin mediates WSSV infection. *Virology* **2007**, *368*, 122–132. [CrossRef] [PubMed]
99. Zhang, J.Y.; Liu, Q.H.; Huang, J. Multiple proteins of white spot syndrome virus involved in recognition of β-integrin. *J. Biosci.* **2014**, *39*, 1–8. [CrossRef]
100. Sun, Z.; Li, S.; Li, F.; Xiang, J. Bioinformatic prediction of WSSV-host protein-protein interaction. *BioMed Res. Int.* **2014**. [CrossRef] [PubMed]
101. Watthanasurorot, A.; Guo, E.; Tharntada, S.; Lo, C.F.; Soderhall, K.; Soderhall, I. Hijacking of host calreticulin is required for the white spot syndrome virus replication cycle. *J. Virol.* **2014**, *88*, 8116–8128. [CrossRef] [PubMed]
102. Wu, W.; Zong, R.; Xu, J.; Zhang, X. Antiviral phagocytosis is regulated by a novel Rab-dependent complex in shrimp *Penaeus japonicus*. *J. Proteom. Res.* **2007**, *7*, 424–431. [CrossRef] [PubMed]
103. Sritunyalucksana, K.; Wannapapho, W.; Lo, C.F.; Flegel, T.W. PMRAB7 is a VP28-binding protein involved in white spot syndrome virus infection in shrimp. *J. Virol.* **2006**, *80*, 10734–10742. [CrossRef] [PubMed]
104. Carrasco-Miranda, J.S.; Lopez-Zavala, A.A.; Arvizu-Flores, A.A.; Garcia-Orozco, K.D.; Stojanoff, V.; Rudiño-Piñera, E.; Brieba, L.G.; Sotelo-Mundo, R.R. Crystal structure of the shrimp proliferating cell nuclear antigen: Structural complementarity with WSSV DNA polymerase pip-box. *PLoS ONE* **2014**, *9*, e94369. [CrossRef] [PubMed]
105. Lin, S.J.; Lee, D.Y.; Wang, H.C.; Kang, S.T.; Hwang, P.P.; Kou, G.H.; Huang, M.F.; Chang, G.D.; Lo, C.F. White spot syndrome virus protein kinase 1 defeats the host cell's iron-withholding defense mechanism by interacting with host ferritin. *J. Virol.* **2015**, *89*, 1083–1093. [CrossRef] [PubMed]
106. Lu, H.; Ruan, L.; Xu, X. An immediate-early protein of white spot syndrome virus modulates the phosphorylation of focal adhesion kinase of shrimp. *Virology* **2011**, *419*, 84–89. [CrossRef] [PubMed]
107. Leu, J.H.; Chen, L.L.; Lin, Y.R.; Kou, G.H.; Lo, C.F. Molecular mechanism of the interactions between white spot syndrome virus anti-apoptosis protein AAP-1 (WSSV449) and shrimp effector caspase. *Dev. Comp. Immunol.* **2010**, *34*, 1068–1074. [CrossRef] [PubMed]
108. Lertwimol, T.; Sangsuriya, P.; Phiwsaiya, K.; Senapin, S.; Phongdara, A.; Boonchird, C.; Flegel, T.W. Two new anti-apoptotic proteins of white spot syndrome virus that bind to an effector caspase (PmCasp) of the giant tiger shrimp *Penaeus monodon*. *Fish Shellfish Immunol.* **2014**, *38*, 1–6. [CrossRef] [PubMed]
109. Wang, Z.; Chua, H.K.; Gusti, A.A.; He, F.; Fenner, B.; Manopo, I.; Wang, H.; Kwang, J. RING-H2 protein WSSV249 from white spot syndrome virus sequesters a shrimp ubiquitin-conjugating enzyme, PvUbc, for viral pathogenesis. *J. Virol.* **2005**, *79*, 8764–8772. [CrossRef] [PubMed]
110. Wang, H.C.; Wang, H.C.; Ko, T.P.; Lee, Y.M.; Leu, J.H.; Ho, C.H.; Huang, W.P.; Lo, C.F.; Wang, A.H. White spot syndrome virus protein ICP11: A histone-binding DNA mimic that disrupts nucleosome assembly. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 20758–20763. [CrossRef] [PubMed]
111. Tonganunt, M.; Saelee, N.; Chotigeat, W.; Phongdara, A. Identification of a receptor for activated protein kinase C1 (Pm-RACK1), a cellular gene product from black tiger shrimp (*Penaeus monodon*) interacts with a protein, VP9 from the white spot syndrome virus. *Fish Shellfish Immunol.* **2009**, *26*, 509–514. [CrossRef] [PubMed]
112. Sangsuriya, P.; Senapin, S.; Huang, W.-P.; Lo, C.-F.; Flegel, T.W. Co-interactive DNA-binding between a novel, immunophilin-like shrimp protein and VP15 nucleocapsid protein of white spot syndrome virus. *PLoS ONE* **2011**, *6*, e25420. [CrossRef] [PubMed]
113. Ma, F.F.; Liu, Q.H.; Guan, G.K.; Li, C.; Huang, J. Arginine kinase of *Litopenaeus vannamei* involved in white spot syndrome virus infection. *Gene* **2014**, *539*, 99–106. [CrossRef] [PubMed]
114. Xie, X.; Yang, F. Interaction of white spot syndrome virus VP26 protein with actin. *Virology* **2005**, *336*, 93–99. [CrossRef] [PubMed]
115. Lu, L.; Kwang, J. Identification of a novel shrimp protein phosphatase and its association with latency-related ORF427 of white spot syndrome virus. *FEBS Lett.* **2004**, *577*, 141–146. [CrossRef] [PubMed]

116. Liu, W.J.; Chang, Y.S.; Huang, W.T.; Chen, I.T.; Wang, K.C.; Kou, G.H.; Lo, C.F. *Penaeus monodon* tata box-binding protein interacts with the white spot syndrome virus transactivator *ie1* and promotes its transcriptional activity. *J. Virol.* **2011**, *85*, 6535–6547. [CrossRef] [PubMed]
117. Ran, X.; Bian, X.; Ji, Y.; Yan, X.; Yang, F.; Li, F. White spot syndrome virus *ie1* and *wsv056* modulate the G1/S transition by binding to the host retinoblastoma protein. *J. Virol.* **2013**, *87*, 12576–12582. [CrossRef] [PubMed]
118. Ruoslahti, E. RGD and other recognition sequences for integrins. *Annu. Rev. Cell Dev. Biol.* **1996**, *12*, 697–715. [CrossRef] [PubMed]
119. Chang, Y.-S.; Liu, W.-J.; Lee, C.-C.; Chou, T.-L.; Lee, Y.-T.; Wu, T.-S.; Huang, J.-Y.; Huang, W.-T.; Lee, T.-L.; Kou, G.-H.; *et al.* A 3D model of the membrane protein complex formed by the white spot syndrome virus structural proteins. *PLoS ONE* **2010**, *5*, e10718. [CrossRef] [PubMed]
120. Spear, P.G.; Longnecker, R. Herpesvirus entry: An update. *J. Virol.* **2003**, *77*, 10179–10185. [CrossRef] [PubMed]
121. Mason, C.P.; Tarr, A.W. Human lectins and their roles in viral infections. *Molecules* **2015**, *20*, 2229–2271. [CrossRef] [PubMed]
122. Huang, Z.-J.; Kang, S.-T.; Leu, J.-H.; Chen, L.-L. Endocytic pathway is indicated for white spot syndrome virus (WSSV) entry in shrimp. *Fish Shellfish Immunol.* **2013**, *35*, 707–715. [CrossRef] [PubMed]
123. Duan, H.; Jin, S.; Zhang, Y.; Li, F.; Xiang, J. Granulocytes of the red claw crayfish *Cherax quadricarinatus* can endocytose beads, *E. coli* and wssv, but in different ways. *Dev. Comp. Immunol.* **2014**, *46*, 186–193. [CrossRef] [PubMed]
124. Huang, J.; Li, F.; Wu, J.; Yang, F. White spot syndrome virus enters crayfish hematopoietic tissue cells via Clathrin-mediated endocytosis. *Virology* **2015**, *486*, 35–43. [CrossRef] [PubMed]
125. Mayor, S.; Presley, J.F.; Maxfield, F.R. Sorting of membrane components from endosomes and subsequent recycling to the cell surface occurs by a bulk flow process. *J. Cell Biol.* **1993**, *121*, 1257–1269. [CrossRef] [PubMed]
126. Lozach, P.-Y.; Huotari, J.; Helenius, A. Late-penetrating viruses. *Curr. Opin. Virol.* **2011**, *1*, 35–43. [CrossRef] [PubMed]
127. Zhang, M.; Chen, L.; Wang, S.; Wang, T. Rab7: Roles in membrane trafficking and disease. *Biosci. Rep.* **2009**, *29*, 193–209. [CrossRef] [PubMed]
128. Kobiler, O.; Drayman, N.; Butin-Israeli, V.; Oppenheim, A. Virus strategies for passing the nuclear envelope barrier. *Nucleus* **2012**, *3*, 526–539. [CrossRef] [PubMed]
129. Huotari, J.; Helenius, A. Endosome maturation. *EMBO J.* **2011**, *30*, 3481–3500. [CrossRef] [PubMed]
130. Del Conte-Zerial, P.; Brusch, L.; Rink, J.C.; Collinet, C.; Kalaidzidis, Y.; Zerial, M.; Deutsch, A. Membrane identity and GTPase cascades regulated by toggle and cut-out switches. *Mol. Syst. Biol.* **2008**, *4*, 206. [CrossRef] [PubMed]
131. Píndaro, Á.-R.; Humberto, M.-R.C.; Javier, M.-B.F.; Marcial, E.-B.C. Silencing pacific white shrimp *Litopenaeus vannamei* LVRAB7 reduces mortality in brooders challenged with white spot syndrome virus. *Aquacult. Res.* **2013**, *44*, 772–782. [CrossRef]
132. Attasart, P.; Kaewkhaw, R.; Chimwai, C.; Kongphom, U.; Namramoon, O.; Panyim, S. Inhibition of white spot syndrome virus replication in *Penaeus monodon* by combined silencing of viral rr2 and shrimp PmRab7. *Virus Res.* **2009**, *145*, 127–133. [CrossRef] [PubMed]
133. Ongvarrasopone, C.; Chanasakulniyom, M.; Sritunyalucksana, K.; Panyim, S. Suppression of PmRab7 by dsRNA inhibits WSSV or YHV infection in shrimp. *Ma. Biotechnol.* **2008**, *10*, 374–381. [CrossRef] [PubMed]
134. Pan, D.; He, N.; Yang, Z.; Liu, H.; Xu, X. Differential gene expression profile in hepatopancreas of WSSV-resistant shrimp (*Penaeus japonicus*) by suppression subtractive hybridization. *Dev. Comp. Immunol.* **2005**, *29*, 103–112. [CrossRef] [PubMed]
135. Grigoriev, I.; Splinter, D.; Keijzer, N.; Wulf, P.S.; Demmers, J.; Ohtsuka, T.; Modesti, M.; Maly, I.V.; Grosveld, F.; Hoogenraad, C.C.; *et al.* Rab6 regulates transport and targeting of exocytotic carriers. *Dev. Cell* **2007**, *13*, 305–314. [CrossRef] [PubMed]
136. Liu, W.J.; Chang, Y.S.; Wang, A.H.; Kou, G.H.; Lo, C.F. White spot syndrome virus annexes a shrimp stat to enhance expression of the immediate-early gene *ie1*. *J. Virol.* **2007**, *81*, 1461–1471. [CrossRef] [PubMed]
137. Ma, G.; Yu, L.; Wang, Q.; Liu, W.; Cui, Y.; Kwang, J. Sf-PHB2, a new transcription factor, drives WSSV *ie1* gene expression via a 12-bp DNA element. *Virol. J.* **2012**, *9*, 1–11. [CrossRef] [PubMed]

138. Qiu, W.; Zhang, S.; Chen, Y.G.; Wang, P.H.; Xu, X.P.; Li, C.Z.; Chen, Y.H.; Fan, W.Z.; Yan, H.; Weng, S.P.; *et al*. *Litopenaeus vannamei* NF-κB is required for WSSV replication. *Dev. Comp. Immunol.* **2014**, *45*, 156–162. [CrossRef] [PubMed]
139. Huang, X.D.; Zhao, L.; Zhang, H.Q.; Xu, X.P.; Jia, X.T.; Chen, Y.H.; Wang, P.H.; Weng, S.P.; Yu, X.Q.; Yin, Z.X.; *et al*. Shrimp NF-κB binds to the immediate-early gene *ie1* promoter of white spot syndrome virus and upregulates its activity. *Virology* **2010**, *406*, 176–180. [CrossRef] [PubMed]
140. Wang, P.H.; Gu, Z.H.; Wan, D.H.; Zhang, M.Y.; Weng, S.P.; Yu, X.Q.; He, J.G. The shrimp NF-κB pathway is activated by white spot syndrome virus (WSSV) 449 to facilitate the expression of WSSV069 (*ie1*), WSSV303 and WSSV371. *PLoS ONE* **2011**, *6*, e24773. [CrossRef] [PubMed]
141. Shi, H.; Yan, X.; Ruan, L.; Xu, X. A novel jnk from *Litopenaeus vannamei* involved in white spot syndrome virus infection. *Dev. Comp. Immunol.* **2012**, *37*, 421–428. [CrossRef] [PubMed]
142. Li, X.Y.; Pang, L.R.; Chen, Y.G.; Weng, S.P.; Yue, H.T.; Zhang, Z.Z.; Chen, Y.H.; He, J.G. Activating transcription factor 4 and x box binding protein 1 of *Litopenaeus vannamei* transcriptional regulated white spot syndrome virus genes *wsv023* and *wsv083*. *PLoS ONE* **2013**, *8*, e62603. [CrossRef] [PubMed]
143. Liu, W.J.; Lo, C.F.; Kou, G.H.; Leu, J.H.; Lai, Y.J.; Chang, L.K.; Chang, Y.S. The promoter of the white spot syndrome virus immediate-early gene *wssv108* is activated by the cellular KLF transcription factor. *Dev. Comp. Immunol.* **2015**, *49*, 7–18. [CrossRef] [PubMed]
144. Li, X.Y.; Yue, H.T.; Zhang, Z.Z.; Bi, H.T.; Chen, Y.G.; Weng, S.P.; Chan, S.; He, J.G.; Chen, Y.H. An activating transcription factor of *Litopenaeus vannamei* involved in *wssv* genes *wsv059* and *wsv166* regulation. *Fish Shellfish Immunol.* **2014**, *41*, 147–155. [CrossRef] [PubMed]
145. Zuo, H.; Chen, C.; Gao, Y.; Lin, J.; Jin, C.; Wang, W. Regulation of shrimp PjCaspase promoter activity by WSSV VP38 and VP41B. *Fish Shellfish Immunol.* **2011**, *30*, 1188–1191. [CrossRef] [PubMed]
146. Li, F.; Li, M.; Ke, W.; Ji, Y.; Bian, X.; Yan, X. Identification of the immediate-early genes of white spot syndrome virus. *Virology* **2009**, *385*, 267–274. [CrossRef] [PubMed]
147. Liu, W.J.; Chang, Y.S.; Wang, C.H.; Kou, G.H.; Lo, C.F. Microarray and RT-PCR screening for white spot syndrome virus immediate-early genes in cycloheximide-treated shrimp. *Virology* **2005**, *334*, 327–341. [CrossRef] [PubMed]
148. He, F.; Ho, Y.; Yu, L.; Kwang, J. WSSV *ie1* promoter is more efficient than CMV promoter to express H5 hemagglutinin from influenza virus in baculovirus as a chicken vaccine. *BMC Microbiol.* **2008**, *8*. [CrossRef] [PubMed]
149. Gao, H.; Wang, Y.; Li, N.; Peng, W.-P.; Sun, Y.; Tong, G.-Z.; Qiu, H.-J. Efficient gene delivery into mammalian cells mediated by a recombinant baculovirus containing a whispovirus *ie1* promoter, a novel shuttle promoter between insect cells and mammalian cells. *J. Biotechnol.* **2007**, *131*, 138–143. [CrossRef] [PubMed]
150. Yan, M.; Li, C.; Su, Z.; Liang, Q.; Li, H.; Liang, S.; Weng, S.; He, J.; Xu, X. Identification of a jak/stat pathway receptor domeless from pacific white shrimp *Litopenaeus vannamei*. *Fish Shellfish Immunol.* **2015**, *44*, 26–32. [CrossRef] [PubMed]
151. Vallabhapurapu, S.; Karin, M. Regulation and function of NF-κB transcription factors in the immune system. *Annu. Rev. Immunol.* **2009**, *27*, 693–733. [CrossRef] [PubMed]
152. Matys, V.; Kel-Margoulis, O.V.; Fricke, E.; Liebich, I.; Land, S.; Barre-Dirrie, A.; Reuter, I.; Chekmenev, D.; Krull, M.; Hornischer, K.; *et al*. Transfac® and its module transcompel®: Transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **2006**, *34*, D108–D110. [CrossRef] [PubMed]
153. Chen, A.J.; Gao, L.; Wang, X.W.; Zhao, X.F.; Wang, J.X. SUMO-conjugating enzyme E2 UBC9 mediates viral immediate-early protein sumoylation in crayfish to facilitate reproduction of white spot syndrome virus. *J. Virol.* **2013**, *87*, 636–647. [CrossRef] [PubMed]
154. Flemington, E.K. Herpesvirus lytic replication and the cell cycle: Arresting new developments. *J. Virol.* **2001**, *75*, 4475–4481. [CrossRef] [PubMed]
155. Kudoh, A.; Fujita, M.; Kiyono, T.; Kuzushima, K.; Sugaya, Y.; Izuta, S.; Nishiyama, Y.; Tsurumi, T. Reactivation of lytic replication from b cells latently infected with Epstein-Barr virus occurs with high s-phase cyclin-dependent kinase activity while inhibiting cellular DNA replication. *J. Virol.* **2003**, *77*, 851–861. [CrossRef] [PubMed]
156. Chen, L.-L.; Wang, H.-C.; Huang, C.-J.; Peng, S.-E.; Chen, Y.-G.; Lin, S.-J.; Chen, W.-Y.; Dai, C.-F.; Yu, H.-T.; Wang, C.-H.; *et al*. Transcriptional analysis of the DNA polymerase gene of shrimp white spot syndrome virus. *Virology* **2002**, *301*, 136–147. [CrossRef] [PubMed]

157. De-la-Re-Vega, E.; Garcia-Orozco, K.D.; Arvizu-Flores, A.A.; Yepiz-Plascencia, G.; Muhlia-Almazan, A.; Hernández, J.; Brieba, L.G.; Sotelo-Mundo, R.R. White spot syndrome virus ORF514 encodes a bona fide DNA polymerase. *Molecules* **2011**, *16*, 532–542. [CrossRef] [PubMed]
158. De-la-Re-Vega, E.; Muhlia-Almazan, A.; Arvizu-Flores, A.A.; Islas-Osuna, M.A.; Yepiz-Plascencia, G.; Brieba, L.G.; Sotelo-Mundo, R.R. Molecular modeling and expression of the Litopenaeus vannamei proliferating cell nuclear antigen (PCNA) after white spot syndrome virus shrimp infection. *Res. Immunol.* **2011**, *1*, 24–30. [CrossRef] [PubMed]
159. Lunt, S.Y.; van der Heiden, M.G. Aerobic glycolysis: Meeting the metabolic requirements of cell proliferation. *Annu. Rev. Cell Dev. Biol.* **2011**, *27*, 441–464. [CrossRef] [PubMed]
160. Warburg, O. On the origin of cancer cells. *Science* **1956**, *123*, 309–314. [CrossRef] [PubMed]
161. Su, M.A.; Huang, Y.T.; Chen, I.T.; Lee, D.Y.; Hsieh, Y.C.; Li, C.Y.; Ng, T.H.; Liang, S.Y.; Lin, S.Y.; Huang, S.W.; *et al*. An invertebrate Warburg effect: A shrimp virus achieves successful replication by altering the host metabolome via the pi3k-akt-mtor pathway. *PLoS Pathog.* **2014**, *10*, e1004196. [CrossRef] [PubMed]
162. Chen, I.T.; Aoki, T.; Huang, Y.T.; Hirono, I.; Chen, T.C.; Huang, J.Y.; Chang, G.D.; Lo, C.F.; Wang, H.C. White spot syndrome virus induces metabolic changes resembling the Warburg effect in shrimp hemocytes in the early stage of infection. *J. Virol.* **2011**, *85*, 12919–12928. [CrossRef] [PubMed]
163. Robey, R.B.; Hay, N. Is Akt the "warburg kinase"?—Akt-energy metabolism interactions and oncogenesis. *Semin. Cancer Biol.* **2009**, *19*, 25–31. [CrossRef] [PubMed]
164. Spangle, J.M.; Munger, K. The human papillomavirus type 16 E6 oncoprotein activates mtorc1 signaling and increases protein synthesis. *J. Virol.* **2010**, *84*, 9398–9407. [CrossRef]
165. Ye, T.; Wu, X.; Wu, W.; Dai, C.; Yuan, J. Ferritin protect shrimp *Litopenaeus vannamei* from WSSV infection by inhibiting virus replication. *Fish Shellfish Immunol.* **2015**, *42*, 138–143. [CrossRef] [PubMed]
166. Kim, C.S.; Kosuke, Z.; Nam, Y.K.; Kim, S.K.; Kim, K.H. Protection of shrimp (*Penaeus chinensis*) against white spot syndrome virus (WSSV) challenge by double-stranded RNA. *Fish Shellfish Immunol.* **2007**, *23*, 242–246. [CrossRef]
167. Fung, T.S.; Liu, D.X. Coronavirus infection, ER stress, apoptosis and innate immunity. *Front. Microbiol.* **2014**, *5*. [CrossRef] [PubMed]
168. Smith, J.A. A new paradigm: Innate immune sensing of viruses via the unfolded protein response. *Front. Microbiol.* **2014**, *5*, 222. [CrossRef] [PubMed]
169. He, B. Viruses, endoplasmic reticulum stress, and interferon responses. *Cell Death Differ.* **2006**, *13*, 393–403. [CrossRef] [PubMed]
170. Taylor, G.M.; Raghuwanshi, S.K.; Rowe, D.T.; Wadowsky, R.M.; Rosendorff, A. Endoplasmic reticulum stress causes EBV lytic replication. *Blood* **2011**, *118*, 5528–5539. [CrossRef] [PubMed]
171. Trujillo-Alonso, V.; Maruri-Avidal, L.; Arias, C.F.; López, S. Rotavirus infection induces the unfolded protein response of the cell and controls it through the nonstructural protein NSP3. *J. Virol.* **2011**, *85*, 12594–12604. [CrossRef] [PubMed]
172. Burnett, H.F.; Audas, T.E.; Liang, G.; Lu, R.R. Herpes simplex virus-1 disarms the unfolded protein response in the early stages of infection. *Cell Stress Chaperones* **2012**, *17*, 473–483. [CrossRef] [PubMed]
173. Chen, Y.H.; Zhao, L.; Pang, L.R.; Li, X.Y.; Weng, S.P.; He, J.G. Identification and characterization of inositol-requiring enzyme-1 and x-box binding protein 1, two proteins involved in the unfolded protein response of *Litopenaeus vannamei*. *Dev. Comp. Immunol.* **2012**, *38*, 66–77. [CrossRef] [PubMed]
174. Luana, W.; Li, F.; Wang, B.; Zhang, X.; Liu, Y.; Xiang, J. Molecular characteristics and expression analysis of calreticulin in Chinese shrimp *Fenneropenaeus chinensis*. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **2007**, *147*, 482–491. [CrossRef] [PubMed]
175. Visudtiphole, V.; Watthanasurorot, A.; Klinbunga, S.; Menasveta, P.; Kirtikara, K. Molecular characterization of calreticulin: A biomarker for temperature stress responses of the giant tiger shrimp *Penaeus monodon*. *Aquaculture* **2010**, *308*, S100–S108. [CrossRef]
176. Luan, W.; Li, F.; Zhang, J.; Wang, B.; Xiang, J. Cloning and expression of glucose regulated protein 78 (GRP78) in *Fenneropenaeus chinensis*. *Mol. Biol. Rep.* **2009**, *36*, 289–298. [CrossRef] [PubMed]
177. Huang, W.-J.; Leu, J.-H.; Tsau, M.-T.; Chen, J.-C.; Chen, L.-L. Differential expression of LVHSP60 in shrimp in response to environmental stress. *Fish Shellfish Immunol.* **2011**, *30*, 576–582. [CrossRef] [PubMed]

178. Lin, Y.-R.; Hung, H.-C.; Leu, J.-H.; Wang, H.-C.; Kou, G.-H.; Lo, C.-F. The role of aldehyde dehydrogenase and HSP70 in suppression of white spot syndrome virus replication at high temperature. *J. Virol.* **2011**, *85*, 3517–3525. [CrossRef] [PubMed]
179. Donnelly, N.; Gorman, A.; Gupta, S.; Samali, A. The EIF2α kinases: Their structures and functions. *Cell. Mol. Life Sci.* **2013**, *70*, 3493–3511. [CrossRef] [PubMed]
180. Xu, J.; Ruan, L.; Shi, H. EIF2α of *Litopenaeus vannamei* involved in shrimp immune response to WSSV infection. *Fish Shellfish Immunol.* **2014**, *40*, 609–615. [CrossRef] [PubMed]
181. Koyama, A.H.; Fukumori, T.; Fujita, M.; Irie, H.; Adachi, A. Physiological significance of apoptosis in animal virus infection. *Microbes Infect.* **2000**, *2*, 1111–1117. [CrossRef]
182. Best, S.M. Viral subversion of apoptotic enzymes: Escape from death row. *Annu. Rev. Microbiol.* **2008**, *62*, 171–192. [CrossRef] [PubMed]
183. Henderson, T.; Stuck, K. Induction of apoptosis in response to white spot syndrome virus in the pacific white shrimp. In *Penaeus Vannamei*; Book of abstracts; Aquaculture America: Tampa, FL, USA, 1999; Abstract 67.
184. Abeer, H.S.; Hassan, M.D.; Shariff, M. DNA fragmentation, an indicator of apoptosis, in cultured black tiger shrimp *Penaeus monodon* infected with white spot syndrome virus (WSSV). *Dis. Aquat. Org.* **2001**, *44*, 155–159.
185. Kanokpan, W.; Kornnika, K.; Supatra Somapa, G.; Prasert, M.; Boonsirm, W. Time-course and levels of apoptosis in various tissues of black tiger shrimp *Penaeus monodon* infected with white-spot syndrome virus. *Dis. Aquat. Org.* **2003**, *55*, 3–10.
186. Wu, J.L.; Muroga, K. Apoptosis does not play an important role in the resistance of "immune" *Penaeus japonicus* against white spot syndrome virus. *J. Fish Dis.* **2004**, *27*, 15–21. [CrossRef] [PubMed]
187. Leu, J.H.; Lin, S.J.; Huang, J.Y.; Chen, T.C.; Lo, C.F. A model for apoptotic interaction between white spot syndrome virus and shrimp. *Fish Shellfish Immunol.* **2013**, *34*, 1011–1017. [CrossRef] [PubMed]
188. Shi, Y. Mechanisms of caspase activation and inhibition during apoptosis. *Mol. Cell* **2002**, *9*, 459–470. [CrossRef]
189. McIlwain, D.R.; Berger, T.; Mak, T.W. Caspase functions in cell death and disease. *Cold Spring Harbor Perspect. Biol.* **2013**, *5*. [CrossRef] [PubMed]
190. Wang, L.; Zhi, B.; Wu, W.; Zhang, X. Requirement for shrimp caspase in apoptosis against virus infection. *Dev. Comp. Immunol.* **2008**, *32*, 706–715. [CrossRef] [PubMed]
191. Xie, X.; Xu, L.; Yang, F. Proteomic analysis of the major envelope and nucleocapsid proteins of white spot syndrome virus. *J. Virol.* **2006**, *80*, 10615–10623. [CrossRef] [PubMed]
192. Jie, Z.; Xu, L.; Yang, F. The c-terminal region of envelope protein VP38 from white spot syndrome virus is indispensable for interaction with VP24. *Arch. Virol.* **2008**, *153*, 2103–2106. [CrossRef] [PubMed]
193. Jesenberger, V.; Jentsch, S. Deadly encounter: Ubiquitin meets apoptosis. *Nat. Rev. Mol. Cell Biol.* **2002**, *3*, 112–121. [CrossRef] [PubMed]
194. Nandi, D.; Tahiliani, P.; Kumar, A.; Chandu, D. The ubiquitin-proteasome system. *J. Biosci.* **2006**, *31*, 137–155. [CrossRef] [PubMed]
195. Freemont, P.S. Ubiquitination: Ring for destruction? *Current Biol.* **2000**, *10*, R84–R87. [CrossRef]
196. Chen, A.J.; Wang, S.; Zhao, X.F.; Yu, X.Q.; Wang, J.X. Enzyme E2 from Chinese white shrimp inhibits replication of white spot syndrome virus and ubiquitinates its ring domain proteins. *J. Virol.* **2011**, *85*, 8069–8079. [CrossRef] [PubMed]
197. He, F.; Fenner, B.J.; Godwin, A.K.; Kwang, J. White spot syndrome virus open reading frame 222 encodes a viral E3 ligase and mediates degradation of a host tumor suppressor via ubiquitination. *J. Virol.* **2006**, *80*, 3884–3892. [CrossRef] [PubMed]
198. Wang, H.C.; Wang, H.C.; Kou, G.C.; Lo, C.F.; Huang, W.P. Identification of ICP11, the most highly expressed gene of shrimp white spot syndrome virus (WSSV). *Dis. Aquat. Org.* **2007**, *74*, 179–189. [CrossRef] [PubMed]
199. Liu, Y.; Wu, J.; Song, J.; Sivaraman, J.; Hew, C.L. Identification of a novel nonstructural protein, VP9, from white spot syndrome virus: Its structure reveals a ferredoxin fold with specific metal binding sites. *J. Virol.* **2006**, *80*, 10419–10427. [CrossRef] [PubMed]
200. Boner, W.; Morgan, I.M. Novel cellular interacting partners of the human papillomavirus 16 transcription/replication factor E2. *Virus Res.* **2002**, *90*, 113–118. [CrossRef]
201. Liu, Y.; Wu, J.; Chen, H.; Hew, C.L.; Yan, J. DNA condensates organized by the capsid protein VP15 in white spot syndrome virus. *Virology* **2010**, *408*, 197–203. [CrossRef] [PubMed]

202. Witteveldt, J.; Vermeesch, A.M.G.; Langenhof, M.; de Lang, A.; Vlak, J.M.; van Hulten, M.C.W. Nucleocapsid protein VP15 is the basic DNA binding protein of white spot syndrome virus of shrimp. *Arch. Virol.* **2005**, *150*, 1121–1133. [CrossRef] [PubMed]
203. Yang, W.M.; Yao, Y.L.; Seto, E. The FK506-binding protein 25 functionally associates with histone deacetylases and with transcription factor YY1. *EMBO J.* **2001**, *20*, 4814–4825. [CrossRef] [PubMed]
204. Kuzuhara, T.; Horikoshi, M. A nuclear FK506-binding protein is a histone chaperone regulating rDNA silencing. *Nat. Struct. Mol. Biol.* **2004**, *11*, 275–283. [CrossRef] [PubMed]
205. Kobayashi, S.; Uchiyama, S.; Sone, T.; Noda, M.; Lin, L.; Mizuno, H.; Matsunaga, S.; Fukui, K. Calreticulin as a new histone binding protein in mitotic chromosomes. *Cytogenet. Genome Res.* **2006**, *115*, 10–15. [CrossRef] [PubMed]
206. Holaska, J.M.; Black, B.E.; Love, D.C.; Hanover, J.A.; Leszyk, J.; Paschal, B.M. Calreticulin is a receptor for nuclear export. *J. Cell Biol.* **2001**, *152*, 127–140. [CrossRef] [PubMed]
207. Hsieh, Y.C.; Chen, Y.M.; Li, C.Y.; Chang, Y.H.; Liang, S.Y.; Lin, S.Y.; Lin, C.Y.; Chang, S.H.; Wang, Y.J.; Khoo, K.H.; *et al*. To complete its replication cycle, a shrimp virus changes the population of long chain fatty acids during infection via the PI3K-Akt-mTOR-HIF1α pathway. *Dev. Comp. Immunol.* **2015**, *53*, 85–95. [CrossRef] [PubMed]
208. Itami, T.; Asano, M.; Tokushige, K.; Kubono, K.; Nakagawa, A.; Takeno, N.; Nishimura, H.; Maeda, M.; Kondo, M.; Takahashi, Y. Enhancement of disease resistance of kuruma shrimp, *Penaeus japonicus*, after oral administration of peptidoglycan derived from *Bifidobacterium thermophilum*. *Aquaculture* **1998**, *164*, 277–288. [CrossRef]
209. Thitamadee, S.; Srisala, J.; Taengchaiyaphum, S.; Sritunyalucksana, K. Double-dose β-glucan treatment in WSSV-challenged shrimp reduces viral replication but causes mortality possibly due to excessive Ros production. *Fish Shellfish Immunol.* **2014**, *40*, 478–484. [CrossRef] [PubMed]
210. Syed Musthaq, S.K.; Kwang, J. Reprint of "evolution of specific immunity in shrimp—A vaccination perspective against white spot syndrome virus". *Dev. Comp. Immunol.* **2015**, *48*, 342–353. [CrossRef] [PubMed]
211. Singh, I.S.B.; Manjusha, M.; Pai, S.S.; Rosamma, P. *Fenneropenaeus indicus* is protected from white spot disease by oral administration of inactivated white spot syndrome virus. *Dis. Aquat. Org.* **2005**, *66*, 265–270. [CrossRef] [PubMed]
212. Syed Musthaq, S.; Madhan, S.; Sahul Hameed, A.S.; Kwang, J. Localization of VP28 on the baculovirus envelope and its immunogenicity against white spot syndrome virus in *Penaeus monodon*. *Virology* **2009**, *391*, 315–324. [CrossRef] [PubMed]
213. Nguyen, A.T.; Pham, C.K.; Pham, H.T.; Pham, H.L.; Nguyen, A.H.; Dang, L.T.; Huynh, H.A.; Cutting, S.M.; Phan, T.N. *Bacillus subtilis* spores expressing the VP28 antigen: A potential oral treatment to protect *Litopenaeus vannamei* against white spot syndrome. *FEMS Microbiol. Lett.* **2014**, *358*, 202–208. [CrossRef] [PubMed]
214. Huang, T.; Zhang, X. Host defense against DNA virus infection in shrimp is mediated by the siRNA pathway. *European J. Immunol.* **2013**, *43*, 137–146. [CrossRef] [PubMed]
215. Zhu, F.; Zhang, X. Protection of shrimp against white spot syndrome virus (WSSV) with beta-1,3-D-glucan-encapsulated VP28-sirna particles. *Ma. Biotechnol.* **2012**, *14*, 63–68. [CrossRef] [PubMed]
216. Balasubramanian, G.; Sarathi, M.; Venkatesan, C.; Thomas, J.; Sahul Hameed, A.S. Oral administration of antiviral plant extract of *Cynodon dactylon* on a large scale production against white spot syndrome virus (WSSV) in *Penaeus monodon*. *Aquaculture* **2008**, *279*, 2–5. [CrossRef]
217. Sudheer, N.S.; Philip, R.; Singh, I.S.B. *In vivo* screening of mangrove plants for anti WSSV activity in *Penaeus monodon*, and evaluation of ceriops tagal as a potential source of antiviral molecules. *Aquaculture* **2011**, *311*, 36–41. [CrossRef]
218. Balasubramanian, G.; Sudhakaran, R.; Syed Musthaq, S.; Sarathi, M.; Sahul Hameed, A.S. Studies on the inactivation of white spot syndrome virus of shrimp by physical and chemical treatments, and seaweed extracts tested in marine and freshwater animal models. *J. Fish Dis.* **2006**, *29*, 569–572. [CrossRef] [PubMed]

219. Ghosh, U.; Chakraborty, S.; Balasubramanian, T.; Das, P. Screening, isolation and optimization of anti–white spot syndrome virus drug derived from terrestrial plants. *Asia-Pac. J. Trop. Biomed.* **2014**, *4*, S118–S128. [CrossRef] [PubMed]

220. Bindhu, F.; Velmurugan, S.; Donio, M.B.S.; Michaelbabu, M.; Citarasu, T. Influence of *agathi grandiflora* active principles inhibit viral multiplication and stimulate immune system in Indian white shrimp *Fenneropenaeus indicus* against white spot syndrome virus infection. *Fish Shellfish Immunol.* **2014**, *41*, 482–492. [CrossRef] [PubMed]

Chapter 3

De novo assembly of the *Carcinus maenas* transcriptome and characterization of innate immune system pathways.

BMC Genomics (2016), *16*(1), 458.

Pages: 83 - 100

RESEARCH ARTICLE Open Access

De novo assembly of the *Carcinus maenas* transcriptome and characterization of innate immune system pathways

Bas Verbruggen[1*†], Lisa K. Bickley[1†], Eduarda M. Santos[1], Charles R. Tyler[1], Grant D. Stentiford[2], Kelly S. Bateman[2] and Ronny van Aerle[3*]

Abstract

Background: The European shore crab, *Carcinus maenas*, is used widely in biomonitoring, ecotoxicology and for studies into host-pathogen interactions. It is also an important invasive species in numerous global locations. However, the genomic resources for this organism are still sparse, limiting research progress in these fields. To address this resource shortfall we produced a *C. maenas* transcriptome, enabled by the progress in next-generation sequencing technologies, and applied this to assemble information on the innate immune system in this species.

Results: We isolated and pooled RNA for twelve different tissues and organs from *C. maenas* individuals and sequenced the RNA using next generation sequencing on an Illumina HiSeq 2500 platform. After *de novo* assembly a transcriptome was generated encompassing 212,427 transcripts (153,699 loci). The transcripts were filtered, annotated and characterised using a variety of tools (including BLAST, MEGAN and RSEM) and databases (including NCBI, Gene Ontology and KEGG). There were differential patterns of expression for between 1,223 and 2,741 transcripts across tissues and organs with over-represented Gene Ontology terms relating to their specific function. Based on sequence homology to immune system components in other organisms, we show both the presence of transcripts for a series of known pathogen recognition receptors and response proteins that form part of the innate immune system, and transcripts representing the RNAi, Toll-like receptor signalling, IMD and JAK/STAT pathways.

Conclusions: We have produced an assembled transcriptome for *C. maenas* that provides a significant molecular resource for wide ranging studies in this species. Analysis of the transcriptome has revealed the presence of a series of known targets and functional pathways that form part of their innate immune system and illustrate tissue specific differences in their expression patterns.

Background

In recent years, large scale sequencing studies have benefitted from the advance of high-throughput sequencing technologies that have resulted in substantial improvement in sequencing efficiency. Additionally, increases in the length and quality of sequencing reads have improved assemblies of sequenced genomes and transcriptomes. Sequencing is a powerful technique allowing for the rapid generation of transcriptome assemblies for any species of interest. Transcriptome sequencing measures expressed sequences only, thus does not have some of the challenges in DNA sequencing (e.g. long repeating sequences) [1]. *De novo* transcriptome assembly removes the need for a reference genome in quantitative RNA-Seq experiments, allowing for the rapid and accurate quantification of transcript abundance in a given biological sample. These aspects are especially useful in studies for organisms with limited genomic resources. Exemplary is the application of *de novo* transcriptome sequencing to a large range of organisms: vertebrates, e.g. brown trout (*Salmo trutta*) [2], invertebrates e.g. sea louse (*Caligus rogercresseyi*) [3], oriental fruit flies (*Bactrocera dorsalis*) [4] and the pollen beetles

* Correspondence: bv213@exeter.ac.uk; ronny.vanaerle@cefas.co.uk
†Equal contributors
[1]Biosciences, College of Life & Environmental Sciences, University of Exeter, Geoffrey Pope Building, Exeter EX4 4QD, UK
[3]Aquatic Health and Hygiene Division, Centre for Environment, Fisheries and Aquaculture Science (Cefas), Weymouth Laboratory, Weymouth, Dorset DT4 8UB, UK
Full list of author information is available at the end of the article

(*Meligethes aeneus*) [5], fungi (*Trichoderma brevicompactum*) [6] and other microorganisms.

Despite the rapid advances in sequence capabilities and in bioinformatics resources for generating high quality assemblies [7–9], *de novo* transcriptome studies in poorly characterized taxonomic groups continue to be challenging because of difficulties with annotation. This is due to the lack of information available on the genes of interest in closely related organisms. The subphylum Crustacea represents one such taxonomic group for which limited information exists. The Ensembl genome database for metazoan species contains mainly Diptera (flies), Nematoda (worms) and Hymenoptera (ants), but information on only a single crustacean: the common water flea, *Daphnia Pulex* [10]. Furthermore, the number of NCBI Entrez records in the invertebrate taxonomic branch shows huge under-representation of crustaceans. In total, there are approximately 2,300,000 nucleotide sequences in the subphylum Crustacea; in comparison the order Hymenoptera which alone contains almost 2,600,000 nucleotide sequences (numbers dated to April 2014). Consequently, subtaxa within the subphylum Crustacea contain less information: Decapoda (shrimps, crabs, lobsters and crayfish) have a total of 478,358 nucleotide and 44,210 protein sequences available.

The European shore crab (or green crab), *C. maenas,* is a keystone species in the European marine environment and is the only crustacean on the Global Invasive Species Database [11], with invasions into Australia, South Africa and the United States [12]. In such locations, *C. maenas* threatens local fishing industries, for example the destruction of the soft-shell clam (*Mya arenaria*) fishery in New England [13]. *C. maenas* is also an important study species for biomonitoring and ecotoxicology [14, 15]. The species has been used in monitoring for heavy metal contamination [16], metal toxicity studies [17–22], and more recently in exposures studies with nanomaterials [23] and microplastics [24]. Pathological studies are a new area wherein *C. maenas* could play a role. A study investigating infection of crustaceans with White Spot Syndrome Virus (WSSV), recognized as the most significant pathogen affecting global shrimp aquaculture, showed that *C. maenas* are relatively resistant to the virus [25–27]. Despite its importance in these research areas, and its biological significance in the environment, the available molecular resources for *C. maenas* are extremely limited. To date, sequence data for this species comprises approximately 15,000 EST sequences and several hundred nucleotide and protein sequences [28].

Given the ecological importance of *C. maenas,* together with its wider general utility for research purposes, we aimed to sequence, assemble and annotate a shore crab transcriptome. We further set out to establish the relative expression profiles of all sequenced transcripts in different body tissues and organs, and to characterize immune pathways against those known for other invertebrates as a resource for future investigations on the response of this host to pathogens.

Results and discussion

RNA sequencing and assembly

Twelve sequence libraries corresponding to 12 pooled tissue samples from adult male and female *C. maenas* were sequenced on an Illumina HiSeq 2500 platform and yielded a total of 138,863,679 paired reads across all tissues. After removal of low quality reads through quality filtering, there were 96,247,762 remaining paired reads. On average 8.0 ± 1.7 million read pairs were obtained for each tissue and the distribution of the reads per pooled transcript sample is presented in Table 1. The filtered RNA-Seq data were used for *de novo* transcriptome assembly using the Trinity pipeline with default parameters. The assembled transcriptome encompassed 196,966,469 bp distributed over 153,669 loci, represented by 212,427 transcripts (Table 2). The transcript lengths had a median of 380 bp and a mean of 992 bp (standard deviation = 1363 bp), and ranged between 201 bp and 24,848 bp (Additional file 1 shows the length distribution of assembled transcripts). The transcriptome N50 was calculated to be 2,102 bp. 75.2 % of the read pairs could be mapped back to the *de novo* assembled transcriptome using the bowtie2 aligner.

A total of 231 out of the 248 highly conserved eukaryotic "core" genes were identified completely (93.15 %) and 245 genes (98.79 %) partially in the transcriptome by the CEGMA pipeline [29], indicating that the transcriptome contains a near complete set of core eukaryotic genes.

Table 1 Number of read pairs obtained for each crab tissue before and after removal of adapter sequences and quality filtering

Tissue sample	Number of read pairs	Number of clean read pairs
Eggs	9,337,648	6,614,044
Epidermis	11,929,821	8,302,718
Eye	13,463,765	9,430,381
Gill	10,110,102	7,234,304
Haemolymph	10,611,241	7,233,253
Heart	9,657,081	6,717,788
Hepatopancreas	9,216,408	6,471,110
Intestine	8,685,232	5,765,077
Muscle	17,251,355	11,749,555
Nerve	14,278,257	9,670,912
Ovary	11,125,170	7,869,190
Testis	13,197,599	9,189,430
Total	138,863,679	96,247,762

Table 2 Transcriptome statistics

Description	Value
Number of loci	153,669
Number of transcripts	212,427
Maximum transcript length (bp)	24,848
Minimal transcript length (bp)	201
Mean transcript length (bp)	992
Standard deviation (bp)	1363
Median transcript length (bp)	380
Total length (bp)	196,966,469
N50 (bp)	2,102

Transcriptome characterization

Several approaches were taken to annotate the assembled transcripts. Firstly, the transcript sequences were compared to existing *C. maenas* EST sequences in the NCBI database using BLASTn. In total, 19,981 sequences (9.4 % of the total number of transcripts) showed high similarity to 4,759 EST sequences (30.6 % of total *C. maenas* ESTs in NCBI; Table 3). This indicates that the majority of transcripts in the assembly were previously un-reported for *C. maenas*. A broader sequence homology search was performed using BLASTx against the NCBI non-redundant *nr* protein database and hits were found for 62,804 (29.6 %) of the transcripts using an e-value threshold of 1e-3. Open reading frames were identified in 58,383 (27.5 %) of transcripts and the majority of the predicted peptides (41,108), corresponding to 70.4 % of all predicted peptides were annotated using the UniProt/Swissprot database (with an e-value cut-off of 1e-5). Furthermore, conserved Pfam domains were assigned to 37,776 (67.4 %) of the peptides and 4,132 (1.9 %) of these peptides appeared to contain signal peptides (Table 3) as determined by SignalP. Transcriptome annotation details can be found in Additional files 2 and 3.

Table 3 Number of annotated transcripts and open reading frames (identified by TransDecoder) using different annotation methods and sequence databases

Input	Annotation method	Number of annotated transcripts
All transcripts	BLASTx – NCBI nr protein	62,804 (29.6 %)
All transcripts	BLASTn – *C.maenas* EST	19,891 (9.4 %)
All transcripts	BLAST2GO	8,091 (3.8 %)
All transcripts	TransDecoder ORF finder	58,383 (27.5 %)
All transcripts	KEGG	30,352 (14.3 %)
Open reading frames	BLASTp – UniProt/ SwissProt	41,108 (70.4 %)
Open reading frames	Pfam	37,776 (67.4 %)
Open reading frames	SignalP	4,132 (1.9 %)
Open reading frames	TmHMM	0 (0.0 %)

Transcriptome functional annotation

Gene Ontology (GO) terms were assigned to 53,766 (25.3 %) of the annotated transcripts and 47.23 % of the annotated predicted peptides (UniProt/Swissprot; Table 3) by BLAST2GO [30]. The most common GO terms were protein binding (10.93 %), cytoplasm (10.93 %), nucleus (10.07 %), plasma membrane (6.55 %) and membrane (6.25 %). The most common annotations for the three gene ontology trees are presented in Table 4, and a full list of transcript annotations is available in Additional file 4.

Taxonomy

The BLASTx output was used as input for MEGAN4 to illustrate the taxonomic origin of BLAST hits for the transcriptome in a phylogenetic tree. A partially collapsed phylogenetic tree is presented in Fig. 1. The taxon with the largest number of sequence homologies was the pancrustacean taxon wherein 21,642 *C. maenas* transcripts showed similarity. Within this taxon, transcripts were split between the crustacean and hexapoda taxa. Since *C. maenas* is a crustacean species it is expected that a large proportion of transcripts show similarity to sequences derived from this taxon. However, due to the limitations in crustacean genomic resources a significant proportion of transcripts mapped to related sequences in the hexapoda taxon instead (containing e.g. *Drosophila melanogaster*). Furthermore, it can be seen that a variety of sequences were derived from micro-organisms (e.g. bacteria, fungi and viruses), which may correspond to transcripts originating from micro-organisms living within the *C. maenas* hosts, and/or may reflect contamination of kits and samples with environmental micro-organisms [31]. To remove these potential contaminating transcripts from the transcriptome we filtered the transcriptome for sequences that mapped to the metazoan taxon. Following the application of this filtering step, a transcriptome encompassing 59,392 transcripts was retained and used in subsequent analysis.

Differential gene expression

Transcript expression in the twelve tissue types was estimated by the RSEM program [32]. Next, differentially expressed transcripts were identified through comparing gene expression profiles of each sampled tissue to the others. The number of differentially expressed (metazoan) transcripts for the various tissues ranged between 1,223 in gill and 2,741 in hepatopancreas (FDR < 0.01; Table 5). All tissues showed enrichment for Gene Ontology (GO) terms; the top five for every tissue are listed in Table 6 (a complete list is presented in Additional file 5).

Table 4 Identification, sequence similarity and Gene Ontology annotation statistics of peptide sequences in the transcriptome

Description		Number of sequences	Percentage of sequences (%)
Transcripts		212,427	
TransDecoder peptides		58,383	
Peptides with Swissprot /Uniprot annotation		41,108	70.41
GO annotated transcripts		53,766	25.31
GO annotated peptides		19,423	47.23
GO tree	GO	Count	%
Cellular Component	cytoplasm	2,122	10.9
	nucleus	1,955	10.1
	plasma membrane	1,272	6.6
	membrane	1,213	6.3
	cytosol	1,195	6.2
Molecular Function	protein binding	2,577	13.3
	binding	1,071	5.5
	ATP binding	755	3.9
	metal ion binding	569	2.9
	protein homodimerization activity	485	2.5
Biological Process	cellular process	597	3.1
	regulation of cellular process	539	2.8
	primary metabolic process	446	2.3
	response to stimulus	419	2.2
	transport	416	2.1

The enriched GO terms often reflected the function of the tissue e.g. structural constituent of cuticle in eggs, angiogenesis in haemolymph and sarcolemma in muscle. In several tissues the link to function is not very clear in the top five, but becomes apparent in other enriched terms. For example, in the eye, phototransduction (FDR = 9.42e−4) and detection of light stimulus (FDR = 1.01e−3) were over-represented; contractile fibre (FDR = 6.72e−3) and sarcomere (FDR = 7.15e−3) were enriched in the heart tissue and finally, the epidermis and ovary tissues yielded only three enriched annotations (Table 6).

Immune pathway characterization in *C. maenas*

Application of *C. maenas* as a model organism to study crustacean infectious diseases requires insight in the organism's immune system. Since crustaceans do not have adaptive immune systems, innate immune strategies will predominate in this organism when responding to pathogenic insults. We investigated the presence of several innate immune system pathways in the *C. maenas* transcriptome and mapping the transcripts to pathways in the KEGG database. In total 30,352 (14.3 %) of transcripts were annotated to a KEGG orthology group (Table 3). The KEGG server [33] allows mapping of the present orthology groups to pathways in the KEGG database and visualization of presence/absence of their components. Li *et al.* 2013 characterized a selection of innate immune pathways in the hepatopancreas transcriptome of the mitten crab *Eriocheir sinensis*, including the RNAi pathway, Toll-like receptor pathway, immune deficiency (IMD) pathway, the JAK-STAT and mitogen activated protein kinase (MAPK) signalling pathways [34]. We characterized the same pathways in the *C. maenas* transcriptome with additions including the endocytosis pathway. The latter is not directly related to the immune response but many viruses utilize its machinery to gain entry to host cells [35]. Its characterization can thus be important for investigations of viral infections.

Pathogen associated molecular pattern recognition

The first stage in immune defence is the identification of invading pathogens by an organism. In this process a distinction between cells from the organism itself and those of the invading pathogens needs to occur. To achieve this, the innate immune system employs a group of pattern recognition receptors (PRRs) that are able to recognize pathogen associated molecular patterns (PAMPs). Examples of PAMPs include lipopolysaccharides, peptidoglycans and β-1,3-glucans [36] and groups of PRRs include gram-negative binding proteins

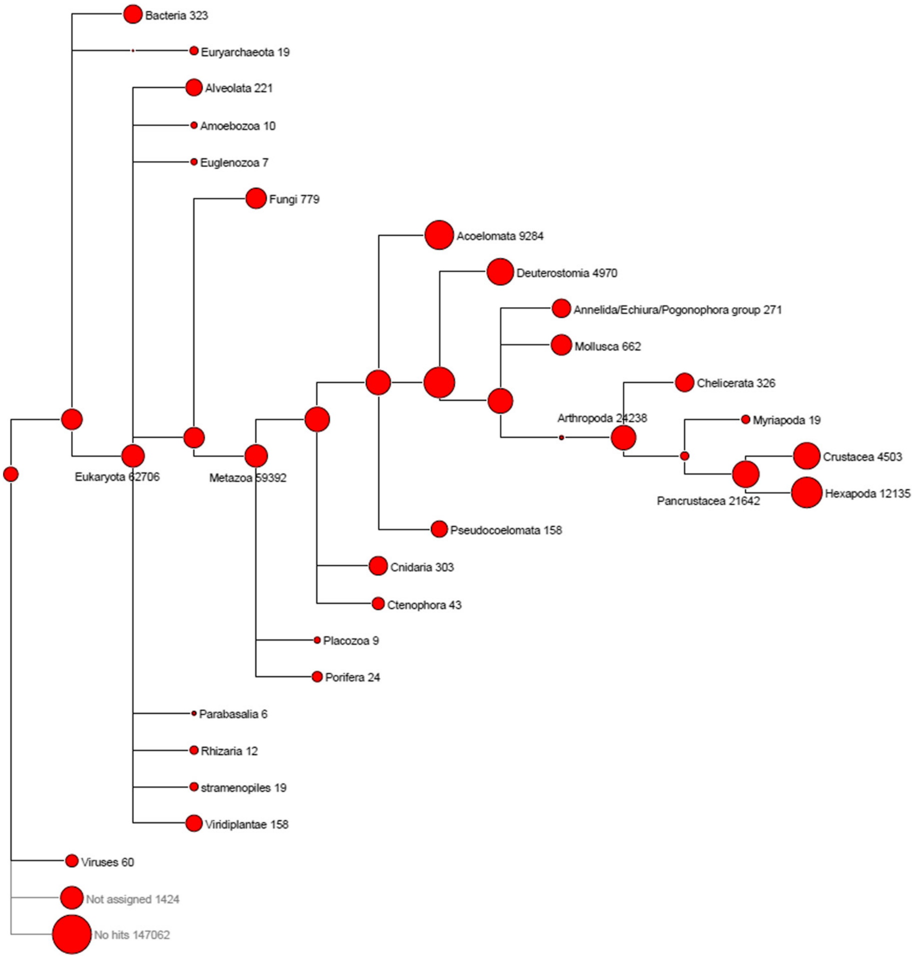


Fig. 1 Taxonomic classifications of *Carcinus maenas* transcripts. Partially collapsed phylogenetic tree produced by MEGAN4. Numbers illustrate the number of transcripts representing each taxa. Within the metazoan taxon, the pancrustacea represented the largest taxonomic group

Table 5 Differentially expressed transcripts in specific tissues

Tissue	Differentially expressed transcripts
eggs	1,605
epidermis	1,339
eye	1,312
gill	1,223
Haemolymph	2,008
heart	1,226
hepatopancreas	2,741
intestine	1,519
muscle	2,200
nerve	1,989
ovary	1,751
testis	1,391

(GNBPs), peptidoglycan recognition proteins (PGRP), thioester containing proteins and lectins [36]. Upon successful pathogen recognition, PRRs initiate immune responses.

C. maenas transcripts that show sequence similarity to known PRR groups are shown in Table 7. Representatives of most groups of PRR have counterparts in the *C. maenas* transcriptome as identified through sequence similarity, often to sequences derived from organisms that are closely related to *C. maenas*. One group that is not represented are the PGRPs, this has also been reported in other crustacean species [37, 38]. Down syndrome cell adhesion molecule (Dscam) is a PAMP recognition protein that has been hypothesized to be involved in immune memory (reviewed in Armitage et al. 2014 [39]). This gene can produce many isoforms, and initial findings suggested that it played an important role in the development of the nervous system in invertebrates where Dscam isoforms aid in the discrimination

Table 6 Top 5 most enriched Gene Ontology terms in specific tissues

Tissue	GO-ID	Term	*P* - value	FDR
Eggs	GO:0042302	structural constituent of cuticle	9.86e-11	1.07e-6
Eggs	GO:0003677	DNA binding	5.25e-7	2.86e-3
Eggs	GO:0006260	DNA replication	2.78e-6	1.01e-2
Eggs	GO:0006261	DNA-dependent DNA replication	5.76e-6	1.57e-2
Eggs	GO:0001708	cell fate specification	1.06e-5	2.30e-2
Epidermis	GO:0018298	protein-chromophore linkage	3.85e-7	2.56e-3
Epidermis	GO:0015772	oligosaccharide transport	7.05e-7	2.56e-3
Epidermis	GO:0015766	disaccharide transport	7.05e-7	2.56e-3
Eye	GO:0003008	system process	7.00e-12	7.61e-8
Eye	GO:0050877	neurological system process	2.36e-11	1.28e-7
Eye	GO:0022834	ligand-gated channel activity	9.64e-9	2.62e-5
Eye	GO:0015276	ligand-gated ion channel activity	9.64e-9	2.62e-5
Eye	GO:0070011	peptidase activity, acting on L-amino acid peptides	1.57e-8	3.42e-5
Gill	GO:0070160	occluding junction	1.71e-6	4.20e-3
Gill	GO:0005344	oxygen transporter activity	1.84e-6	4.20e-3
Gill	GO:0015671	oxygen transport	1.84e-6	4.20e-3
Gill	GO:0015669	gas transport	1.84e-6	4.20e-3
Gill	GO:0005923	tight junction	2.65e-6	4.20e-3
Haemolymph	GO:0001525	angiogenesis	5.76e-11	6.27e-7
Haemolymph	GO:0048514	blood vessel morphogenesis	1.80e-9	9.79e-6
Haemolymph	GO:0001568	blood vessel development	1.25e-8	4.54e-5
Haemolymph	GO:0001944	vasculature development	3.97e-8	1.08e-4
Haemolymph	GO:0009653	anatomical structure morphogenesis	1.46e-7	1.65e-4
Heart	GO:0016328	lateral plasma membrane	1.78e-7	1.94e-3
Heart	GO:0006768	biotin metabolic process	1.28e-6	3.04e-3
Heart	GO:0004736	pyruvate carboxylase activity	1.28e-6	3.04e-3
Heart	GO:0005344	oxygen transporter activity	1.67e-6	3.04e-3
Heart	GO:0015671	oxygen transport	1.67e-6	3.04e-3
Hepatopancreas	GO:0016491	oxidoreductase activity	6.35e-16	6.91e-12
Hepatopancreas	GO:0003824	catalytic activity	2.87e-11	1.56e-7
Hepatopancreas	GO:0044710	single-organism metabolic process	2.06e-10	7.47e-7
Hepatopancreas	GO:0005576	extracellular region	5.68e-10	1.20e-6
Hepatopancreas	GO:0005764	lysosome	6.61e-10	1.20e-6
Intestine	GO:0016337	cell-cell adhesion	5.72e-9	6.22e-5
Intestine	GO:0005548	phospholipid transporter activity	7.29e-8	3.97e-4
Intestine	GO:0006022	aminoglycan metabolic process	1.56e-7	4.06e-4
Intestine	GO:0015917	aminophospholipid transport	2.09e-7	4.06e-4
Intestine	GO:0004012	phospholipid-translocating ATPase activity	2.09e-7	4.06e-4
Muscle	GO:0042383	sarcolemma	1.93e-11	2.10e-7
Muscle	GO:0031674	I band	7.82e-11	4.25e-7
Muscle	GO:0006811	ion transport	2.87e-10	1.04e-6
Muscle	GO:0030018	Z disc	1.94e-9	5.29e-6
Muscle	GO:0044449	contractile fiber part	2.54e-9	5.52e-6

Table 6 Top 5 most enriched Gene Ontology terms in specific tissues *(Continued)*

Nerve	GO:0015277	kainate selective glutamate receptor activity	1.39e-14	1.51e-10
Nerve	GO:0004872	receptor activity	2.54e-12	1.38e-8
Nerve	GO:0048172	regulation of short-term neuronal synaptic plasticity	5.16e-12	1.87e-8
Nerve	GO:0004970	ionotropic glutamate receptor activity	1.02e-11	2.77e-8
Nerve	GO:0048168	regulation of neuronal synaptic plasticity	4.92e-11	1.07e-7
Ovary	GO:0016459	myosin complex	1.43e-7	1.56e-3
Ovary	GO:0018298	protein-chromophore linkage	1.67e-6	9.10e-3
Ovary	GO:0036002	pre-mRNA binding	1.37e-5	4.96e-2
Testis	GO:0008499	UDP-galactose:beta-N-acetylglucosamine beta-1,3-galactosyltransferase activity	5.07e-17	5.52e-13
Testis	GO:0035250	UDP-galactosyltransferase activity	1.45e-16	7.86e-13
Testis	GO:0005797	Golgi medial cisterna	1.38e-15	5.00e-12
Testis	GO:0048531	beta-1,3-galactosyltransferase activity	6.12e-15	1.66e-11
Testis	GO:0008378	galactosyltransferase activity	1.26e-14	2.75e-11

Table 7 *Carcinus maenas* pathogen associated molecular pattern recognition genes

PRP group	Transcript	Identity (%)	Length	E-value	Query	Ancestor
GNBP	comp44152_c0_seq1	42.06	340	1.00e-65	gi|300507044 : gram-negative binding protein [*Artemia sinica*]	Crustacea
	comp44453_c0_seq (1–2)	58.06	341	3.00e-123	gi|62122584 : GNBP [*Oryzias latipes*]	Bilateria
	comp74133_c0_seq1	44.8	346	8.00e-88	gi|62122584 : GNBP [*Oryzias latipes*]	Bilateria
	comp83740_c0_seq (1–5)	46.02	339	5.00e-87	gi|62122584 : GNBP [*Oryzias latipes*]	Bilateria
	comp19734_c0_seq1	41.55	142	8.00e-32	gi|62122584 : GNBP [*Oryzias latipes*]	Bilateria
	comp136078_c0_seq1	62.96	81	3.00e-26	gi|62122584 : GNBP [*Oryzias latipes*]	Bilateria
	comp75261_c0_seq1	27.57	243	6.00e-22	gi|62122584 : GNBP [*Oryzias latipes*]	Bilateria
TECP	comp85313_c2_seq1	39.94	318	6.00e-63	gi|385049105 : thioester containing protein 3, partial [*Daphnia parvula*]	Crustacea
	comp65627_c0_seq1	46.34	246	3.00e-58	gi|54644242 : Thioester-containing protein 6 [*Drosophila pseudoobscura pseudoobscura*]	Pancrustacea
	comp87629_c0_seq4	74.36	234	6.00e-101	gi|331031264 : TEP isoform 2 [*Pacifastacus leniusculus*]	Pleocyemata
	comp74624_c1_seq1	40.65	310	8.00e-56	gi|385049099 : thioester containing protein 3, partial [*Daphnia pulex*]	Crustacea
	comp65627_c1_seq1	36.78	590	6.00e-118	gi|54644242 : Thioester-containing protein 6 [*Drosophila pseudoobscura pseudoobscura*]	Pancrustacea
	comp74624_c2_seq1	37.83	534	1.00e-105	gi|54644242 : Thioester-containing protein 6 [*Drosophila pseudoobscura pseudoobscura*]	Pancrustacea
	comp85313_c0_seq1	38.14	430	1.00e-80	gi|568250870 : thioester-containing protein [*Anopheles darlingi*]	Pancrustacea
	comp103781_c0_seq1	43.36	113	4.00e-22	gi|54644242 : Thioester-containing protein 6 [*Drosophila pseudoobscura pseudoobscura*]	Pancrustacea
C-Type Lectin	comp69837_c0_seq1	43.15	146	1.00e-25	gi|558633447 : C-type lectin [*Marsupenaeus japonicus*]	Decapoda
	comp86095_c0_seq (1–2)	43.92	148	2.00e-25	gi|558633447 : C-type lectin [*Marsupenaeus japonicus*]	Decapoda
	comp68699_c0_seq1	38.89	144	1.00e-24	gi|558633447 : C-type lectin [*Marsupenaeus japonicus*]	Decapoda
	comp87731_c3_seq (2–3)	33.78	225	8.00e-25	gi|657397985 : C-type lectin receptor-like tyrosine-kinase plant [*Medicago truncatula*]	Eukaryota
	comp88573_c0_seq (1–2)	57.5	80	4.00e-22	gi|676264911 : C-type lectin domain family 3 member A [*Fukomys damarensis*]	Bilateria
	comp90611_c0_seq1	64.56	158	5.00e-60	gi|575878533 : C-type lectin [*Scylla paramamosain*]	Portunoidea

of neuritis [39]. Dscam isoforms were later found to be able to recognise pathogens, aiding in phagocytosis [40]. In concordance with this hypothesis, the *C. maenas* Dscam gene appears to encode many isoforms, and in total 242 transcripts with significant similarity to Dscam sequences in NCBI were found in the transcriptome.

The immune responses initiated by these PRRs can occur at a transcriptional level, e.g. activation of Toll and IMD can aid in phagocytosis e.g. Dscam binding, or can initiate proteolytic cascades leading to melanization.

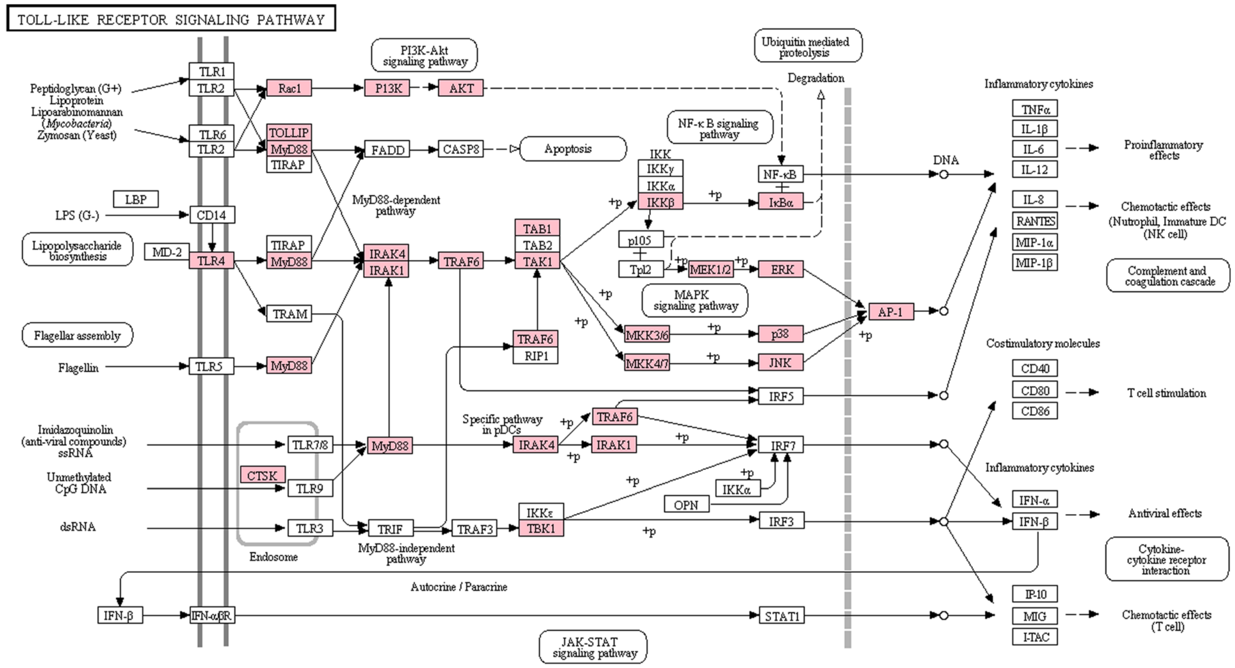
Toll-like receptor pathway

The Toll receptor pathway is a signalling route that responds to the presence of PAMPs by ultimately activating Nf-κB [41]. In mammals, Toll-like receptors (TLR) bind to PAMPs resulting in dimerization. Upon forming dimers, the TLRs recruit MyD88 and subsequently IRAK kinases. After IRAK kinases activate TRAF6, its binding to TAK1 and IKKβ ultimately frees Nf-κB to diffuse into the nucleus [42]. In invertebrates, such as *D. melanogaster,* the mechanism is slightly different, and instead of directly binding PAMPs, TLRs respond to the Toll ligand Spätzle [41].

The KEGG database contains a version of the Toll-like receptor pathway which was used to visualize the coverage of this pathway in the *C. maenas* transcriptome (see Fig. 2). Homologues were found for most of the components in the paths from TLR to NF-κB and activator protein-1 (AP-1). Since KEGG is targeted towards vertebrate genes and pathways, a characterization of an invertebrate Toll signalling pathway was also performed (see Methods for pathway analysis strategy). Components of the *D. melanogaster* Toll signalling pathway were taken from Li *et al.* [34] and Kingsolver *et al.* [41] and investigated for presence and expression in the assembled transcriptome. Transcripts with significant sequence similarity to most of the Toll pathway components were found in the transcriptome (Additional file 6). Tube, an IRAK homolog, was not identified in the *C. maenas* transcriptome. Successfully identified transcripts were found to be expressed across all tissues (Additional file 7), and the median expression values varied from 82.4 FPKM for myD88 to 5576.7 FPKM for Toll.

IMD pathway

The IMD pathway is also activated upon pathogen recognition, in particular by Gram-negative bacteria. Similar to Toll-like receptors, the binding of peptidoglycan by PGRPs leads to dimerization [41]. After the dimerization, the signal is transmitted through IMD, as well as FADD and DREDD. Activation of DREDD leads to poly-ubiquitination of IMD [41], binding of TAK1 and assembly of the IKK complex. Relish phosphorylation is promoted by IKK, and an event followed by cleavage of Relish by DREDD cause translocation of the



04620 6/7/13
(c) Kanehisa Laboratories

Fig. 2 Toll-like receptor signalling pathway coverage. The Toll-like receptor signalling pathway in the KEGG database. Proteins in the pathway are depicted by boxes while arrows depict signalling routes. Pathway components with homologues in the *Carcinus maenas* transcriptome are highlighted in pink

N-terminal end to the nucleus where it regulates the expression of effector molecules [41]. Since the KEGG database does not contain the IMD pathway, the KEGG TNF-signalling pathway was used instead. As for the Toll-like receptor pathway, homologues also were found for most constituents of the TNF-signalling pathway (see Fig. 3). Manual identification of IMD pathway components derived from Kingsolver *et al.* [41] showed that FADD was the only absent component in the *C. maenas* transcriptome (Additional file 6). IMD itself was only expressed in three out of twelve tissues (eye, ovary and haemolymph) whereas the rest of the IMD pathway was expressed across all tissue types (Additional file 8).

JAK-STAT signalling pathway

The JAK-STAT signalling pathway mediates the response to chemical messenger molecules like cytokines. It has been shown that STAT signalling is activated upon WSSV infection in shrimp [43]. JAK tyrosine kinases bind to cytokine receptors and upon ligand binding they phosphorylate tyrosine residues on those receptors [44]. STAT is able to bind and subsequently be phosphorylated by JAK [44]. Following phosphorylation, STAT forms dimers, translocates to the nucleus and organizes the response to the signalling molecule by altering gene expression [44]. Inhibitors of JAK-STAT signalling are present at several stages and include dominant negative co-receptors, prevention of STAT recruitment by SOCS (suppressor of cytokine signalling) and protein inhibitors of activated STAT (PIAS) [44]. The KEGG reference pathway and coverage in the transcriptome are presented in Fig. 4. Most of the components of the JAK-STAT pathway have a homologue in the *C. maenas* transcriptome. The pathway in Fig. 4 shows that only the cytokine receptor was not identified by the KEGG annotation. However one transcript (comp79993_c0_seq2) showed highly significant sequence homology to the cytokine receptor of *Harpegnathos saltator* (e = 3.00e−74) and the domeless receptor of *Tribolium castaneum* (e = 2.00e−51).

Response proteins

The signalling cascade through the IMD, Toll and JAK-STAT pathways results in a transcriptional immune response mediated by transcription factors like STAT and NF-kB. One part of this immune response includes antimicrobial peptides (e.g. anti-lipopolysaccharide factor (ALF) and lysozyme), which have evolved to attack pathogens [45, 36]. In addition to antimicrobial peptides, the

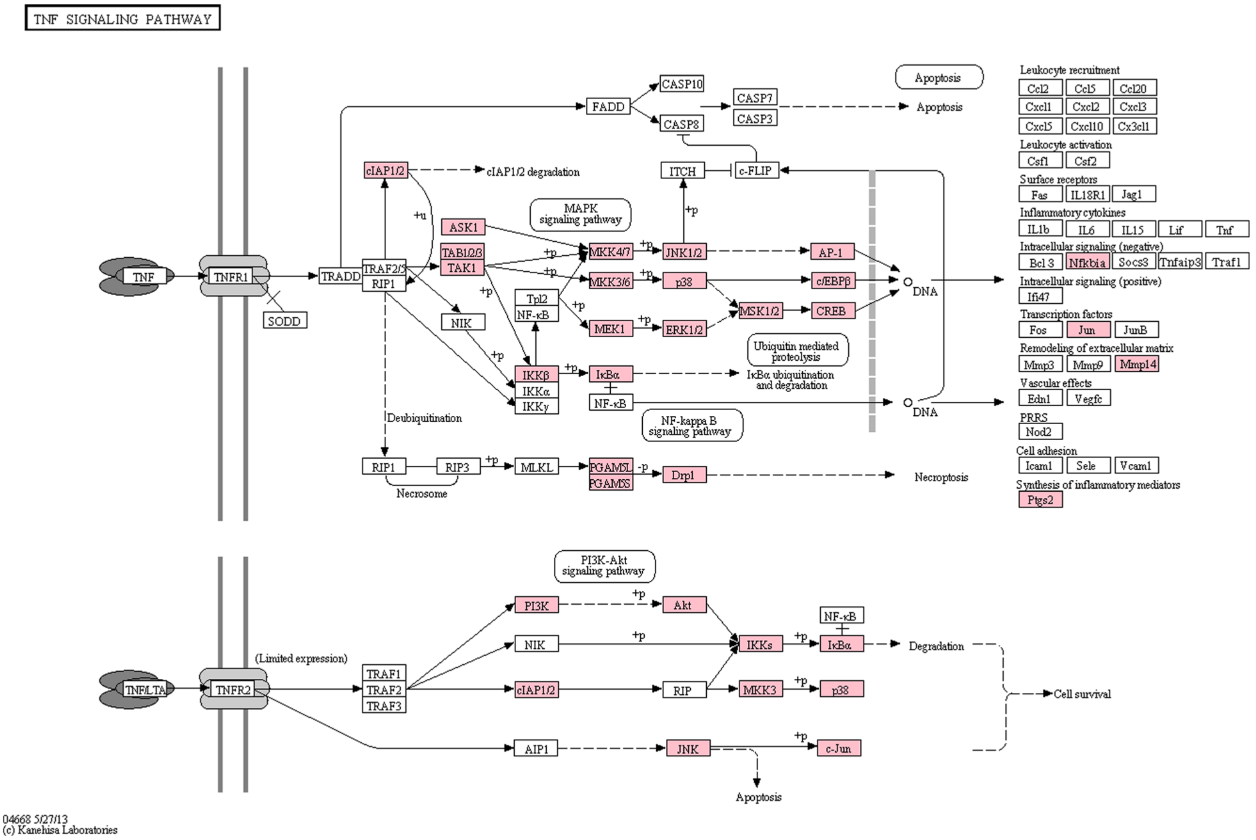


Fig. 3 TNF signalling pathway. Overview of the KEGG TNF signalling pathway, components depicted as in Fig. 2. Components with homologues in the *Carcinus maenas* transcriptome are highlighted in pink

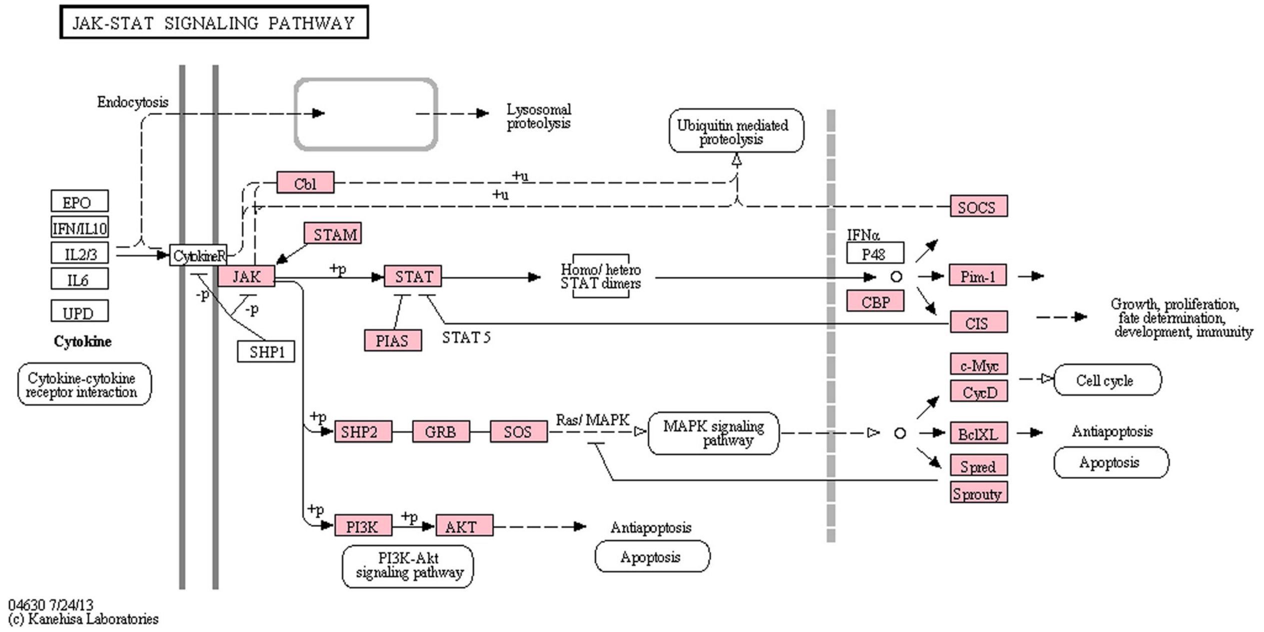


Fig. 4 JAK-STAT signalling pathway. Overview of the KEGG JAK-STAT signalling pathway. Components with homologues in the *Carcinus maenas* transcriptome are highlighted in pink

innate immune system also employs nitric oxide as a defensive molecule. Nitric oxide is an important redox activated signalling molecule and can be produced in large concentrations by nitric oxide synthase 2 (NOS-2), an enzyme synthesized as a response to PRR activation [46]. Response proteins identified in the *C. maenas* transcriptome are listed in Table 8 along with their target pathogen type, as described in Tassanakajon et al. [45]. Neither penaeidins [47] nor stylicins [48] were identified for *C. maenas* and we hypothesise that both are probably limited to penaeid shrimp species. The antimicrobial arsenal of *C. maenas* includes ALF, lysozyme, crustins, carcinin and inducible nitric oxide synthase. It is possible that the *C. maenas* transcriptome also contains novel anti-microbial peptides but to identify them will require exposure studies to trigger their activation.

Melanization pathway

The *C. maenas* innate immune system also contains a more direct response to pathogen infection in the form of the melanization pathway. Activated within minutes after infection, melanization damages and encapsulates invading

Table 8 *Carcinus maenas* Immune system response proteins

Response protein	Transcript	Identity (%)	Length	E-value	Query	Ancestor
ALF	comp79835_c0_seq2	65.98	97	2.00e-34	gi|302138013 : anti-lipopolysacharide factor [*Fenneropenaeus indicus*]	Decapoda
Crustin	comp88229_c1_seq1	56.36	110	8.00e-31	gi|162945361 : crustin antimicrobial peptide [*Scylla paramamosain*]	Portunoidea
	comp91133_c0_seq1	65.38	78	7.00e-24	gi|255653868 : crustin 1 [*Panulirus japonicus*]	Pleocyemata
Carcinin	comp88229_c1_seq1	86.36	110	1.00e-49	gi|18157188 : carcinin [*Carcinus maenas*]	Carcinus maenas
Lysozyme	comp83352_c1_seq4	41.13	124	4.00e-23	gi|675374133 : Lysozyme 1, partial [*Stegodyphus mimosarum*]	Arthropoda
	comp83352_c1_seq2	41.13	124	4.00e-23	gi|675374133 : Lysozyme 1, partial [*Stegodyphus mimosarum*]	Arthropoda
	comp83352_c1_seq1	41.13	124	4.00e-23	gi|675374133 : Lysozyme 1, partial [*Stegodyphus mimosarum*]	Arthropoda
iNOS	comp89503_c2_seq (1–26)	52.6	308	1.00e-96	gi|13359094 : nitric oxide synthase 2 [*Meriones unguiculatus*]	Pancrustacea

pathogens with melanin [49]. The production of melanin from phenols and quinones generates reactive oxygen species that are damaging to the pathogen. Synthesis of quinones is catalyzed by the phenol oxidase (PO) enzyme. PO is readily available as a precursor (proPO) that is activated through proteolysis, ensuring a fast response time. Recognition of PAMPs by PRRs leads to activation of a serine protease cascade that ends with the activation of PO [45, 49, 50]. The proteolytic cascade is regulated by serpins that act as serine protease inhibitors [49]. Members of the melanization pathway as described in Tang 2009 [49] and transcripts with significant sequence similarity are listed in Additional file 6. The upstream proteases of proPO: MP1, Sp7 and the activating enzyme PPAE and prophenoloxidase itself are identified. Transcripts coding for the transcription factors serpent and lozenge, controlling the expression of proPO [49], and Peroxinectin, a protein that is associated with the proPO pathway and aids in cellular adhesion of haemocytes to pathogens [51] were also found. The expression of proPO varied across tissues (see Fig. 5), and was particularly high in the hepatopancreas and ovary.

RNAi pathway

RNA interference (RNAi) is one of the major antiviral pathways within the invertebrate innate immune system [52]. The pathway produces small interfering RNA molecules (siRNAs) from virus derived dsRNA [41]. In short, dsRNA is recognized by Dicer proteins that subsequently cleave it to 21 nucleotide (nt) siRNAs. siRNAs are loaded into the RISC complex, which utilizes argonaute (Ago) protein to cleave viral RNAs targeted by the siRNA, and thus silencing expression [41]. The RNAi pathway can also be employed to silence specific genes in cells and forms the basis of antiviral immunity strategies, a topic explored in La Fauce *et al.* 2012 [53]. Identification of components of the RNAi pathway was based on those listed in Wang *et al.* 2014 [52], results are shown in Table 9. *D. melanogaster* has distinct functions for dicer-1 and dicer-2, the first being involved in the miRNA pathway and the latter in siRNA [54, 41, 55]. Both dicer-1 and dicer-2 were identified in *C. maenas* suggesting that a similar division of tasks could exist in this organism.

Endocytosis pathway

The endocytosis pathway plays a crucial role in viral challenges. Whereas some viruses are able to enter the cytosol directly, the majority require uptake via endocytosis [35]. Viral particles can enter endosomes via various endocytotic mechanisms (e.g. clathrin-mediated endocytosis,

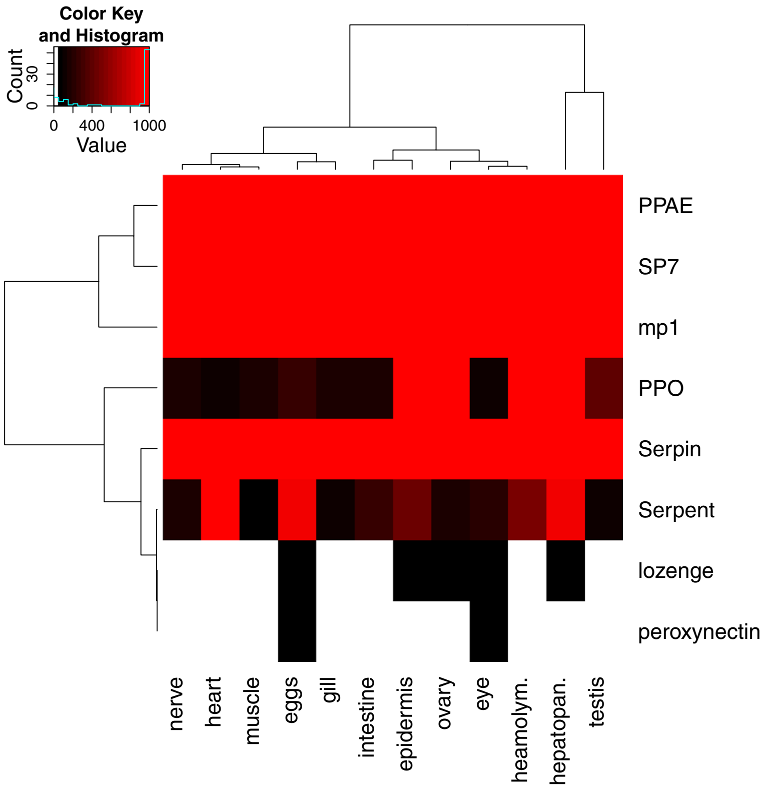


Fig. 5 Melanization pathway expression. Expression of melanization pathway components in twelve *Carcinus maenas* tissues. The expression values are presented in FPKM, values of 0 are coloured white and values over 10000 FPKM are binned together

Table 9 *Carcinus maenas* RNAi pathway components

RNAi	Transcript	Identity (%)	Length	E-value	Query	Ancestor
TRBP	comp79785_c0_seq (1–2)	83.97	343	2.00e-167	gi|332271591 : TAR RNA-binding protein isoform 1 [*Marsupenaeus japonicus*]	Decapoda
	comp79200_c0_seq (1–2)	36.74	460	4.00e-77	gi|110825988 : probable methyltransferase TARBP1 [*Homo sapiens*]	Bilateria
	comp49673_c0_seq1	46.34	205	2.00e-41	gi|444174849 : TAR RNA-binding protein 1 [*Penaeus monodon*]	Decapoda
R2D2	comp79785_c0_seq (1–2)	48.86	350	8.00e-81	gi|619831236 : R2D2 [*Bemisia tabaci*]	Pancrustacea
	comp49673_c0_seq1	38.32	167	6.00e-24	gi|619831236 : R2D2 [*Bemisia tabaci*]	Pancrustacea
drosha	comp87202_c0_seq1	93.37	829	0	gi|396941645 : drosha [*Marsupenaeus japonicus*]	Decapoda
Dicer2	comp90354_c0_seq (1–11)	47.73	1253	0	gi|402534262 : Dicer-2 [*Marsupenaeus japonicus*]	Decapoda
Dicer1	comp85246_c1_seq1	77.95	1578	0	gi|195424855 : dicer-1 [*Litopenaeus vannamei*]	Decapoda
	comp90354_c0_seq (5–6)	31	658	1.00e-83	gi|195424855 : dicer-1 [*Litopenaeus vannamei*]	Decapoda
	comp55144_c0_seq1	61.06	113	3.00e-37	gi|283827860 : dicer-1 [*Marsupenaeus japonicus*]	Decapoda
	comp77864_c(1–2)_seq (1–2)	83.67	98	2.00e-40	gi|195424855 : dicer-1 [*Litopenaeus vannamei*]	Decapoda
ago2	comp81967_c(1–2)_seq1	37.16	802	5.00e-139	gi|563729913 : argonaute2 [*Penaeus monodon*]	Decapoda
	comp41784_c0_seq1	52.74	876	0	gi|563729913 : argonaute2 [*Penaeus monodon*]	Decapoda
	comp76466_c0_seq1	42.51	821	0	gi|563729913 : argonaute2 [*Penaeus monodon*]	Decapoda
ago1	comp81967_c1_seq1	89.45	758	0	gi|321468117 : putative Argonaute protein [*Daphnia pulex*]	Crustacea
	comp41784_c0_seq1	43.65	811	0	gi|321468117 : putative Argonaute protein [*Daphnia pulex*]	Crustacea
	comp76466_c0_seq1	41.58	671	4.00e-148	gi|321468117 : putative Argonaute protein [*Daphnia pulex*]	Crustacea

caveolar-mediated endocytosis, or micropinocytosis). Decreasing pH in the endosome environment is a cue to the viral particles, which then penetrate into the cytosol [35]. This indicates that there are important interactions between components of the endocytosis pathway and viral proteins, e.g. cellular Rab7 can interact with the VP28 protein of the White Spot Syndrome Virus [56]. Therefore, information on the sequences and expression of the *C. maenas* endocytic system may aid in the study of viral infection. The mechanisms of endocytosis, maturation of endosomes and related signalling molecules are depicted in the KEGG pathway shown in Fig. 6. The number of identified components demonstrates that *C. maenas* contains an endocytic system that closely resembles this canonical KEGG pathway. The KEGG annotation did not yield transcripts similar to caveolin, an important constituent of caveolar-mediated endocytosis. However a tBLASTn search of NCBI caveolin protein sequences in the transcriptome identified similarity between 'comp141181_c0_seq1' and caveolin-3-like isoform *X*2 (XP_006615923.1, *Apis dorsata*, e = 1e−15). Expression of components of the endocytosis pathway is visualized in Fig. 7, and most of these components were expressed across all tissues. The muscle tissue showed an endocytosis expression profile that differs from the other tissues.

Conclusions

We produced an assembled transcriptome for *C. maenas* that consists of 153,699 loci and 212,427 transcripts and provides a significant molecular resource for wide studies into both basic and applied biology for this species. Comparisons run in the NCBI-nr database showed 30 % of *C. maenas* transcripts had significant homology against known sequences, but a large number were novel transcripts that have yet to be characterized. Expression analysis revealed tissues and organ transcript specificity that mapped with gene ontology annotations relating to specific tissue/organ-related functions. Of particular relevance for studies into pathogenesis and disease, we identified the presence of a series of known targets and functional pathways including the RNAi pathway, Toll-like receptor signalling, IMD and JAK-STAT pathways that form part of their innate immune system.

Methods

mRNA preparation

Four individual *Carcinus maenas* were collected from Newton's Cove, Weymouth, UK and placed on ice prior to dissecting tissues and organs of interest (including gill, hepatopancreas, epidermis, eyes, intestine, haemolymph, muscle, heart, nerve, ovary, testis and eggs). All tissues and organs were immediately snap-frozen in

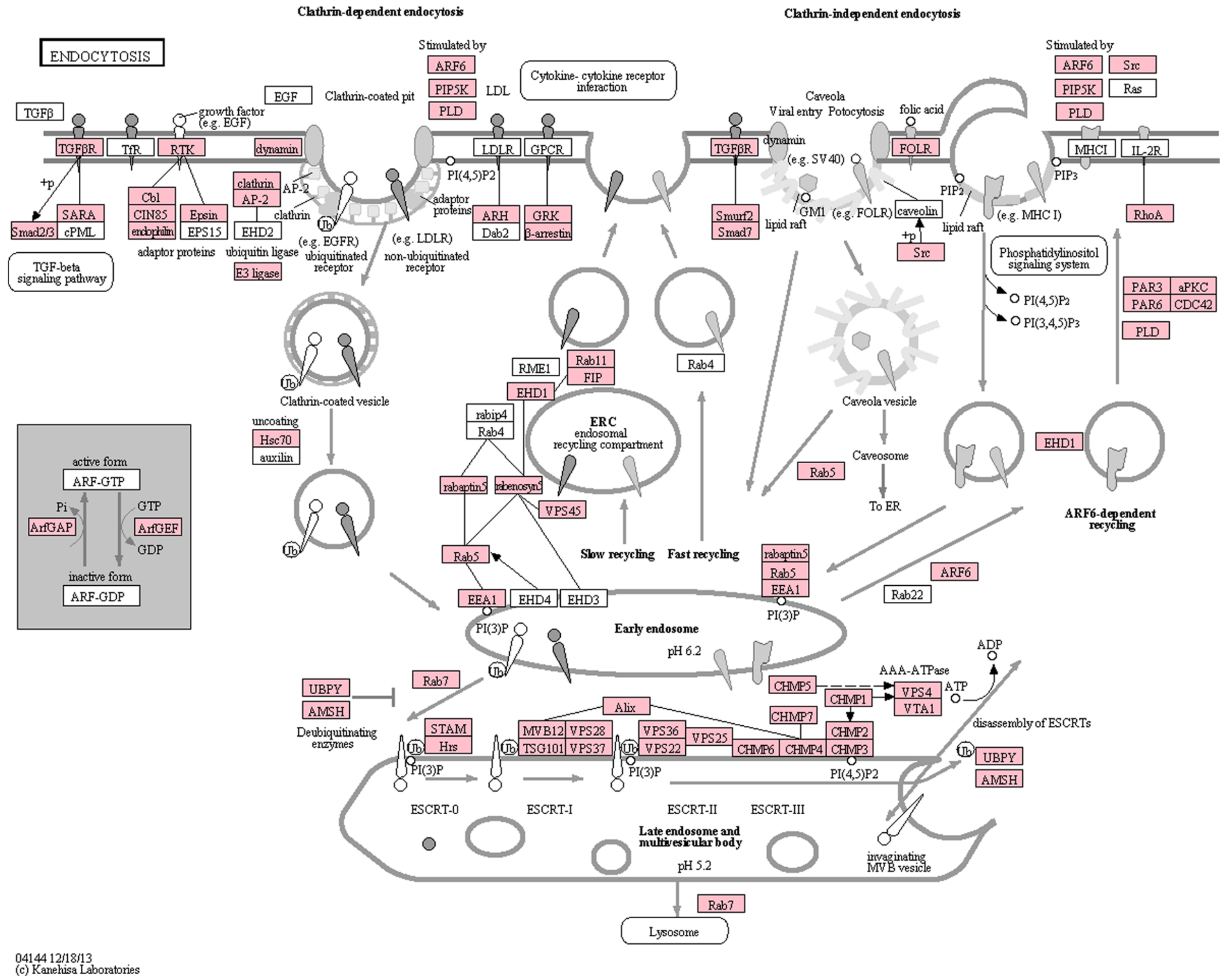


Fig. 6 Endocytosis pathway. Overview of the KEGG endocytosis pathway. Components with homologues in the *Carcinus maenas* transcriptome are highlighted in pink

liquid nitrogen and transported to the University of Exeter for sample preparation and analysis.

RNA was extracted using Qiagen's miRNeasy mini kit, with on column DNase digestion, according to the manufacturer's instructions. RNA quality was measured using an Agilent 2100 Bioanalyzer with RNA 6000 nano kit (Agilent Technologies, CA, USA). cDNA libraries for each tissue were constructed using 2.5 μg of RNA pooled from the four sampled individuals. ERCC Spike-In control mixes (Ambion via Life Technologies, Paisley, UK) were added to control for technical variation during sample preparation and sequencing, and analysed using manufacturer's guidelines. mRNA purification was performed via poly (A) enrichment using Tru-Seq Low Throughput protocol and reagents (Illumina, CA, USA). Finally, cDNA libraries were constructed using Epicentre's ScriptSeq v2 RNA-seq library preparation kit (Illumina). Each tissue was labelled with a unique barcode sequence to enable multiplexing of all samples across one lane whilst ensuring sequencing data from each tissue could be separated for analysis. Sequencing was performed on an Illumina HiSeq 2500 with the 2 × 100 bp paired-end read module.

Transcriptome assembly

Prior to transcriptome assembly, the sequence reads were processed to remove those with low confidence (as assigned by the sequencer). The first 12 bp were trimmed from the reads to remove bias caused by random hexamer priming [57] and Illumina adapters were removed using Trimmomatic [58]. Trimmomatic was also used for quality trimming of the 3' end of the reads using a sliding window (4 bp with a minimal Phred quality of 30). Reads shorter that 70 bp were discarded. Only read pairs where both reads passed the desired quality threshold were retained. Read pairs of all tissue libraries were pooled and used for *de novo* transcriptome assembly using the Trinity

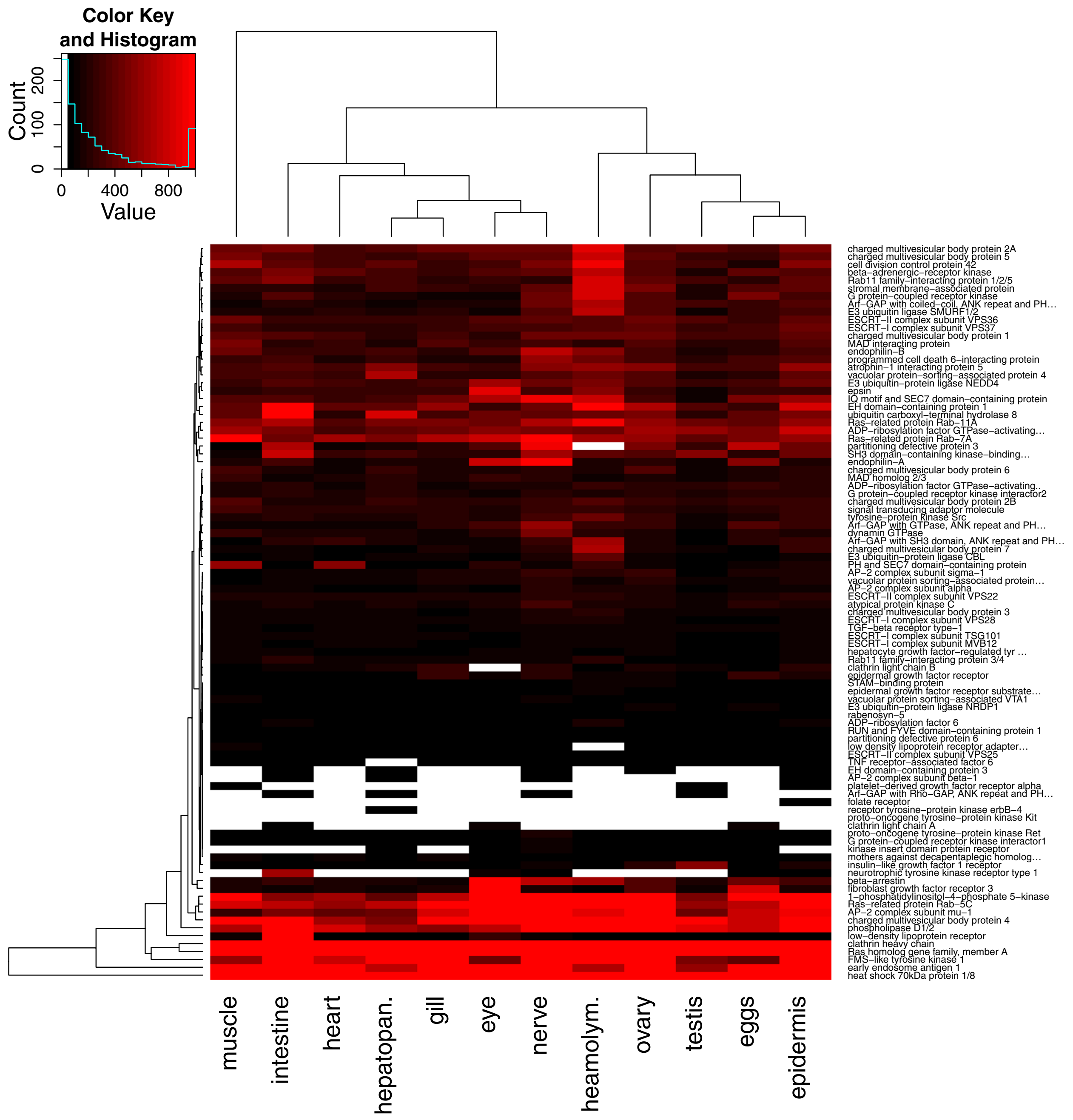


Fig. 7 Endocytosis pathway expression. Expression of endocytosis pathway components in twelve *Carcinus maenas* tissues. The expression values are in FPKM, values of 0 are coloured white and values over 1000 FPKM are binned together

(2013-02-25 release) software package [9]. Transcripts with a length of 200 nucleotides or less were removed from the assembly. General transcriptome statistics, including maximal transcript length, mean transcript length and N50, of the resulting transcriptome were calculated with a custom R script. This Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GBXE00000000. The version described in this paper is the first version, GBXE01000000.

Transcriptome characterization

The Trinotate suite (2013-08-26 release) [59] was used to annotate transcripts. Peptide coding regions were found through transdecoder and BLASTp v 2.2.28 (release 2013-07, e-value cutoff of 1e-5) was used to find sequence homology to UniProt/SwissProt. HMMR 3.1.b1 [60] and the Pfam database (version 27.0) were used to identify conserved protein domains. Additionally transmembrane regions were predicted with TMHMM-2.0c [61] and potential signal peptides identified with SignalP 4.1 [62]. Furthermore, homology searches were performed using BLASTx v 2.2.28 against the NCBI non-redundant (nr) protein database with an e-value cutoff of 1e-3 and BLASTn against all available *C. maenas* ESTs in the NCBI database (2013-05-10; 15,558 ESTs in total), with an e-value cutoff of 1e-3 and retaining the best 20 hits. The presence of highly conserved core eukaryotic genes was assessed using CEGMA 2.5 [29, 63]. Functional annotation analysis was conducted by assigning Molecular Function, Biological Process and Cellular Component Gene Ontology annotations to transcripts with BLAST2GO (v2.7.0) [30]. Finally, taxonomic classifications of the transcripts were determined and visualized using MEGAN 4 [64], and transcripts that did not map to the metazoan taxon were removed from the transcriptome assembly.

Differential gene expression analysis

For each tissue, reads were mapped to the *Carcinus* transcriptome (including non-metazoan transcripts) using bowtie2 [65] and RSEM [32] to obtain overall transcript expression values. Differential transcript expression was performed by comparing each tissue to the other eleven tissues, treating the latter as biological replicates. The calculations were performed with RSEM based on the edgeR package [66] with a dispersion parameter of 0.4 which is recommended for analysis without replicates. Transcripts with an FDR < 0.01 were treated as differentially expressed. The lists of differentially expressed genes for each tissue were analysed for enrichment of Gene Ontology categories using BLAST2GO, and terms were deemed significant when FDR < 0.05.

Pathway analysis

KEGG ontology groups were assigned to assembled transcripts through the KEGG Automatic Annotation Server (KAAS) web service [33]. Next, the presence of components of reference pathways related to immune responses, including the toll-like receptor signalling pathway (map04620), TNF signalling pathway (map04668), JAK-STAT signalling pathway (map04630) and the endocytosis pathway (map04144) were visualized through the KAAS web service [33].

Since KEGG is focused on vertebrate pathways an additional, more flexible, pathway annotation strategy was required. For identification of a pathway component (e.g. Spätzle in the invertebrate Toll signalling pathway) the following steps were followed: 1. Protein sequences for the component were downloaded from the NCBI protein database based on a search query. 2. These sequences were used as input in a tBLASTn search against the assembled transcriptome (cut-off 1e-20). 3. For every transcript with BLAST hits, a filter was applied to select the best three query sequences based on first taxonomic distance to a reference taxon (tax_id = 6759, *Carcinus maenas*) and secondly the e-value. 4. When necessary, manual filtering to remove irrelevant sequences that were returned from NCBI. An R-script that performs this analysis is supplied in Additional file 9.

Expression of pathway components was derived by adding the RSEM-derived FPKM values for transcripts that were annotated to the component (either through KEGG annotation or the annotation stratagem explained above).

Availability of supporting data

The data set supporting the results of this article is available in the genbank Transcriptome Shotgun Assembly Sequence Database repository (http://www.ncbi.nlm.nih.gov/genbank/tsa) under the accession GBXE00000000. The version described in this paper is the first version, GBXE01000000.

Additional files

Additional file 1: Cmaenas_transcript_lengths.pdf. This file contains a histogram of transcript lengths to illustrate the presence of fragments and full length transcripts in the transcriptome.

Additional file 2: Cmaenas_trinotate_annotation_report.txt. Output of the Trinotate annotation pipeline, tabular format. This file contains annotation information derived from the Trinotate annotation pipeline as decribed in the Methods section.

Additional file 3: Cmaenas_NCBIblastx.txt. Blastx results of transcripts to NCBI nr database, tabular format. This file contains information on sequence similarity between transcripts in the transcriptome and sequences in the NCBI non-redundant database.

Additional file 4: Cmaenas_transcriptome_GO_annot.txt. Transcript Gene Ontology annotation, tabular format. This file contains Gene Ontology annotations for transcripts.

Additional file 5: Tissue_GO_Enrichment.xlsx. Enriched Gene Ontology terms for analyzed tissues. This file shows which Gene Ontology terms are enriched for tissue specific differentially expressed genes.

Additional file 6: Pathway_components.xlsx. This file contains sequence similarities between components of immune pathways and the transcriptome.

Additional file 7: toll_pathway_heatmap.pdf. Heatmap of expression values for components of the Toll-like signalling pathway.

Additional file 8: imd_pathway_heatmap.pdf. Heatmap of expression values for components of the IMD signalling pathway.

Additional file 9: Pathway_annotation.R. R script used to identify transcripts with significant sequence similarity to genes/proteins of interest.

Abbreviations

WSSV: White spot syndrome virus; GO: Gene ontology; FDR: False discovery rate; IMD: Immune deficiency; MAPK: Mitogen activated protein kinase; PRR: Pattern recognition receptors; PAMP: Pathogen associated molecular patterns; GNBP: Gram-negative binding proteins; PGRP: Peptidoglycan recognition proteins; TLR: Toll like receptor; FPKM: Fragments per kilo bases of exons per million mapped reads; ALF: Anti-lipopolysaccharide factor; PO: Phenol oxidase; RNAi: RNA interference.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

BVE conducted the bioinformatics analysis and wrote the first version of the manuscript. LKB conducted the sample collection from wild crabs (together with KSB), RNA extraction and library construction for the sequencing experiments. EMS, CRT, GDS and RvA designed the study and supervised the laboratory experiments and data analysis. All authors contributed towards the design of the study and the final version of the manuscript.

Acknowledgements

The authors thank Audrey Farbos, Karen Moore and Konrad Paszkiewicz for facilitating the sequencing experiments. This work was funded by the Cefas Seedcorn Contract #DP318 (to GDS) and the University of Exeter's Open Innovation Platform (to CRT, EMS and RvA). The Exeter Sequencing Facility was funded by a Wellcome Trust Institutional Strategic Support Award (WT097835MF).

Author details

[1]Biosciences, College of Life & Environmental Sciences, University of Exeter, Geoffrey Pope Building, Exeter EX4 4QD, UK. [2]European Union Reference Laboratory for Crustacean Diseases, Centre for Environment, Fisheries and Aquaculture Science (Cefas), Weymouth Laboratory, Weymouth, Dorset DT4 8UB, UK. [3]Aquatic Health and Hygiene Division, Centre for Environment, Fisheries and Aquaculture Science (Cefas), Weymouth Laboratory, Weymouth, Dorset DT4 8UB, UK.

Received: 2 December 2014 Accepted: 29 May 2015
Published online: 16 June 2015

References

1. Ge X, Chen H, Wang H, Shi A, Liu K. *De novo* assembly and annotation of *Salvia splendens* transcriptome using the illumina platform. PLoS One. 2014;9(3):e87693.
2. Uren Webster TM, Bury N, van Aerle R, Santos EM. Global transcriptome profiling reveals molecular mechanisms of metal tolerance in a chronically exposed wild population of brown trout. Environ Sci Technol. 2013;47(15):8869–77.
3. Gallardo-Escárate C, Valenzuela-Muñoz V, Nuñez-Acuña G. RNA-Seq analysis using *de novo* transcriptome assembly as a reference for the salmon louse *caligus rogercresseyi*. PLoS One. 2014;9(4):e92239. doi:10.1371/journal.pone.0092239.
4. Yang W-J, Yuan G-R, Cong L, Xie Y-F, Wang J-J. *De novo* cloning and annotation of genes associated with immunity, detoxification and energy metabolism from the fat body of the oriental fruit fly, *Bactrocera dorsalis*. PloS One. 2014;9(4):e94470.
5. Zimmer CT, Maiwald F, Schorn C, Bass C, Ott MC, Nauen R. A *de novo* transcriptome of European pollen beetle populations and its analysis, with special reference to insecticide action and resistance. Insect Mol Biol. 2014;23:511–26.
6. Shentu X-P, Liu W-P, Zhan X-H, Xu Y-P, Xu J-F, Yu X-P, et al. Transcriptome sequencing and gene expression analysis of *Trichoderma brevicompactum* under different culture conditions. PLoS One. 2014;9(4):e94203.
7. Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, et al. SOAPdenovo-trans: *de novo* transcriptome assembly with short RNA-Seq reads. Bioinformatics. 2014;30(12):1660–6.
8. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;28(8):1086–92.
9. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–52.
10. EMBL-EBI. Ensembl Metazoa. EMB-EBI. 2014. http://metazoa.ensembl.org/info/website/species.html. 2014.
11. Global invasive species database. http://www.issg.org/database/welcome/. Accessed 2013.
12. Darling JA, Bagley MJ, Roman J, Tepolt CK, Geller JB. Genetic patterns across multiple introductions of the globally invasive crab genus Carcinus. Mol Ecol. 2008;17(23):4992–5007.
13. Perry H. *Carcinus maenas*. USGS nonindigenous aquatic species database. 2014.
14. Hänfling B, Edwards F, Gherardi F. Invasive alien Crustacea: dispersal, establishment, impact and control. BioControl. 2011;56(4):573–95.
15. Jebali J, Chicano-Galvez E, Fernandez-Cisnal R, Banni M, Chouba L, Boussetta H, et al. Proteomic analysis in caged Mediterranean crab (*Carcinus maenas*) and chemical contaminant exposure in Teboulba Harbour, Tunisia. Ecotoxicol Environ Saf. 2014;100:15–26.
16. Klassen L. A biological synopsis of the European green crab, *Carcinus maenas*. Can Manuscr Rep Fish Aquat Sci. 2007;2818:vii. +75pp.
17. Ben-Khedher S, Jebali J, Houas Z, Naweli H, Jrad A, Banni M, et al. Metals bioaccumulation and histopathological biomarkers in *Carcinus maenas* crab from Bizerta lagoon, Tunisia. Environ Sci Pollut Res Int. 2014 Mar;21(6):4343-57
18. Elumalai M, Antunes C, Guilhermino L. Enzymatic biomarkers in the crab *Carcinus maenas* from the Minho River estuary (NW Portugal) exposed to zinc and mercury. Chemosphere. 2007;66(7):1249–55.
19. Ghedira J, Jebali J, Banni M, Chouba L, Boussetta H, López-Barea J, et al. Use of oxidative stress biomarkers in *Carcinus maenas* to assess littoral zone contamination in Tunisia. Aquat Biol. 2011;14(1):87–98.
20. Chen CY, Dionne M, Mayes BM, Ward DM, Sturup S, Jackson BP. Mercury bioavailability and bioaccumulation in estuarine food webs in the Gulf of Maine. Environ Sci Technol. 2009;43(6):1804–10.
21. Rainbow PS, Black WH. Cadmium, zinc and the uptake of calcium by two crabs, *Carcinus maenas* and *Eriocheir sinensis*. Aquat Toxicol. 2005;72(1–2):45–65.
22. Pedersen KL, Bach LT, Bjerregaard P. Amount and metal composition of midgut gland metallothionein in shore crabs (*Carcinus maenas*) after exposure to cadmium in the food. Aquat Toxicol. 2014;150:182–8.
23. Windeatt KM, Handy RD. Effect of nanomaterials on the compound action potential of the shore crab, *Carcinus maenas*. Nanotoxicology. 2013;7(4):378–88.
24. Watts AJ, Lewis C, Goodhead RM, Beckett SJ, Moger J, Tyler CR, et al. Uptake and retention of microplastics by the shore crab *Carcinus maenas*. Environ Sci Technol. 2014;48(15):8823–30.
25. Stentiford GD, Bonami JR, Alday-Sanz V. A critical review of susceptibility of crustaceans to taura syndrome, yellowhead disease and white spot disease and implications of inclusion of these diseases in European legislation. Aquaculture. 2009;291(1–2):1–17.
26. Stentiford GD, Neil DM, Peeler EJ, Shields JD, Small HJ, Flegel TW, et al. Disease will limit future food supply from the global crustacean fishery and aquaculture sectors. J Invertebr Pathol. 2012;110(2):141–57.
27. Bateman KS, Tew I, French C, Hicks RJ, Martin P, Munro J, et al. Susceptibility to infection and pathogenicity of White Spot Disease (WSD) in non-model crustacean host taxa from temperate regions. J Invertebr Pathol. 2012;110(3):340–51.
28. NCBI taxonomy *Carcinus maenas*. NCBI. 2014. http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=6759.
29. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23(9):1061–7.
30. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21(18):3674–6.
31. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;12:87.
32. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinf. 2011;12:323.
33. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 2007;35(Web Server issue):W182–5.
34. Li X, Cui Z, Liu Y, Song C, Shi G. Transcriptome analysis and discovery of genes involved in immune pathways from hepatopancreas of microbial challenged mitten crab *Eriocheir sinensis*. PLoS One. 2013;8(7):e68233.
35. Mercer J, Schelhaas M, Helenius A. Virus entry by endocytosis. Annu Rev Biochem. 2010;79:803–33.
36. Christophides GK, Vlachou D, Kafatos FC. Comparative and functional genomics of the innate immune system in the malaria vector *Anopheles gambiae*. Immunol Rev. 2004;198(1):127–48.
37. McTaggart SJ, Conlon C, Colbourne JK, Blaxter ML, Little TJ. The components of the *Daphnia pulex* immune system as revealed by complete genome sequencing. BMC Genomics. 2009;10:175.
38. Liu H, Wu C, Matsuda Y, Kawabata S, Lee BL, Soderhall K, et al. Peptidoglycan activation of the proPO-system without a peptidoglycan receptor protein (PGRP)? Dev Comp Immunol. 2011;35(1):51–61.

39. Armitage SA, Peuss R, Kurtz J. Dscam and pancrustacean immune memory—a review of the evidence. Dev Comp Immunol. 2015 Feb;48(2):315-23.
40. Ng TH, Chiang YA, Yeh YC, Wang HC. Review of Dscam-mediated immunity in shrimp and other arthropods. Dev Comp Immunol. 2014;46(2):129–38.
41. Kingsolver MB, Huang Z, Hardy RW. Insect antiviral innate immunity: pathways, effectors, and connections. J Mol Biol. 2013;425(24):4921–36.
42. Kawai T, Akira S. The role of pattern-recognition receptors in innate immunity: update on toll-like receptors. Nat Immunol. 2010;11(5):373–84.
43. Chen WY, Ho KC, Leu JH, Liu KF, Wang HC, Kou GH, et al. WSSV infection activates STAT in shrimp. Dev Comp Immunol. 2008;32(10):1142–50.
44. Morin-Poulard I, Vincent A, Crozatier M. The JAK-STAT pathway in blood cell formation and immunity. JAKSTAT. 2013;2(3):e25700.
45. Tassanakajon A, Somboonwiwat K, Supungul P, Tang S. Discovery of immune molecules and their crucial functions in shrimp immunity. Fish Shellfish Immunol. 2013;34(4):954–67.
46. Coleman JW. Nitric oxide in immunity and inflammation. Int Immunopharmacol. 2001;1(8):1397–406.
47. Destoumieux D, Bulet P, Loew D, Van Dorsselaer A, Rodriguez J, Bachere E. Penaeidins, a new family of antimicrobial peptides isolated from the shrimp *Penaeus vannamei* (Decapoda). J Biol Chem. 1997;272(45):28398–406.
48. Rolland JL, Abdelouahab M, Dupont J, Lefevre F, Bachere E, Romestand B. Stylicins, a new family of antimicrobial peptides from the pacific blue shrimp *Litopenaeus stylirostris*. Mol Immunol. 2010;47(6):1269–77.
49. Tang H. Regulation and function of the melanization reaction in *Drosophila*. Fly. 2009;3(1):105–11.
50. Tang H, Kambris Z, Lemaitre B, Hashimoto C. Two proteases defining a melanization cascade in the immune system of *Drosophila*. J Biol Chem. 2006;281(38):28097–104.
51. Liu CH, Cheng W, Chen JC. The peroxinectin of white shrimp *Litopenaeus vannamei* is synthesised in the semi-granular and granular cells, and its transcription is up-regulated with *Vibrio alginolyticus* infection. Fish Shellfish Immunol. 2005;18(5):431–44.
52. Wang PH, Huang T, Zhang XB, He JG. Antiviral defense in shrimp: from innate immunity to viral infection. Antiviral Res. 2014;108:129–41. doi:10.1016/j.antiviral.2014.05.013.
53. La Fauce K, Owens L. RNA interference with special reference to combating viruses of crustacea. Indian J Virol. 2012;23(2):226–43.
54. Lee YS, Nakahara K, Pham JW, Kim K, He Z, Sontheimer EJ, et al. Distinct roles for *Drosophila* dicer-1 and dicer-2 in the siRNA/miRNA silencing pathways. Cell. 2004;117(1):69–81.
55. Bernstein E, Caudy AA, Hammond SM, Hannon GJ. Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature. 2001;409(6818):363–6.
56. Verma AK, Gupta S, Verma S, Mishra A, Nagpure NS, Singh SP, et al. Interaction between shrimp and white spot syndrome virus through PmRab7-VP28 complex: an insight using simulation and docking studies. J Mol Model. 2013;19(3):1285–94.
57. Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. 2010;38(12):e131.
58. Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, Stitt M, et al. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Res. 2012;40(Web Server issue):W622–7.
59. Trinotate. http://trinotate.github.io/.
60. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011;39(Web Server issue):W29–37.
61. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. J Mol Biol. 2001;305(3):567–80.
62. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8(10):785–6.
63. Parra G, Bradnam K, Ning Z, Keane T, Korf I. Assessing the gene space in draft genomes. Nucleic Acids Res. 2009;37(1):289–97.
64. Huson DH, Mitra S, Ruscheweyh HJ, Weber N, Schuster SC. Integrative analysis of environmental sequences using MEGAN4. Genome Res. 2011;21(9):1552–60.
65. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
66. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.

Chapter 4

De novo Assembly of the European shore crab (*Carcinus maenas*) genome.

Supplementary material available in the ‘Chapter 4’ folder on the DVD

Chapter 4: De novo Assembly of the European shore crab (*Carcinus maenas*) genome

4.1 Abstract

The European shore crab (*Carcinus maenas*) is an invasive species that has impacted on ecosystems across the many regions of the globe. *C. maenas* has caused millions of dollars of damage to some wild fisheries through predation and extirpation of other crustacean populations. Its success as an invasive species is, in part, due to its resilience and ability to adapt to different environmental conditions. These characteristics, together with the ubiquitous range of this species in temperate costal habitats, have played a key part in *C. maenas* being adopted widely in scientific experiments including in biomonitoring, ecotoxicology and host-pathogen interactions. We aimed to generate a genomics resource to facilitate future studies using *C. maenas* as a model organism. We sequenced and annotated a *de novo* draft genome assembly for this species using an Illumina HighSeq 2500 platform. We isolated DNA from one individual *C. maenas* and constructed paired-end and mate-pair libraries for sequencing. Approximately 470 million pairs of 150 bp sequencing reads were generated, corresponding to 70x estimated coverage of the crab genome. A *de novo* draft genome assembly consisting of 338,980 scaffolds and covering 362 Mb (36 % of estimated genome size) was produced, using SOAP-denovo2 coupled with the BESST scaffolding system. The assemblies generated were highly fragmented due to the presence of repetitive areas in the *C. maenas* genome. Using a combination of *ab initio* predictors, previous RNA-sequencing data and curated *C. maenas* sequences we produced a model encompassing 10,355 genes. Furthermore, we identified 185 miRNA precursors of which 31 showed sequence similarity to known miRNAs, and 154 were potentially novel miRNAs. We investigated whether the *C. maenas* genome contained endogenous viral elements but found no evidence supporting such integration events. Within the *C. maenas* genome we showed the presence and subsequently identified the introns/exons of *Dscam*, a gene that has recently been shown to play an important role in the invertebrate immune system. The assembled draft genome and accompanying gene model generated in this project are a valuable molecular resource for studies involving *C. maenas*.

4.2 Introduction

The European green crab or shore crab, *Carcinus maenas*, is a marine invertebrate that commonly lives in marine or estuarine habitats. It is an important invasive species that has spread from its native range in Europe and northern Africa to Australia, South Africa and the Americas. Furthermore, it is the only crustacean listed on the global index of invasive species [1]. At invasion sites it can cause significant damage through predation, targeting bivalves in particular, and extirpation of local populations of crustaceans. Economic losses due to *C. maenas* invasions can be very significant [2]. Lafferty and Kuris estimated potential losses to green crab predation in the USA to be around $44 million per year and the Oregon Dungeness Crab Commission defined the potential impact of *C. maenas* on Dungeness crab population in 2005 at $50 million [2-4]. The success of *C. maenas* as an invasive species can be attributed to its resilience, both to changes in the environment and to pathogens. These characteristics enable this species to become widespread in coastal areas across the globe and make it an interesting model species to address a wide range of research questions including adaptation, ecotoxicology and disease resistance.

C. maenas has become a popular experimental test organism, particularly in the field of aquatic ecotoxicology [5]. The species has a high tolerance to environmental conditions (air exposure, temperature, salinity, starvation) [2] and has reasonable tolerance to environmental contaminants. The crab lives in estuarine and coastal habitats that are often polluted by a wide variety of chemicals derived from different sources (urban, industrial, agricultural and maritime) [5, 6]. The species has been used in a large number of studies investigating the effects of metals, endocrine disruptors, pesticides and other pollutants [5]. These studies aid in understanding such effects in invertebrates, other than the popular *Daphnia* species. In addition the shore crab is easy to maintain in the lab because of its robustness and size.

A major reason for the success of *C. maenas* as an invasive species is its resilience to infections by viruses, bacteria and parasites. In its natural habitat, the shore crab carries a large pathogen load but when moved into novel habitats the load reduces, thus facilitating settlement in these new environments. In addition pathogens in the new environment are more specialized towards species native to that environment. This process, the Enemy Release Hypothesis, has been described for many invasive species [7]. When comparing infections within aquatic Crustacea the shore crab appears to be amongst the most resilient [2] to for example *Hematodinium* [8] and to White Spot Syndrome Virus (WSSV) [2, 9]. Of particular interest to this thesis, WSSV is one of the most devastating viruses to the global aquaculture industry [10], and has caused enormous damage to global crustacean aquaculture, most notably to shrimp farming where it has resulted in a loss of 10 % of global production [11]. Despite intensive research, the successful development of an effective treatment for WSSV still remains elusive. The mechanisms of resistance to WSSV in *C. maenas* may offer the opportunity to identify disease treatments opportunities in the more susceptible and economically important crustacean species.

Rodrigues and Pardal (2014) have suggested that a genome sequencing program is essential for the development of *C. maenas* as a model species in cutting edge ecotoxicology research [5]. Through knowledge of its genome sequence, the interpretation of current results and the design of future research studies will be greatly facilitated. Until recently the genomic resources for *C. maenas* were limited to a few hundred protein sequences and around 15,000 expressed sequence tag sequences. In a previous study, we expanded these resources by generating a *de novo* transcriptome wherein we show the presence of many of the genes that form part of pathways connected to the innate immune system [12]. However an assembled genome is still lacking. Such a genome will open further avenues for research into pathogen resistance, including the role of miRNAs and viral inserts in a host genome that convey resistance.

Here, we employed next generation genome sequencing technology to produce a draft genome assembly for *C. maenas*. Within the assembly we identify likely locations of genes and their exons and summarize them in a gene model. Other structures like repeat regions and potential viral inserts were also investigated. In order to facilitate the use of this genome assembly as a resource to study host-pathogen interactions we focused on mechanisms of response to disease. It is well documented that within invertebrates the RNA interference (RNAi) is an important component of the immune response, especially against viruses [13]. RNAi can be used to regulate host responses to pathogens or directly inhibit expression of viral transcripts [11]. In order to facilitate studies focused on miRNAs in the context of responses to pathogen infection, we identified miRNA precursors present in the *C. maenas* genome. In addition, given the recent discoveries associated with Down syndrome cell adhesion molecule (Dscam) which is hypothesized to play an important role in invertebrate "immune memory,[14], we made a gene model for Dscam in order to investigate its potential for hyper-variability which is instrumental to its role in immune memory.

4.3 Materials and Methods

4.3.1 DNA isolation and sequencing

Crabs were placed on ice prior to dissecting muscle tissue, which was immediately snap-frozen in liquid nitrogen until DNA extraction and sequencing analysis. Foozen muscle tissue fragments were grounded with a mortar and pestle in liquid nitrogen. DNA was extracted using the Qiagen's Blood and Tissue kit, with on column RNase A digestion, according to the manufacturer's instructions with the following modifications: after incubation with proteinase K, the lysate was transferred to a new microcentrifuge tube to perform RNase digestion, and following 15 minutes incubation at room temperature, DNA was eluted with 100µL Tris-HCl (pH 9) at 37°C. DNA quality was assessed by gel electrophoresis (on a 1.5% agarose gel) and its quantity assessed using the Qubit dsDNA BR Assay kit (ThermoFisher Scientific, UK). Paired-end DNA libraries were constructed using the NEBNext® library preparation kits (Illumina) according to the manufacturers' protocol. The insert sizes for the paired-end libraries were: 300-500 bp and 500-700 bp. Mate pair libraries were constructed with the gel-free Nextera Mate Pair library preparation kit (Illumina) according to the protocol. Mate pair libraries size selected for insertsizes of 3500-5624 bp and 7361-11741 bp as measured on a Bioanalyzer. The paired-end and mate pair libraries were quantified and multiplexed in two lanes on an Illumina HiSeq 2500 (Exeter sequencing service).

4.3.2 Data pre-processing

The quality of the sequencing data was assessed with FastQC v 0.10.1 [15]. After this assessment, the reads were pre-processed in order to remove any remaining adapter sequences and reduce low confidence information. The paired-end library sequences were trimmed using Trimmomatic v 0.32 [16] to remove Illumina TruSeq adapters, repeat reads, the first 12 bases and the last 3 bases of each sequence, the low confidence bases using a sliding window of 4 and minimal Phred quality of 20 and the reads with a length below 25. The adapters in the mate-pair library sequences were removed by NextClip v 1.3 [17] (to obtain a minimal read length of 25, remove PCR duplicates and retain only read pairs where the adapter was identified in at least one of the reads). After this initial filtering the mate pairs were quality-trimmed by Trimmomatic v 0.32 [16] using the same settings as described above for the paired-end libraries. Finally, the quality of the data was re-assessed with FastQC v 0.10.1.

4.3.3 De novo genome scaffold assembly and quality assessment

The assembly of the reads from paired-end and mate-pair libraries into genome scaffolds was optimized for contig quantity and length distribution by using various combinations of *de novo* genome assemblers, scaffolders and parameter settings. The following assemblers were used: ALLPATHS v 51875 [18, 19], Platanus v 1.2.1 [20], SPAdes v 3.6.1 [21] and SOAP-denovo2 v 2.04-r240 [22], the latter being combined with BESST v 1.3.7 scaffolder [23]). Assembly statistics were calculated in QUAST v 2.3 [24]. Based on these statistics the SOAP-BESST and ALLPATHS assemblies produced the highest quality/largest scaffolds and were selected for further analysis. These assemblies were subsequently assessed for their coverage of *Carcinus maenas* transcriptome reads. *C. maenas* RNA-sequencing reads from Verbruggen et al. 2015 [12](SRA accession: SRX691208, SRX698361, SRX710630, SRX713887, SRX732924, SRX744550, SRX744554, SRX744557, SRX744558, SRX744559, SRX744561, SRX744562) were aligned to the genome scaffolds using bowtie2 v 2.2.6 (default settings) [25]. The presence of highly conserved genes in the *de novo* genome scaffolds were assessed by CEGMA v. 2.5 [26]. In addition, BUSCO v 1.1b [27] was used to test assembly quality by investigating the presence or absence of a set of single copy orthologues.

4.4.4 Genome scaffold annotation

Following analysis of quality criteria, the SOAP-denovo2 + BESST scaffolder assembly (> 500 bp) was chosen as the highest quality assembly and selected for further downstream analyses. Simple repeats were identified using RepeatMasker version 4.0.5 [28]. Arthropod repeats were downloaded from RepBase (Oct 2015) and BLASTN 2.2.28+ (e-value < $1e^{-10}$, minimal 75 % sequence identity) was used to locate retro-elements in the *de novo* genome scaffold. Protein sequences from the Malacostraca class were downloaded from GenBank and used as guide for gene model prediction. These sequences were aligned to the *de novo* genome scaffold with TBLASTN 2.2.28+ (e-value < $1e^{-10}$). Significant combinations of protein and scaffold sequences were used as input for GeneWise 2-4-1 to derive introns and exons [29]. *Ab initio* gene prediction was performed with GeneMark-ES 3.4.8 [30]. Assembled transcripts from Verbruggen et al. 2015 [12] were aligned to the genome scaffold in PASA 2.0.2 [31] and used to model gene structures. Further gene structures were derived by AUGUSTUS 2.7 [32], the prediction process was informed by Illumina RNA-sequencing reads from Verbruggen et al. 2015 [12] according to the protocol described in [33]. The predictions from GeneWise, GeneMark-ES, PASA and AUGUSTUS were combined by EVidenceModeler v 1.1.1 [34].

The presence of miRNA precursors was investigated using the miRCandRef [35] pipeline using the paired-end DNA sequencing reads with an insert size of 300-500 bp and small RNA sequencing data from a pilot experiment conducted at Exeter/Cefas. The pilot experiment included data from *C. maenas* individuals injected with either saline solution or White Spot Syndrome Virus (unpublished). The pipeline was executed on the Genomic HyperBrowser server [36]. Identified miRNAs were annotated through a BLAST search optimized for small sequences, as suggested by the miRCandRef pipeline [35]. The potential mRNA targets of miRNAs were predicted *in silico* through miRanda v 3.3a [37] and MicroTar v 0.9.6 [38], based on the *C. maenas* transcriptome sequences identified in Verbruggen et al. 2015 [12]. Predictions were filtered based on a free energy threshold of -10 kcal/mol.

4.4.5 Analysis of viral inserts

Integration viral sequences were assessed both at the nucleotide and protein level. Viral genomic sequences were downloaded from RefSeq (release 72) [39] and the pre-processed paired-end reads were aligned to these sequences using bowtie2 [25], not allowing for discordant mapping. From the resulting SAM/BAM-files the coverage of the viral genomes were computed in BEDTools genomeCoverageBed v 2.17.0 [40]. The sequencing reads mapping with a high mapping quality (MAPQ > 40) to the viral genomes were manually investigated for likelihood of a genuine insertion in contradiction to artefacts such as repeat regions. Furthermore, all viral protein sequences were downloaded from RefSeq (release 72) [39]. Sequence similarity between pre-processed DNA sequencing reads and viral proteins was computed with DIAMOND [41] (BLASTx, e-value < $1e^{-10}$, percent identity >= 90 %) and successful alignments were again quality assessed to avoid false positives. To investigate the *C. maenas* genome scaffolds for the presence of potential WSSV sequences specifically, the pre-processed paired-end sequencing reads were aligned to the genome of the Chinese WSSV isolate (AF332093) using bowtie2 2.1.0 [25] (default and local alignment) and bwa 0.7.12 [42]. Successful alignments were investigated in the genome browser IGV [43].

4.4.6 Gene model for Dscam

Identification of scaffolds that might contain a *Dscam* gene was performed using BLASTN 2.2.28+ and using *C. maenas Dscam* (GenBank: HG964670.1) as query and an e-value threshold of $1e^{-10}$. Once potential scaffolds were identified, Dscam protein sequences were aligned using GeneWise 2-4-1 [29]. *Dscam* sequences included: *C. maenas Dscam* (GenBank: CDO91660.1), *E. sinensis Dscam* (GenBank: AGL39311.1) and *D. melanogaster Dscam* (GenBank: AAF71926.1). Alignments between scaffolds and *Dscam* sequences were produced using BLAT v 36 [44] and visualised using IGV 2.3 [43]. GeneMark-HMM v 3.4.8 predictions for scaffolds were obtained from the scaffold annotations step. Pfam domain analysis [45] was performed on both protein sequences obtained from GeneWise alignments and nucleotide sequence derived from GeneMark-HMM predictions.

4.5 Results and Discussion

4.5.1 Draft Genome Assembly - Quality Control

Two paired-end libraries with insert sizes 300-500 and 500-700 bp and two mate pair libraries with insert sizes 3500-5624 bp and 7361-11741 bp were created from *C. maenas* DNA, and a total of 682 million paired-end and 258 million mate-pair reads were generated (Table 1). Assuming an average read-length of 150bp and considering that the genome size of *C. maenas* is estimated to be around 1-1.2 Gb [46, 47], the estimated coverage was approximately 60x for paired-end and 19x for mate-pair libraries. The quality of the sequencing data for the 300-500 bp paired-end and 3500-5624 bp mate-pair libraries (Run 1) are shown in Figure 1 and Figure 2 respectively. There was a significant drop in Phred[1] quality around 80 bp into the sequencing reads. The same trend was present in the 500-700bp paired-end and 7361-11741 bp mate-pair libraries. While drops in quality over the sequencing reads are to be expected in sequencing datasets obtained from Illumina sequencing, the observed drops in sequencing quality were greater than expected, indicating issues with the sequencing process (possibly to do with loss of quality of sequencing reagents during storage). Since working with data from a poor quality sequencing run can have severe impacts in later stages of the analysis, another separate sequencing run (Run 2) using the same libraries was performed (except the 7361-11741 bp mate-pair libraries due to insufficient quantity).

[1] Phred quality scores Q are defined as a property which is logarithmically related to the base-calling error probabilities P. $Q= -10 \log_{10} P$

Table 1 Number of paired-end and mate-pair reads obtained before (raw reads) and after quality-trimming (processed reads)

Library	Number of reads	Number of reads after quality trimming (% of original reads)
Paired-end 300-500 (Run 1)	89,022,217	79,938,270 (90 %)
Paired-end 300-500 (Run 2)	95,325,569	88,033,605 (90 %)
Paired-end 500-700 (Run 1)	75,202,448	66,821,819 (89 %)
Paired-end 500-700 (Run 2)	81,640,231	73,673,023 (90 %)
Mate-pair 3500-5624 (Run 1)	63,224,417	40,162,634 (64 %)
Mate-pair 3500-5624 (Run 2)	42,848,052	27,198,748 (63 %)
Mate-pair 7361-11741 (Run 1)	23,701,171	10,761,155 (45 %)

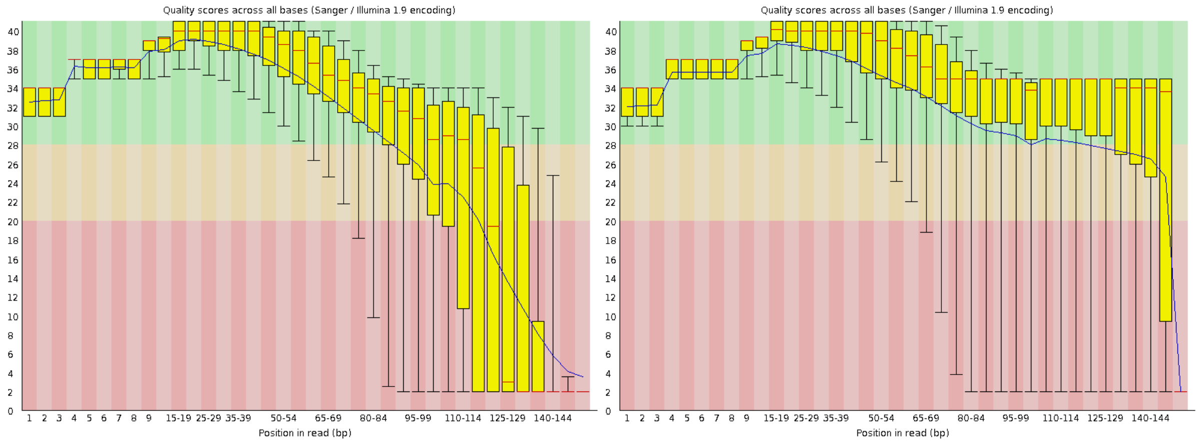


Figure 1 FastQC per base sequence quality for 300-500bp Paired-end sequencing data Run 1). Left: first read in pair. Right: second read in pair. Quality scores are in Phred (Quality = -10 $\log_{10}$ Probability for incorrect basecall) and shown on the y-axis. The x-axis denotes the position in the sequencing reads. Quality drops progressively along the length of the sequencing reads.

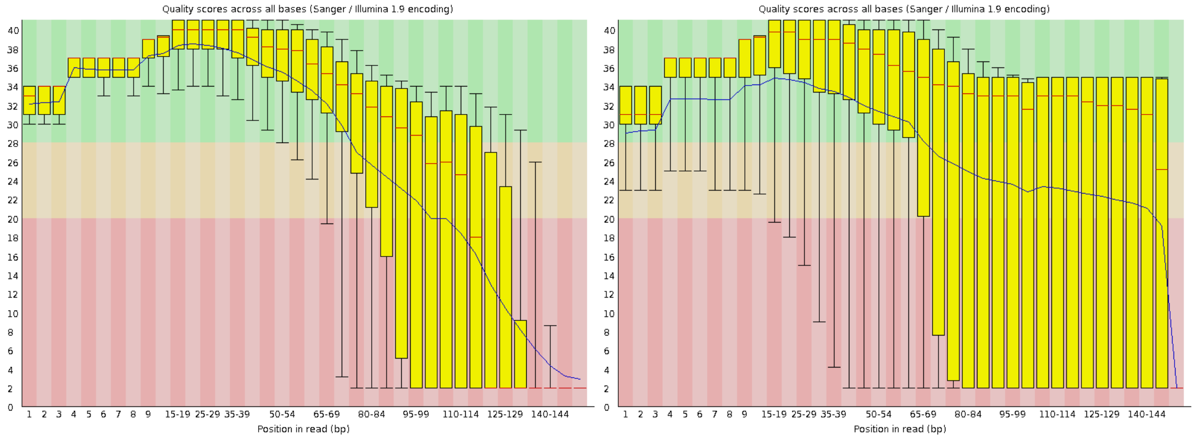


Figure 2 FastQC per base sequence quality for 3500-5624bp Mate-pair sequencing data (Run 1). Left: first read in pair. Right: second read in pair. Quality scores are in Phred (Quality = -10 $\log_{10}$ Probability for incorrect basecall) and shown on the y-axis. The x-axis denotes the position in the sequencing reads. Quality drops progressively along the length of the sequencing reads.

The quality of re-sequenced samples (Run 2) is shown in Figure 3. Comparison of the quality scores for the paired-end libraries indicated that the second sequencing run significantly improved the quality of the data. The quality of the sequences of the repeated mate-pair library also improved, but significant drops in quality were still present. This could be indicative of an issue with the mate-pair libraries themselves, which are often considered difficult to produce. Because of limited amount of libraries available and insufficient high quality DNA from the same individual crab, no more additional sequencing runs were performed and further quality control was performed bioinformatically (e.g. by removing bases with low confidence scores from the reads).

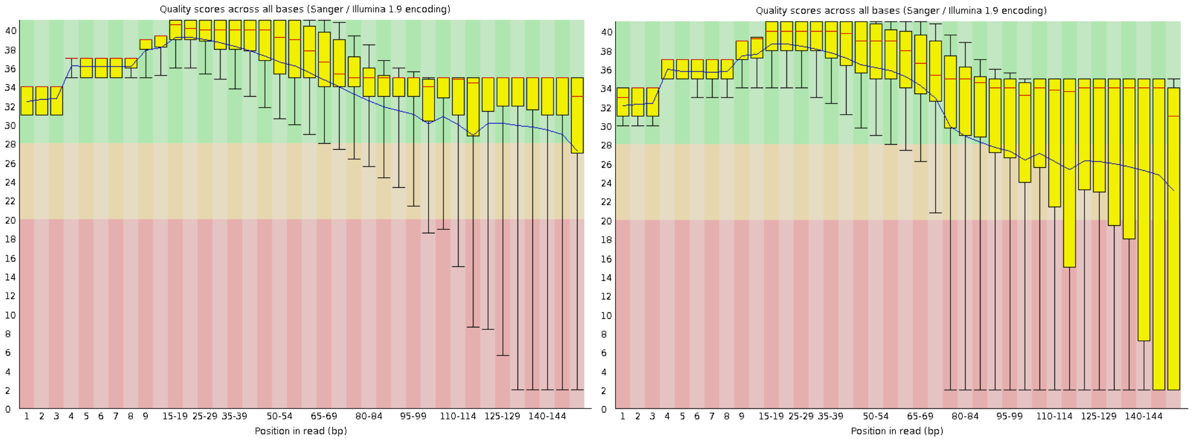


Figure 3 FastQC per base sequence quality for re-sequenced samples (Run 2). Left: first read in 300-500bp paired-end library. Right: first read in 3500-5624bp mate-pair library. Quality scores are in Phred (Quality = -10 $\log_{10}$ Probability for incorrect basecall).

Sequencing adapters and low quality sequences were removed from both datasets (Run 1 and Run 2). This cleaning step resulted in a significant loss of sequencing reads, particularly for the mate-pair libraries. Table 1 shows that overall 90 % reads were retained for the paired-end libraries while in the mate-pair libraries around 37-55 % of pairs did not pass the quality thresholds. While nearly equal amounts of read-pairs were retained in both sequencing runs, after quality-trimming, the remaining reads from Run 1 were significantly shorter. For example: the reduction in the total number of bases for the first reads in the paired-end library with insert size 300-500bp was 53 % in Run 1 compared to 74.3 % in Run 2 (see Table S 1). The genome size of *C. maenas* is estimated to be around 1-1.2 Gb [46, 47]. Taking 1 Gb as reference, the coverage of the DNA sequencing libraries of the *C. maenas* genome was 69.6 x, 64.6 x for paired-end and 5x for matepair libraries (Table 2). During the pre-processing step another issue was identified. It appears that the *C. maenas* DNA sequencing reads from both runs were enriched for AG/AC/TG/TC repeat reads as is illustrated by the GC content graphs in Figure 4. A further peak was present around 66% GC content, also indicative of repeats such as CAG. Additional quality filtering was performed to deal with this over-representation of repeat reads. Overall, the genome of *C. maenas* appears to be AT rich, with an average GC content around 44 %. Genome assemblies of other aquatic invertebrates e.g. like the Chinese mitten crab (*Eriocheir sinensis*) [48], draft genome assembly GC content of ~43 %, and *Daphnia pulex* with a GC content of 42 % [49]. Within the group of arthropods there is more variation, e.g. the red flour beetle (*Tribolium castaneaum*) with 34 % and *Drosophila* species approaching 41 % [50]. Coding regions within the genome often have a higher GC content; therefore an overall GC content of 44 % is another hint at the repetitive nature of the *C. maenas* genome [51].

Table 2 Genome coverage after quality-trimming

Library	**Length Read 1 (bp)**	**Length Read2 (bp)**	**Total length (bp)**	**Coverage (x)**
PE 300-500	7,075,894,970	8,610,319,484	15,686,214,454	15.7
PE 300-500_re	10,622,989,288	10,622,989,317	21,245,978,605	21.2
PE 500-700	5,652,491,870	6,840,897,257	12,493,389,127	12.5
PE 500-700_re	7,781,512,294	7,531,279,299	15,312,791,593	15.3
MP 3500-5624	1,168,869,646	1,304,617,677	2,473,487,323	2.5
MP 3500-5624_re	942,637,509	912,979,760	1,855,617,269	1.9
MP_7361-11741	265,830,974	294,561,463	560,392,437	0.6
Total	33,510,226,551	36,117,644,257	69,627,870,808	69.6

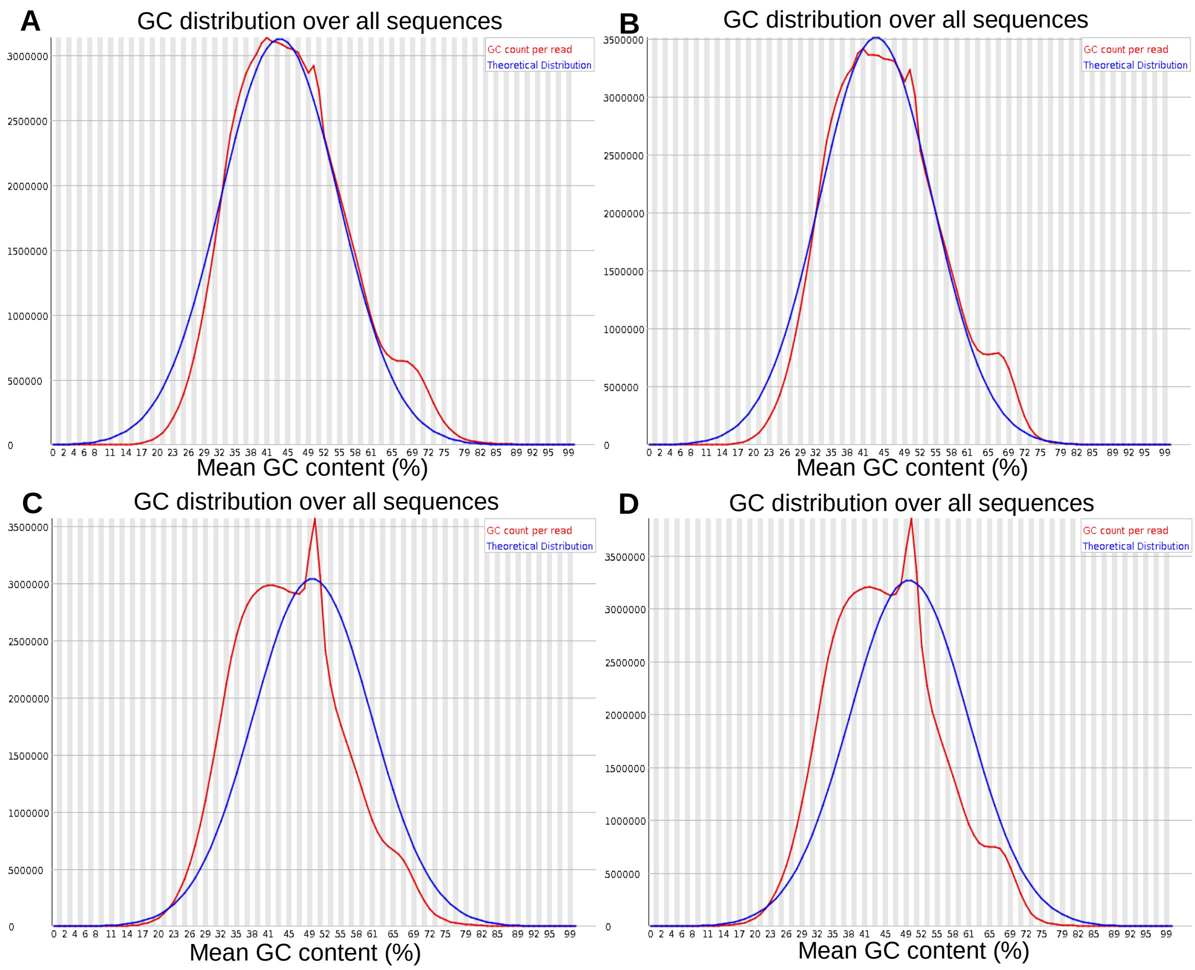


Figure 4 GC content of raw and processed sequencing reads. GC content distributions for raw and preprocessed sequence data. A) raw data for 300-500 bp paired-end library Run1. B) raw data for 300-500 bp paired-end library Run2. C) preprocessed data for 300-500 bp paired-end library Run1. D) preprocessed data for 300-500 bp paired-end library Run2.

4.5.2 De novo genome scaffold assembly

The filtered DNA-sequencing reads were used as input for a *de novo* genome assembly. When using Illumina sequence reads, the quality/reliability/completeness of *de novo* genome assembly is heavily dependent on the chosen assembler. Therefore, several assemblers, including ALLPATHS, platanus, SPAdes and SOAP-denovo2 and SOAP-denovo2+BESST scaffolder, were applied to the data and results compared. Although there is no universal consensus on how *de novo* assemblies can be compared, there are some statistics that are indicators of quality, including the number of contiguous sequences and/or scaffolds, the N50[2] and the total length of the generated genome scaffolds. The graphs in Figure 5 show the statistics for all the generated assemblies.

[2] The N50 length is defined as the length N for which 50% of all bases in the sequences are in a sequence of length L < N

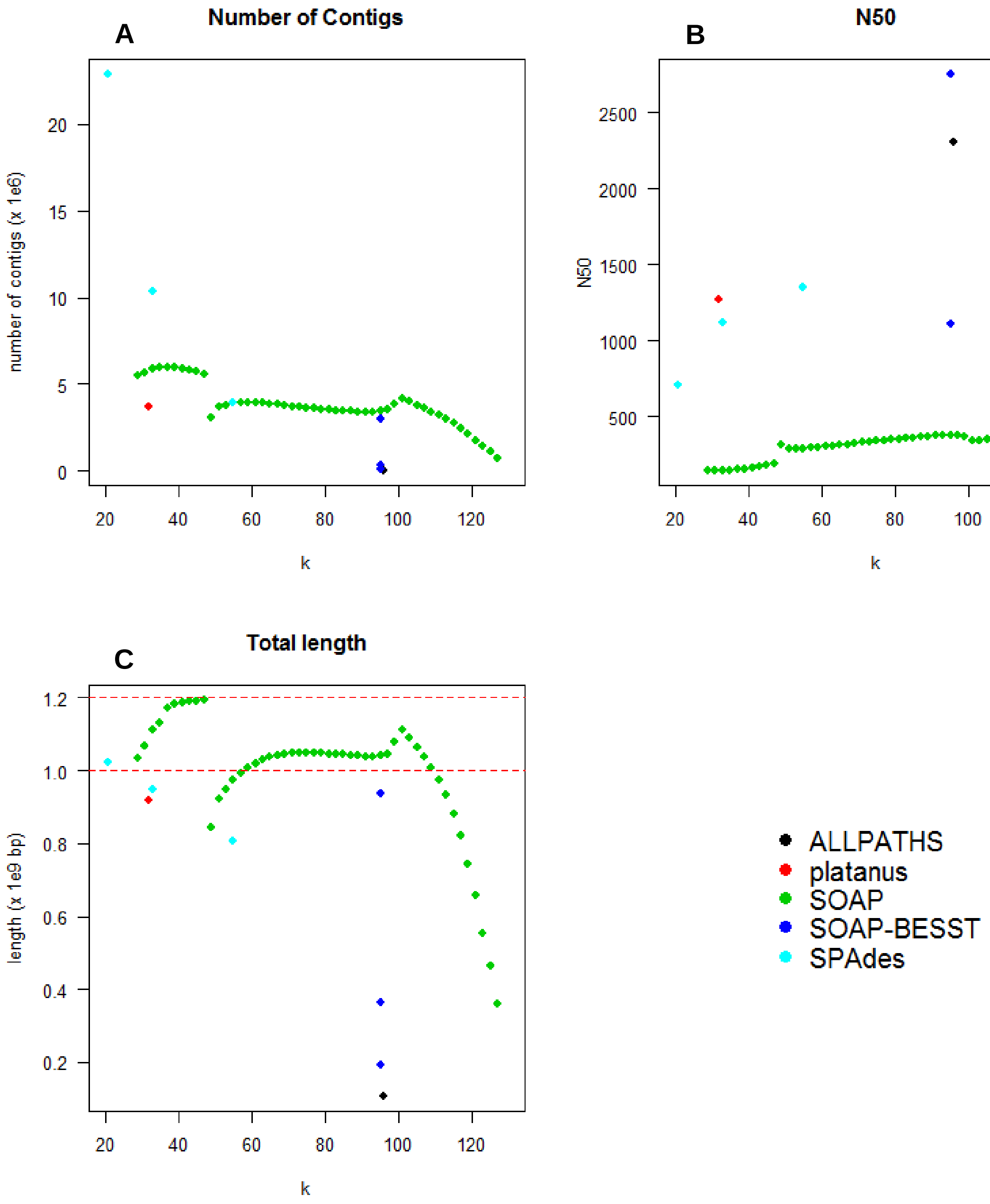


Figure 5 *C. maenas de novo* genome assembly statistics. A) Number of scaffolds in the *de novo* genome assemblies. The kmer parameter is shown on the x-axis and millions of contigs on the y-axis. B) N50 statistic for *de novo* assemblies. Kmer parameter on the x-axis and N50 (bp) on the y-axis. C) The total length of the assembly. Kmer parameter on the x-axis and total length in billion bp on the y-axis. Red lines indicate the estimated size of the *C. maenas* genome (1 – 1.2 Gb).

The total lengths of the *de novo* assemblies varied significantly. The majority of assemblies showed lengths over 1 Gb whereas others fell significantly short of that. The main reason behind this is that the assemblers put a threshold on contig/scaffold length in the final output. The total length of the *C. maenas* genome is estimated at 1-1.2 Gb and most assemblies had a total length within this range. However, there was a significant degree of fragmentation with most assemblies sporting over 3 million scaffolds. Comparing the results of *C. maenas* assemblies to the recently published draft genome of the Chinese mitten crab (*Eriocheir sinensis*), which encompasses 17,553 scaffolds covering a total of 1.66 Gb [48] illustrates this [48]. Another published aquatic crustacean genome, the cherry shrimp *Neocaridina denticulate*, showed similar results to that of *C. maenas* [52], with a highly fragmented assembly consisting of over 3,3 million contigs with an N50 of 400 that cover a total of 1.2 Gb of an estimated genome size of 3 Gb.

The N50 statistics demonstrate the fragmentation of the *C. maenas* genome assemblies was similar to that obtained for *E. sinensis*: between 500-3000 bp in the first against 224kb in the latter [48]. The ALLPATHS and SOAP-BESST assemblies for *C. maenas* resulted in the lowest degree of fragmentation, illustrated by the N50 and scaffold count in Figure 5. This lower degree of fragmentation, which is still large when compared to the *E. sinensis* draft genome, comes at the cost of reduced coverage of the estimated genome size for *C. maenas*. Since very small (< 500 bp) scaffolds are unlikely to contain full genes as they are shorter than the vast majority of genes [53] they are uninformative in relation to generating gene models for *C. maenas.* Thus further analyses only focussed on the ALLPATHS and SOAP-BESST assemblies since they encompassed the larger scaffolds and thus the most meaningful information.

The significant fragmentation in the *de novo* assemblies can have multiple causes. Firstly, biased GC content can have detrimental effects on assembly. In addition, during the sequencing process itself strong GC bias can yield lower coverage [54]. GC % biased areas lead to fragmented assembled scaffolds as was shown by Chen *et al.* 2013 for seven *de novo* genome assemblers (including ALLPATHS and SOAP-denovo) [54]. Indeed, the *C. maenas* raw- and preprocessed sequencing reads show a GC% distribution around 40 % (Figure 4) thus there are reads with difficult GC % content. A second cause for fragmentation in assemblies is the presence of repeated regions in the genome. The challenges that repeats pose to genome assembly have been reviewed by Treangen and Salzberg 2012 [55]. In brief, repeats that are longer than the read length create gaps in the assembly since assemblers cannot distinguish these based on their sequence [55]. The repeats cause the formation of branches in the De Bruijn graphs used by the *de novo* assemblers which can lead to false joins, erroneous copy numbers or fragmented assemblies with small contigs [55]. Tandem repeat regions can often be collapsed into fewer copies which can have impact on studies involving copy number variation in these regions [55]. Some of these issues can be overcome through the use of paired-end and mate-pair reads which are able to produce "scaffolds" that connect distant regions in the genome to each other. In the present study, mate-pair reads with insert sizes ranging between 3509-5624 bp and 7361-11741 bp were included for this reason. However, as is illustrated in Figure 4, there are large peaks at 50 % GC and again at 66 % GC indicating that many reads were derived from simple repeats, an indication that the *C. maenas* genome is rich in such structures. The assembly results indicate that the available mate-pair libraries were not sufficient to overcome all of these problems, thus resulting in a fragmented assembly. Sequencing of *C. maenas* DNA on platforms that can offer longer read lengths would improve the quality of the de novo draft assembly. Single molecule sequencing platforms, such as the PacBio sequencer can offer reads ranging from 10kb to 15kb and can thus traverse more read regions. However, at the time of performing this study PacBio sequencing was prohibitively expensive and generation of a dataset with 20x coverage of the *C. maenas* genome or higher would cost over £ 50,000. However it has been shown that hybrid sequencing and assembly strategies, using both high coverage Illumina sequencing and long reads from the PacBio system at a lower depth, can be successful [56], and this in the future may lead to the opportunity to incorporate this data with longer reads and reduce the fragmentation of the current assembly.

Based on the quality of the assemblies obtained, the ALLPATHS and SOAP-BESST assemblies were assessed further. In a previous experiment RNA sequencing of *C. maenas* was performed and the reads from that study were aligned to the assembled scaffolds [12] (note that the current DNA-sequencing data originated from a different crab than the crabs used in Verbruggen *et al.* 2015). The alignment rate of these mRNA reads is a good indicator of the coverage of the transcriptome across the genome. Furthermore, these alignments can provide indications about the locations of introns and exons in the genome assemblies. Alignment rates of the RNA-Seq reads (derived from a variety of tissues including muscle, intestine, heart, hepatopancreas, gill, eye, nerve, haemolymph, ovary, testis, eggs and epidermis) varied along the different assemblies and was generally lower as the size of the assembly decreased (Table 3). Around 85 % of all RNA-sequencing reads mapped to the SOAP-BESST assembly which shows that both datasets agreed to a large extend, despite originating from different individuals. It is encouraging that filtering out the smaller scaffolds does not dramatically reduce the alignment rate. For example, removal of all scaffolds under 500 bp in length (~90 % of scaffolds) only resulted in a 17 % reduction of aligning RNA sequencing reads. This indicates that a large portion of *C. maenas* genes are present within the larger scaffolds.

Table 3 Bowtie2 Alignment of RNA sequencing reads from Verbruggen *et al.* 2015 to assembled genome scaffolds

Assembly	Number of Contigs	Total length (bp)	Aligned RNAseq reads (%)
SOAP k95 – BESST	3,020,931	937,113,974	85.6
SOAP k95 – BESST (> 500bp)	338,980	362,671,342	68.6
SOAP k95 – BESST (> 1000bp)	82,100	194,190,891	52.8
ALLPATHS	52,355	108,312,348	32.9

The good agreement with the RNA-sequencing data shows that the *de novo* assemblies, despite being fragmented, contain relevant biological information. Based on this observation a further judgement of the quality of assembly could be made through observing the presence of highly conserved genes (Table 4). Parra *et al.* 2007 created a list of core eukaryotic genes and a tool (CEGMA) that can be used for identifying these genes within sequences[26]. Running CEGMA on the draft assemblies showed presence of ~80 % of core eukaryotic genes, the full sequence could be retrieved for 30 % of core eukaryotic genes while ~50 % were partially retrieved. Like the RNA-sequencing alignment, these percentages were largely retained when the threshold for scaffold lengths increased. A different assessment of assembly completeness, the presence of universal single copy orthologs for metazoans and arthropods (BUSCO, as selected by Simao *et al.* 2015 [27]; Table 5), showed similar results. Overall half of the orthologue groups were identified, but a large portion was fragmented (Table 5). Combined, these results show that while the genome scaffolds comprise large portions of RNA-sequencing reads, the identification of whole genes was still problematic due to fragmentation. For comparison, the percentage of aligned core eukaryotic genes to the *E. sinensis* draft genome was 66.9 % [48]. This puts the percentages for *C. maenas* in a better perspective since despite being significantly less fragmented; the *E. sinensis* genome contains 12 % less core eukaryotic genes. The *N. denticulate* draft genome recovered 99 .3 % of core genes, but this number cannot be compared directly since the parameters for CEGMA were changed to be less stringent in that assessment (tBLASTn e-value threshold to 1e^{-3}) [52]. Based on the transcriptome mapping and gene conservation results we chose the SOAP-BESST (> 500 bp) genome assembly for further annotation since it offered the best compromise between genome size, number of contigs and mapping RNA-sequencing transcripts.

Table 4 Presence of core eukaryotic genes in the genome scaffolds as determined by CEGMA.

Assembly	Full CEGMA genes (%)	Partial CEGMA genes (%)
SOAP k95 – BESST	29	52
SOAP k95 – BESST (> 500 bp)	28	51
SOAP k95 – BESST (> 1000 bp)	28	47

Table 5 Presence of single copy orthology groups in genome scaffolds as determined by BUSCO.

Assembly	Busco group (# groups)	Complete	Duplicated	Fragmented	Missing
SOAP k95 – BESST	metazoan (843)	193 (23 %)	10 (1 %)	203 (24 %)	447 (53 %)
SOAP k95 – BESST (> 500 bp)	Metazoan (843)	189 (22 %)	10 (1 %)	211 (25 %)	443 (53 %)
SOAP k95 – BESST (> 1000 bp)	metazoan (843)	192 (23 %)	11 (1 %)	189 (22 %)	462 (55 %)
SOAP k95 – BESST	arthropods (2675)	538 (20 %)	25 (1 %)	482 (18 %)	1,655 (62 %)
SOAP k95 – BESST (> 500 bp)	arthropods (2675)	570 (21 %)	25 (1 %)	471 (18 %)	1,634 (61 %)
SOAP k95 – BESST (> 1000 bp)	arthropods (2675)	570 (21 %)	25 (1 %)	453 (17 %)	1,652 (62 %)

4.5.3 De novo genome scaffold annotation

An extensive analysis pipeline was developed to annotate the scaffolds (Figure 6), which included a series of bioinformatics tools and several external datasets, including RNA and small RNA sequencing reads and an assembled *C. maenas* transcriptome (from Verbruggen *et al.* 2015) and curated gene and protein sequences from public databases (Figure 6). The pipeline analysed the presence of repeat regions, retrotransposons, miRNAs and gene models. The assembly chosen for annotation was the SOAP-denovo2 assembly (k = 95) combined with the BESST scaffolder, filtered for scaffolds larger than 500 bp (supplementary file 1: CMaenas-SOAP95-BESST_500bp.fa).

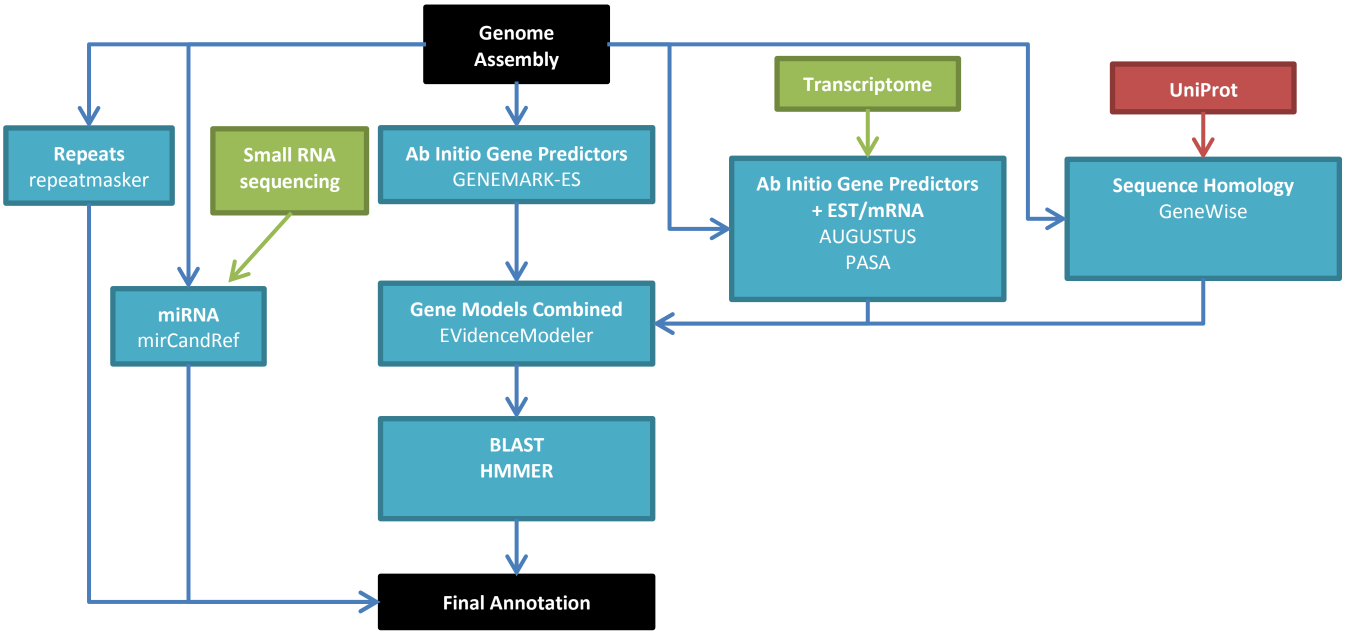


Figure 6 Genome annotation pipeline used in this study. The genome assembly was annotated using several software packages, combined with external data sources when appropriate. Transcriptome data included both sequencing reads and assembled transcripts from Verbruggen et al. 2015. Small RNA sequencing data derived from a pilot experiment (van Aerle et al., unpublished). UniProt data consists of Malacostrata protein sequences. Color indicator: black = input/output, green = sequencing data, red = external database and blue = bioinformatics software.

4.5.4 Repeats and Transposable Elements

Repeat structures and transposable elements in the genome were identified with RepeatMasker 4.0.5 [28]. The total percentage of tandem repeat regions in the *C. maenas* genome could not be assessed via RepeatMasker because the repeats were not represented by their true length in the assembly. Nonetheless, an indication could be given of the repeat content of available scaffold sequences and a summary is given in Table 6. It appears that even though most repeats should not be present in their full length, around 7.5 % of the scaffold consisted of simple repeat sequences, mainly representing low complexity regions and simple repeats (e.g. microsatellites). Table 6 also lists the transposable elements discovered by RepeatMasker. The most abundant class of transposable elements were the long interspersed elements (LINEs). Indeed, LINEs are the most reported retrotransposons amongst all crustacean species [57]. The same study illustrated that Copia retrotransposon elements are scarce compared to Gypsy retrotransposon elements [57]. Through PCR the presence of 35 Copia and 45 Gypsy elements in 15-18 crustacean species was shown [57]. Repeat structures in *C. maenas* showed a similar trend. The total number of Copia elements (911) was low as compared to Gypsy (13840). Furthermore the diversity of Gypsy elements was larger: 269 Gypsy elements to 60 Copia elements. These numbers are large compared to the elements identified by Piednoël et al 2013 [57]. A possible explanation is that, compared to PCR detection, the results are inflated through the identification method employed by RepeatMasker – resulting in false positives. Most of the classified elements had between 25 % and 50 % substitution compared to the consensus sequence. A conservative BLAST search identified only two Gypsy elements and no Copia elements (Table 7).

Table 6 Repeat elements present in the *C. maenas* scaffolds as identified and classified by RepeatMasker

RepeatMasker		**Number of elements**	**Length occupied**	**Percentage of sequence**
SINEs		79	7,142	0
	ALUs	1	145	0
	MIRs	35	9,091	0
LINEs		6,646	2,760,013	0.76
	LINE1	386	62,765	0.02
	LINE2	213	36,402	0.01
	L3/CR1	6,012	2,650,712	0.73
LTR		3,905	851,054	0.23
	ERVL	316	45,089	0.01
	ERVL-MaLRs	654	118,836	0.03
	ERV_classI	2,380	528,118	0.15
	ERV_classII	34	5,978	0
DNA elements		3,606	1,047,817	0.29
	hAT-Charlie	1,265	371,738	0.1
	TcMar-Tigger	1,611	450,832	0.12
Unclassified		124,397	13,800,909	3.81
Simple repeats		410,556	27,227,176	7.51
Total		549,189	45,694,111	12.6

Table 7 Repeat elements present in the *C. maenas* scaffolds identified by BLASTn

Repeat element	Length (bp)	> 75 % identity (count)	> 90 % identity (count)
Art1_Cis	3223	3	
Chapaev3-1_HR	2434	5	
Chapaev3-1_SM	2407	4	
EnSpm-1_LSal	1417	5	2
EnSpm-2_LMi	2299	107	56
EnSpm-N1_LMi	1559	1	
Gypsy-10_LVa-I	3580	3	
Gypsy-14_DAn-I	6105	4	4
hAT-17_LSal	2233	34	21
hAT-17A_LSal	2445	43	26
hAT-26_SM	3040	1	
hAT-N45_LSal	208	8	8
hAT-N6_LSal	1302	143	121
Helitron-like-11_Hmel	1284	1	1
Helitron-like-13_Hmel	3491	14	13
Helitron-like-6b_Hmel	1343	14	13
LIN9_SM	5304	2	
Mariner-N31_LSal	634	10	7
piggyBac-4_BF	9533	1	
R1-2_PBa	5457	8	8
R2-1_SM	5417	4	
REP-6_LMi	16159	2	
RTE-3_LVa	3654	4	
Sat-1_LVa	6803	63	63
SAT-1_SK	1015	4	
tRNA-1_LSal	83	3	
VENSMAR1	1293	3	

4.5.5 miRNA detection

Micro-RNAs (miRNAs) are short (20-25 bp) non-coding RNA molecules that can regulate the translation of mRNA, either through blocking translation or facilitating degradation [58]. miRNAs play important roles in the regulation of many biological processes, including developmental timing, cell death, cell proliferation, haematopoiesis and patterning of the nervous system [58]. Additionally, it has been shown that during virus-host interactions both host and viral miRNAs play an important part that determines the fate of the host cell [11]. All animals have the ability to produce miRNAs which are produced from precursor molecules that are characterized by specific hairpin structures. Detection of sequences with the potential to form these secondary structures enables *in silico* prediction of miRNAs in *de novo* assembled genomes. Small RNA sequencing data for *C. maenas* derived from a pilot experiment (van Aerle *et al.*, unpublished) was used in the miRCandRef pipeline, resulting in identification of 185 miRNA precursor structures in the *C. maenas* genome scaffolds. Comparing these miRNA sequences to known miRNAs in miRbase and Genbank showed that 31 out of 185 miRNAs had significant sequence homology (Table S2). This indicated that potentially many novel putative miRNAs have been identified based on the evidence of genome and miRNA sequencing. Table S2 shows the miRNAs that were identified and in some cases their annotation.

Prediction of the targets of miRNAs *in silico* was based on free energy calculations between miRNA and known *C. maenas* transcripts. Two target predictors were used and results compared (Table 8), hits for both predictors can be found in supplementary file 2: Cmaenas_miRNA_targetprediction.rar. Both predictors yielded potential targets for all miRNAs, but analysis showed that agreement between predictors was low. This is because predictor algorithms tend to produce many false positive results [59]. The total number of potential miRNA-mRNA combinations predicted by miRanda was significantly higher than predictions derived by MicroTar (1,953,229 for miRanda compared to 7,691 for MicroTar). Comparing predictions from both methods resulted in 2,746 shared miRNA-mRNA combinations. The general level of disagreement indicates the issues with *in silico* prediction, especially when seeking targets in a whole transcriptome rather than in a list of selected genes. The capricious nature of these *in silico* predictions indicate that miRNA-mRNA combinations should be verified through additional studies and drawing conclusions based on these results would not be fruitful.

Table 8 miRNA Target prediction with miRanda and MicroTar

Predictor	Total miRNA-transcript predictions (< -10 kcal/mol)	Number of miRNAs for which potential targets were found	Hits in common with the other predictor
miRanda	1,953,229	185 (100 %)	2746 (0.0014 %)
MicroTar	7,691	185 (100 %)	2746 (35.7 %)

4.5.6 Gene model generation and analysis

A gene model contains the locations of genes, including their exons and introns, on the scaffold. Prediction of genes in the *C. maenas* genome assembly was based on *ab initio* algorithms, i.e. predictions of the presence of genes based purely on the sequence information. Other indications of gene locations could be derived through the inclusion of existing information on *C. maenas* gene sequences. Combining the genome scaffold sequences with known transcript sequences should result in more reliable predictions. The annotation workflow of Figure 6 shows inclusion of known *C. maenas* protein sequences and transcriptome data. The transcriptome data is derived from Verbruggen et al. 2015 [12] and used in two ways: firstly, the raw sequence data is used in combination with the Augustus *ab initio* gene predictor to inform on the location of exon boundaries [32, 33]; secondly, the assembled transcripts were aligned to the genome scaffold using PASA [34]. The final gene model was created by combining the results of each individual prediction method through Evidence Modeller [34]. Some statistics on the generated predictions are summarized in Table 9. The *ab initio* gene predictors found an almost identical number of genes in the scaffolds: 71,394 for Augustus and 71,802 for GeneMark. The number of predicted exons on these genes was found to diverge, with GeneMark identifying more exons than Augustus. It is possible that the exon boundary information derived from the RNA-sequencing data prevented unnecessary splitting of exons in Augustus, resulting in reduced exon numbers. Through BLASTx and GeneWise 95 % of Malacostraca proteins could be aligned to genome scaffolds. There was a degree of redundancy due to orthologues. Alignment of assembled transcripts to the genome scaffold proved difficult since validated matches were only identified for around 5,500 transcripts. PASA utilizes only near perfect alignments. These alignments are required to align with a specified percent identity (typically 95%) along a specified percent of the transcript length (typically 90%). Each alignment is required to have consensus splice sites at all inferred intron boundaries, including (GT/GC donor with an AG acceptor, or the AT-AC U12-type dinucleotide pairs) [34]. The fragmentation of the genome scaffold could cause such PASA alignments to fail since transcripts might span multiple scaffolds.

Table 9 Gene model evidence statistics

Prediction Type	**Prediction Tool**	**Gene**	**Exons**
Gene Prediction	GeneMark.hmm	71,802	193,476
Gene Prediction	Augustus	71,394	125,561
		Aligned malacostraca proteins	
Protein Alignment	GeneWise	72,711	
		Aligned transcripts	**Scaffolds**
Transcript Alignment	blat	5,263	19,218
	gmap	5,766	21,974

The evidence presented in Table 9 was combined in Evidence Modeller, resulting in a single gene model for the *C. maenas* genome assembly. The produced gene model (Suppl. File S3) indicated presence of genes on 8,624 scaffolds out of 338,980 scaffolds with a total of 10,355 genes predicted. This number is lower compared to the *E. sinensis* draft genome wherein 14,436 genes were identified with AUGUSTUS [48]. Observing the length of the scaffolds with successful annotations showed that typically only the larger scaffolds contain predicted genes (Figure 7). This is in agreement with the evidence files provided to Evidence Modeller. Larger scaffolds have a better chance of successful protein and transcript alignments, which are the most reliable form of evidence for annotation. The 10,355 genes sequences were compared to the NCBI non-redundant database, using BLASTx resulting in 5,615 predicted gene sequences having significant sequence similarity (e-value below 1e^{-10}). Thus, 54 % of predicted genes could be annotated through BLASTx which is higher than a similar search of assembled transcriptome sequences performed in Verbruggen et al. 2015 (~ 30 %; all of the BLASTx results can be found in Suppl. File S4). Lastly a total of 6,470 (62 %) of predicted genes were found to encode for proteins containing conserved Pfam protein domains (Suppl. File S5).

In summary, the *C. maenas* genome scaffold produced a gene model that reflects issues with the assembly. The degree of fragmentation hampered the prediction abilities of both *ab initio* and guided algorithms. This is particularly shown by the fact that only 2.5 % of scaffolds contained predicted genes. Significant enhancement of the *C. maenas* gene model would require improvement of the assembly, either through additional (long read) sequencing or novel assembly methods.

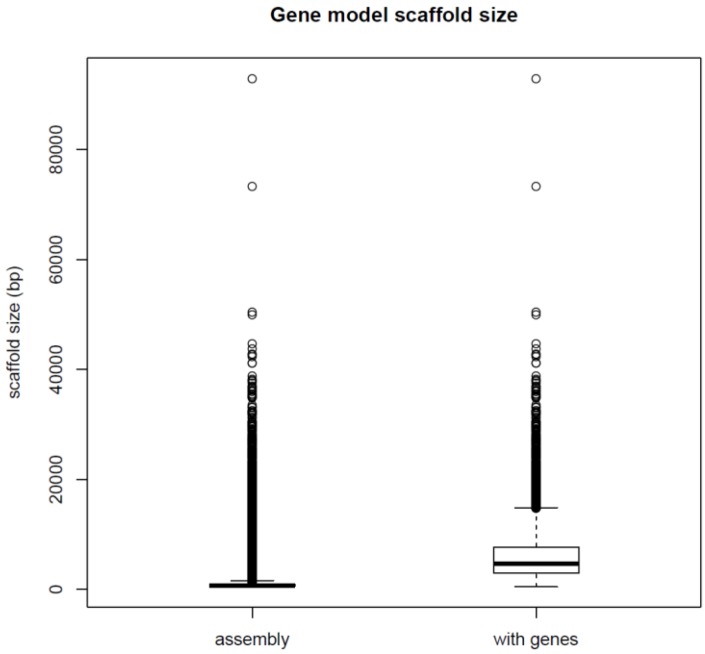


Figure 7 Gene model scaffold size. Scaffold length distributions for the whole assembly and the scaffolds that contain a predicted gene according to the gene model.

4.5.7 Detection of viral inserts

It has been observed that over time it is possible for fragments of viral genomes to become integrated into the host genome [60-62]. Retroviruses are obliged to integrate their genomes into the host genomes and therefore carry machinery capable of achieving this. However on rare occasions it is possible for other types of viruses to integrate into the host genome and even the germ line, these are labelled as endogenous viral elements (EVEs) or viral 'fossils' [61]. For dsDNA viruses like WSSV a possible scenario for EVE formation would include double strand breaks in both the viral and host genomes within a germ line cell. Subsequently, the sequence derived from the viral genome would have to be ligated into the host genome, resulting in integration. Finally, this genotype has to survive and persist in following host generations.

It is possible for some EVEs to produce RNA products. For example, integrated viral ORFs with viable promoters could yield mRNA [61]. Furthermore it could be possible for EVEs to yield small RNA molecules and it has been hypothesized that such small RNAs could convey host resistance through RNAi mechanisms. A viral integration event of WSSV into the *C. maenas* genome could thus be the basis of the resistance of *C. maenas* to WSSV infection [9]. Viral integration events have been reported for a WSSV-like virus in the Jamaican bromeliad crab *Metopaulias depressus* [63]. In their work Rozenberg et al. 2015 showed the presence of WSSV-like ORFs in 454 sequencing data of the Jamaican crab and through *de novo* assembly showed a viral and host elements on the same contig [63].

Identification of WSSV EVEs in the *C. maenas* genome was investigated through aligning pre-processed paired-end DNA sequencing reads to the Chinese WSSV isolate genome. The alignment of reads was investigated in the IGV browser. Visual inspection indicated that no WSSV-related sequences were present in the *C. maenas* DNA sequencing data. Reads that did successfully align were limited to repeat regions. Even when the bowtie2 aligner was used in local mode, which allows for partial alignment of reads, no regions of the WSSV genome were found sufficiently covered. This leads to the conclusion that an integration event as observed in *M. depressus* has not occurred in *C. maenas,* based on the sequence of the individual analysed. Through a similar method the presence of EVEs of viruses other than WSSV was investigated. All viruses with a known genome sequence in RefSeq were used as reference for read alignment, and coverage calculations were used to identify viruses with significant numbers of mapping reads. Table 10 shows the viral genomes with the highest count of mapping sequencing reads. Enterobacteria phage PhiX, used as spike-in during the DNA-sequencing, was successfully identified, as expected. Its successful identification shows that viral sequences can be identified through this analysis pipeline. Reads mapping to other viruses in Table 10 were investigated to see whether these were legitimate indications of EVE. It appeared that reads mapping to these genomes are all repeating reads, for example $(TAC)_n$. Thus reads with GC content ranging between 31-35, 48-52 and 64-68 were removed (Table 11) and the remaining reads mapped again to the viral genomes. Besides Enterobacteria phage PhiX the number of mapped reads to viruses significantly reduced. Additionally, sequences that mapped to viruses other than PhiX still represented repeats that were marginally outside the set GC content ranges.

Table 10 Number of DNA-sequencing reads mapping to viral genomes

NCBI Identifier	Name	Number of reads
gi|9626372|ref|NC_001422.1|	Enterobacteria phage phiX174 sensu lato, complete genome	965 560
gi|725948879|ref|NC_025417.1|	Staphylococcus phage Team1, complete genome	431 010
gi|564292828|ref|NC_023009.1|	Staphylococcus phage Sb-1, complete genome	50 222
gi|448824857|ref|NC_020231.1|	Caviid herpesvirus 2 strain 21222, complete genome	40 850
gi|20198505|ref|NC_002512.2|	Murid herpesvirus 2, complete genome	26 013

Table 11 Number of DNA-sequencing reads mapping to viral genomes after removal of low-complexity reads filtered

NCBI Identifier	Virus species	Number of reads
gi|9626372|ref|NC_001422.1|	Enterobacteria phage phiX174 sensu lato, complete genome	562 255
gi|508181800|ref|NC_021312.1|	Phaeocystis globosa virus strain 16T, complete genome	669
gi|664651935|ref|NC_024474.1|	Pigeon adenovirus 1 complete genome, strain IDA4	327
gi|213159268|ref|NC_011588.1|	Oryctes rhinoceros virus, complete genome	224
gi|131840030|ref|NC_009127.1|	Cyprinid herpesvirus 3, complete genome	223

The search for EVEs was widened by including protein level searches. To this end a search was conducted between paired DNA sequencing reads and RefSeq protein sequences. The results were aggregated based on the virus species of the best DIAMOND BLASTx hit (given certain thresholds, see Methods). Results confirmed the presence of the Enterobacteria phage PhiX (Table 12). The only other listed virus was the Cotesia congregate bracovirus which had over 2 500 DNA sequencing reads mapping to its histone (YP_184795.1). Given that all reads map to a single protein of this virus and that histone sequences are prevalent amongst viruses and hosts, the data obtained did not provide sufficient evidence to suggest the presence of an EVE for this virus in the *Carcinus* genome.

Table 12 Number of reads of the *C. maenas* DNA sequencing dataset with similarity to viral protein sequences.

Virus species	Number of reads
Enterobacteria phage phiX174 sensu lato	330 878
Enterobacteria phage ID2 Moscow/ID/2001	33 888
Enterobacteria phage G4 sensu lato	33 850
Enterobacteria phage ID18 sensu lato	25 570
Enterobacteria phage WA13 sensu lato	17 507
Enterobacteria phage St-1	11 292
Enterobacteria phage alpha3	10 951
Cotesia congregata bracovirus	2 738
Enterobacteria phage lambda	147
Enterobacteria phage HK630	136

In conclusion, the identification of Entereobacteria phage PhiX in *C. maenas* DNA sequencing data indicated that the pipelines were successful in picking up viral sequences. However, besides PhiX there were no significant indications of viral sequences and thus no evidence for the presence of potential EVEs in the crab genome.

4.5.8 Carcinus maenas Down syndrome cell adhesion molecule *(Dscam)*

Traditionally, it was thought that acquired resistance to pathogens did not occur in organisms limited to an innate immune system, mainly due to the lack of the ability to produce antibodies. It has been shown, however, that arthropods like crabs, lobsters and shrimps display a form of immune specificity and immune memory [64]. When shrimp are exposed to antigens or pathogens they exhibit a highly-specific immune response on subsequent challenges with the same pathogens [64-67]. The biology behind this observation has not been elucidated completely, but there is some evidence for the involvement of Down syndrome cell adhesion molecules (Dscam) in this process [14]. A specific immune response requires receptors with a high degree of variability, capable of adjusting to specific pathogen challenges. Indeed, Dscam is a receptor that boasts an extremely high degree of diversity, possibly due to alternative splicing [68].

The Dscam receptor conforms to the following pattern: 9 immunoglobulin (Ig) domains - 4 fibronectin type (FN) domains - 1 Ig domain - 2 fibronectin domains – transmembrane (TM) domain – cytoplasmic tail [64]. While the structure of the Dscam gene is similar across both vertebrates and invertebrates, a notable difference is that only the arthropod Dscam displays hyper-variability [64]. The reasons for variation are tandem arrays for three Dscam Ig exons and one transmembrane domain. Due to alternative splicing of these exons a large number of Dscam variants can be produced. Since Dscam might be playing an important role in the *C. maenas* immune system and thus in WSSV infection/resistance it was of interest to identify whether a gene model for Dscam was identifiable in the *C. maenas* genome scaffold.

The sequence of *C. maenas Dscam* (GenBank: HG964670.1) was aligned to the *C. maenas* genome to identify scaffolds with significant sequence similarity. From this search scaffold_192 was identified to contain a *C. maenas Dscam* gene. To identify the location of exons alignments with GeneWise and BLAT alignments of several invertebrate Dscam protein sequences were visualized in IGV (Figure 8).

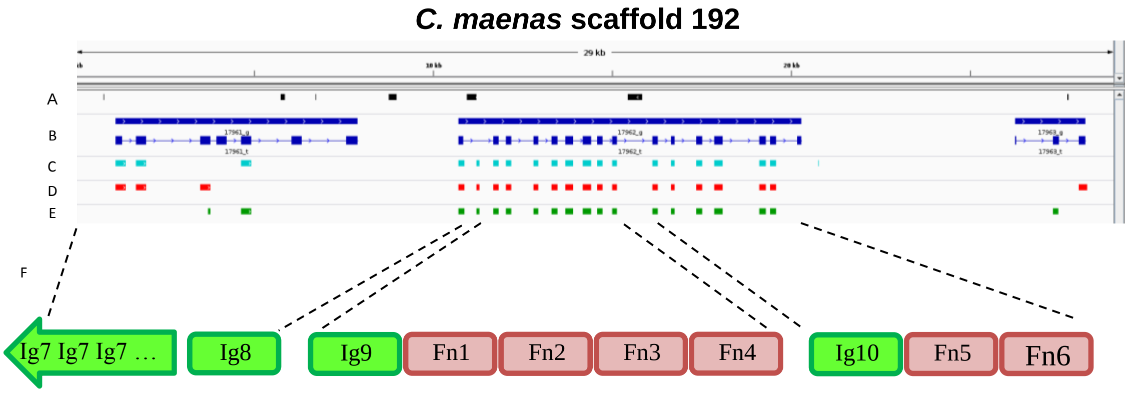


Figure 8 Alignment of invertebrate Dscam to *C. maenas* scaffold_192. The sequence of *C.maenas* scaffold 192 is visualized as a line with length indicators. Annotation is provided in the five lines below (A-E). A) location of 'N' nucleotides, i.e. gaps in the scaffold B) Genes predicted by GeneMark HMM, exons indicated by blocks and introns by lines. C) BLAT alignment of *C. maenas* Dscam (GenBank: HG964670.1), blocks indicate sequence similarity. D) BLAT alignment of *E. sinensis* Dscam (GenBank: AGL39311.1), blocks indicate similarity. D) BLAT alignment of *D. melanogaster* Dscam (GenBank: AAF71926.1), blocks indicate similarity. F) Pfam-A domains identified in the Dscam exons covered by regions within *C. maenas* scaffold 192. The Immunoglobulin domains are indicated by Ig and Fibronectin domains by 'Fn'. The scaffold does not cover the complete Dscam gene but ends near the likely Ig7 tandem array.

The GeneMark track in Figure 8 shows a discrepancy compared to the BLAT alignments. Whereas the BLAT results are spread across the whole scaffold, the GeneMark track is split into three separate genes. The gene model produced by EVidence Modeller aligns with the depicted GeneMark prediction. Given that the three known Dscam sequence all stem from the same gene it can be inferred that the split into three separate genes is erroneous. It was to be expected that the *ab initio* gene predictions are imperfect even after integration of other forms of evidence and therefore would benefit from manual curation. The sequences of *C. maenas*, *E.sinensis* and *D. melanogaster* Dscam align to similar regions across scaffold_192, indicating that the blocks in Figure 8 C-E are likely to be genuine exons. The sequences of these exons were extracted and used as input for a Pfam-A domain search, the result of which is shown in Figure 8 F. The sequence of identified Pfam-A domains follows the tail end of the Dscam blueprint: 9 Immunoglobulin domains – 4 Fibronectin domains – 1 Immunoglobulin domain – 2 Fibronectin domains. The Pfam-A domains show that not all Immunoglobulin domains are represented on scaffold 192. The scaffold thus is unlikely to contain the full *C. maenas Dscam* gene, which should extend in 5' direction on the genome. The hypervariable exon tandem arrays of invertebrate *Dscam* are located in the 2nd, 3rd and 7th Ig domains [69]. Unfortunately these are not incorporated on *C. maenas* scaffold 192. However the sequence that has been identified can be used in PCR or gene walking experiments to fill in the gaps of the *C. maenas Dscam* gene. Once that sequence information is available the changes in Dscam splicing during pathogen exposures can be assessed, for example with RNA sequencing data.

In the case of WSSV infection this has not been shown thus far. For example, work by Watthanasurorot *et al.* 2011 showed that injection of the signal crayfish *Pacifastacus leniusculus* with WSSV did not change *PlDscam* expression, furthermore the silencing of *PlDscam* did not have an effect on WSSV infection progression. Positive examples are available though, including the alteration of *Anopheles gambiae Dscam* (*AgDscam*) splicing upon exposure to Plasmodium parasites resulting in expression of *AgDscam* isoforms with higher affinity to the pathogen [70]. With the *Dscam* sequence available, studies for alternative splicing of *C. maenas Dscam* in response to pathogens, WSSV or other, can be monitored and outcome related to pathogenicity.

4.6 Summary

DNA was isolated from the muscle tissue of a shore crab (*C. maenas*) and subjected to next-generation sequencing, designed to yield a 50 x paired-end and 19x mate pair coverage of its 1 Gb genome. Assembly of the data into a draft genome showed that it was highly fragmented. The high degree of fragmentation appeared to be independent of the assembly method. Comparing this assembly to draft assemblies of other aquatic crustaceans showed a similar degree of fragmentation for *N. denticulate* but not in *E. sinensis*. Evidence in GC % of the sequencing data points toward a repeat-rich genome, which are notoriously difficult to assemble from short read NGS data. The generated draft genome assembly could be improved through additional sequencing coverage to resolve repeated regions. Such sequencing should preferably be performed on platforms that offer longer sequencing reads which are more capable of traversing the repeating regions. Alternatively, should new assemblers become available they can applied to the current dataset which could result in an improved draft genome. Shorter genomic features, such as miRNAs and their precursors could be identified and many novel putative miRNAs were found in *C. maenas*. The prediction of targets for miRNAs *in silico* proved challenging. However with reliable annotation of the UTRs of the genes, which are the locations where miRNAs tend to bind, improvement in *in silico* predictions might be possible. Nevertheless, it is likely that targets will have to be confirmed through laboratory experiments rather than via computational prediction. This genome sequencing project was performed in order to aid the study of viral infection and immunity and therefore we characterized the gene model for *C. maenas Dscam*. Analysis of the domains showed that only a partial segment of *Dscam* was present on scaffold 192, but this information can be used as a basis to inform on further experiments to sequence the full length of this gene. With the *Dscam* sequence available, studies for alternative splicing of *C. maenas Dscam* in response to pathogens, WSSV or other, can be monitored and outcome related to pathogenicity and invertebrate immune memory. The ability to instill a specific immune response to WSSV through manipulation of *Dscam* could be very valuable to crustacean aquaculture.The draft genome produced and described in this chapter is the first one produced for *C. maenas* and one of the few available for aquatic invertebrates. In the short term toxicology, evolutionary and pathogenic studies in this important invasive species can be supported by the availability of this genomic resource. On the longer term continued refinement of the genome build and generation of additional aquatic invertebrate genomes can accelerate the understanding of this group of species.

4.7 Supplementary Files

S1. CMaenas-SOAP95-BESST_500bp.fa: *C. maenas* genome scaffold as produced by SOAP-denovo2 (k=95) and BESST scaffolder.

S2. Cmaenas_miRNA_targetprediction.rar: miRNA target prediction with target annotation.

S3. Cmaenas_evm.gff3: gene model generated by EVidence Modeler for *C. maenas* assembly

S4. Cmaenas_evm_CDS_BLASTx.txt: Annotation of *C. maenas* evm CDS using BLASTx to NCBI nr database

S5. Cmaenas_evm_CDS_peptide.hmmtable: Pfam domains in *C. maenas* evm CDS

4.8 Supplementary Data

Table S 1 Total library lengths (in bp) before and after quality trimming

Library	length (bp) using raw reads	length (bp) using trimmed reads	Percentage remaining after quality trimming
PE_300-500_R1	13,353,332,550	7,075,894,970	53
PE_300-500_R2	13,353,332,550	8,610,319,484	64.5
PE_300-500_re_R1	14,298,835,350	10,622,989,288	74.3
PE_300-500_re_R2	14,298,835,350	10,622,989,317	74.3
PE_500-700_R1	11,280,367,200	5,652,491,870	50.1
PE_500-700_R2	11,280,367,200	6,840,897,275	60.6
PE_500-700_re_R1	12,246,034,650	7,781,512,294	63.5
PE_500-700_re_R2	12,246,034,650	7,531,279,299	61.5

MP_3500-5624_R1	9,483,662,550	1,168,869,646	12.3
MP_3500-5624_R2	9,483,662,550	1,304,617,677	13.8
MP_3500-5624_re_R1	6,427,207,800	942,637,509	14.7
MP_3500-5624_re_R2	6,427,207,800	912,979,760	14.2

Table S 2 Predicted *C. maenas* miRNAs and their annotations as determined by BLAST

miRNA	Sequence	Annotation	Hit
mirc_1	aagagagcuauccgucgacagu	miRNA_hit	*Drosophila yakuba* mir-281-2 precursor RNA (Dyak\mir-281-2), miRNA
mirc_3	guugugaccguuauaaugggca	miRNA_hit	TPA: *Capitella teleta* microRNA cte-mir-2001 precursor
mirc_4	uaggaacuucauaccgugcucu	miRNA_hit	*Drosophila melanogaster* strain Zimbabwe-109 mir-276a miRNA gene,
mirc_5	ugagaucauugugaaagcugauu	miRNA_hit	TPA: *Tribolium castaneum* microRNA tca-bantam-3p
mirc_7	aauugcacuagucccggccugc	miRNA_hit	*Drosophila melanogaster* strain Zimbabwe-109 mir-92b miRNA gene,
mirc_9	cgugcagaacgaauguccgca	miRNA_hit	PREDICTED: *Saimiri boliviensis boliviensis* LON peptidase N-terminal
mirc_11	ugacuagauccacacucaucca	miRNA_hit	*Apis mellifera* microRNA mir-279d (Mir279d), microRNA
mirc_12	aauggcacuggaagaauucacgg	miRNA_hit	*Drosophila yakuba* mir-263a precursor RNA (Dyak\mir-263a), miRNA
mirc_24	uugguaacuccaccaccguuggc	miRNA_hit	*Apis mellifera* microRNA mir-2765 (Mir2765), microRNA
mirc_30	uggcagugugguuagcugguugu	miRNA_hit	TPA: *Drosophila melanogaster* microRNA dme-miR-34-5p
mirc_32	uaucacagccagcuuugaugagc	miRNA_hit	*Apis mellifera* microRNA mir-2b (Mir2b), microRNA
mirc_34	ucucacuaucuugucuuucacg	miRNA_hit	*Acyrthosiphon pisum* microRNA mir-71 (Mir71), microRNA
mirc_35	ugaguauuacaucagguacuggu	miRNA_hit	*Nasonia vitripennis* microRNA mir-12 (Mir12), microRNA
mirc_37	uaucacagccaccuuugaugagcu	miRNA_hit	*Drosophila melanogaster* strain Zimbabwe-109 mir-2a-2 miRNA gene,
mirc_41	ugacuagaucuacacucauca	miRNA_hit	*Acyrthosiphon pisum* microRNA mir-279b (Mir279b), microRNA
mirc_48	cuuggcacuggaagaauucacag	miRNA_hit	*Acyrthosiphon pisum* microRNA mir-263b (Mir263b), microRNA
mirc_52	uggaauguaaagaaguauggag	miRNA_hit	*Drosophila yakuba* mir-1 precursor RNA (Dyak\mir-1), miRNA
mirc_53	uacuggccugcuaagucccaag	miRNA_hit	*Apis mellifera* microRNA mir-193 (Mir193), microRNA
mirc_54	agauauguuugauauucuugguug	miRNA_hit	*Bombyx mori* microRNA mir-190 (Mir190),

			microRNA
mirc_55	auaaagcuagauuaccaaagc	miRNA_hit	*Drosophila yakuba* mir-79 precursor RNA (Dyak\mir-79), miRNA
mirc_57	augcauugucguugcauugca	miRNA_hit	*Strongylocentrotus purpuratus* microRNA mir-33 (Mir33), microRNA
mirc_62	cuaaguacuagugccgcagga	miRNA_hit	*Drosophila melanogaster* strain Zimbabwe-109 mir-252 miRNA gene,
mirc_66	ccagaucuaacucuuccagcuca	miRNA_hit	TPA: *Capitella teleta* microRNA cte-mir-750 precursor
mirc_80	gaagcucguuucuacagguaucu	miRNA_hit	*Drosophila melanogaster* strain Zimbabwe-109 mir-993 miRNA gene,
mirc_95	ugagauucaacuccuccaacuuag	miRNA_hit	TPA: *Bombyx mori* microRNA bmo-miR-1175-3p
mirc_120	ucagguaccugauguagcgcgc	miRNA_hit	*Drosophila yakuba* mir-275 precursor RNA (Dyak\mir-275), miRNA
mirc_125	auauuguccugucacagcag	miRNA_hit	TPA: *Drosophila melanogaster* microRNA dme-miR-1000-5p
mirc_146	cuugugcgugugacagcggcu	miRNA_hit	*Drosophila yakuba* mir-210 precursor RNA (Dyak\mir-210), miRNA
mirc_155	caaugcccuuggaaaucccaaa	miRNA_hit	*Bombyx mori* microRNA mir-2788 (Mir2788), microRNA
mirc_164	uagccucuccucggcuuugucu	miRNA_hit	TPA: *Tribolium castaneum* microRNA tca-miR-282-5p
mirc_177	cucacaaaguggcuguuguaug	miRNA_hit	*Nasonia vitripennis* microRNA mir-2a (Mir2a), microRNA
mirc_2	uacccuguagauccgaauuugu	Blast_hit	*Drosophila busckii* chromosome 3R sequence
mirc_6	cggacauucguucugcacgccu	Blast_hit	*Pseudomonas cremoricolorata* strain ND07, complete genome
mirc_8	ucuuugguuaucuagcuguauga	Blast_hit	PREDICTED: *Haplochromis burtoni* uncharacterized LOC102292141
mirc_10	caucuuaccggacagcauuaga	Blast_hit	*Drosophila busckii* chromosome 2R sequence
mirc_13	cggcaucuguuggaguacaguag	Blast_hit	TPA_asm: *Aspergillus nidulans* FGSC A4 chromosome II
mirc_14	caugcagaacguaugucugca	Blast_hit	*Spirometra erinaceieuropaei* genome assembly S_erinaceieuropaei
mirc_15	ugacuagauccauacucaucu	Blast_hit	*Nippostrongylus brasiliensis* genome assembly N_brasiliensis_RM07_v1_5_4
mirc_16	gagcugcccaaugaagggcug	Blast_hit	*Apteryx australis mantelli* genome assembly AptMant0, scaffold
mirc_17	ugacuagaggacuacucaucc	Blast_hit	*Oryza sativa* Japonica Group DNA, chromosome 4, cultivar: Nipponbare,
mirc_18	ucuuuggugaucuagcuguauga	Blast_hit	PREDICTED: *Haplochromis burtoni* uncharacterized LOC102292141
mirc_19	auagguagcuuugaguccagag	Blast_hit	PREDICTED: *Fukomys damarensis* KIAA0141 ortholog (Kiaa0141), transcript

mirc_20	uauugcacuuuccccggccu	Blast_hit	PREDICTED: *Pan paniscus* HMG box domain containing 3 (HMGXB3),
mirc_21	cggacauucguucugcacgcc	Blast_hit	*Pseudomonas cremoricolorata* strain ND07, complete genome
mirc_22	ugacuagacacuuacucaucug	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 3 sequence
mirc_23	agagaauccguauggggagag	Blast_hit	*Bifidobacterium bifidum* strain BF3, complete genome
mirc_25	ccagcuaguuucccaguuuuug	Blast_hit	PREDICTED: *Colobus angolensis palliatus* monoglyceride lipase
mirc_26	uaucacaguccuaguuaccuag	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 2 sequence
mirc_27	cgugcagaaugaauguccgca	Blast_hit	PREDICTED: *Orussus abietinus* cell division cycle protein 27 homolog
mirc_28	ccauuaccuucuuccccucuu	Blast_hit	Uncultured delta proteobacterium clone Ar-cDNA-16S-P1-50 16S
mirc_29	ugacuagagucucacucaucca	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 2 sequence
mirc_31	ccguuaagucugcuguggugug	Blast_hit	*Trichobilharzia regenti* genome assembly T_regenti_v1_0_4 ,scaffold
mirc_33	ucccugagacccuuucuuguga	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome X sequence
mirc_36	ucagguacuaugugacucugca	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome X sequence
mirc_38	ugacuagagauucacacucaucca	Blast_hit	*Schistocephalus solidus* genome assembly S_solidus_NST_G2 ,scaffold
mirc_39	gugagcaaaguuucaggugugu	Blast_hit	*Schistosoma curassoni* genome assembly S_curassoni_Dakar ,scaffold
mirc_40	uaagcguauggcuuuuccccuc	Blast_hit	*Drosophila busckii* chromosome 2L sequence
mirc_42	uagcaccacaggauucagcau	Blast_hit	*Bifidobacterium catenulatum* DSM 16992 = JCM 1194 = LMG 11043
mirc_43	ccgugcuagacgaucguaagug	Blast_hit	*Echinostoma caproni* genome assembly E_caproni_Egypt ,scaffold
mirc_44	uuuugauuguugcucagaaggcc	Blast_hit	*Drosophila busckii* chromosome 3L sequence
mirc_45	uucguugucgucgaaaccugca	Blast_hit	*Drosophila busckii* chromosome X sequence
mirc_46	auguagguguguuacucccac	Blast_hit	*Drosophila melanogaster* chromosome X
mirc_47	uagcaccaguggauucagcaug	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 2 sequence
mirc_49	cggaaaaagauucacucgcagg	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 24 sequence
mirc_50	uaucacaguguuaguuaccuuca	Blast_hit	*Sparus aurata* clone contig01026 genomic sequence
mirc_51	uugagcaaagcuucagggguuu	Blast_hit	*Nomascus siki* isolate J03 mitochondrial sequence
mirc_56	ucggugggauuaucguccguu	Blast_hit	*Dichelobacter nodosus* VCS1703A, complete

			genome
mirc_58	aaauaucagcugguaaauuugg	Blast_hit	*Schistocephalus solidus* genome assembly S_solidus_NST_G2 ,scaffold
mirc_59	guucgagucucggugggac	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 1 sequence
mirc_60	uuuacgaccuucuagcacggu	Blast_hit	*Schistocephalus solidus* genome assembly S_solidus_NST_G2 ,scaffold
mirc_61	uagccucauuaucaguguuaca	Blast_hit	*Homo sapiens*, clone RP11-74D11, complete sequence
mirc_63	uagcaccaugugaauucaguaca	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 12 sequence
mirc_64	uuuacgaccauuuagcacggu	Blast_hit	*Aggregatibacter aphrophilus* strain W10433, complete genome
mirc_65	ugggcgcccgacaggugcaugc	Blast_hit	*Bifidobacterium bifidum* strain BF3, complete genome
mirc_67	cuaaguaguagugccgcagguaa	Blast_hit	*Onchocerca flexuosa* genome assembly O_flexuosa_Cordoba ,scaffold
mirc_68	uuuauauucuucuacuguccu	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 2 sequence
mirc_69	cuucuucuuguucuucuucuac	Blast_hit	PREDICTED: *Cicer arietinum* protein LST8 homolog (LOC101511790),
mirc_70	cuccgugauacaguugaggcug	Blast_hit	*Heligmosomoides polygyrus* genome assembly H_bakeri_Edinburgh
mirc_71	aacgucacgucgccggcagacu	Blast_hit	*Capsaspora owczarzaki* ATCC 30864 transmembrane protein mRNA
mirc_72	auccgugguucagugguagaauuc	Blast_hit	*Archaeon* GW2011_AR10, complete genome
mirc_73	ucuuugguauuaccaggaugcaug	Blast_hit	*Mus musculus* BAC clone RP24-132O18 from chromosome 15, complete
mirc_74	ugugacuaggguaccuguuacc	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 10 sequence
mirc_75	gaaacgaagggcuugucau	Blast_hit	*Gongylonema pulchrum* genome assembly G_pulchrum_Hokkaido ,scaffold
mirc_76	cguugaguaccguucguucuuc	Blast_hit	*Trichobilharzia regenti* genome assembly T_regenti_v1_0_4 ,scaffold
mirc_77	cuccgugauacaguugaggcag	Blast_hit	*Heligmosomoides polygyrus* genome assembly H_bakeri_Edinburgh
mirc_78	aggcaagacuccggcguagcug	Blast_hit	*Anabaena* sp. 90 chromosome chANA01, complete sequence
mirc_79	auuuuagucucuauggucagac	Blast_hit	*Cyprinus carpio* genome assembly common carp genome ,scaffold
mirc_81	aaaaugcccguggugauaauug	Blast_hit	*Cyprinus carpio* genome assembly common carp genome, scaffold
mirc_82	gacgggacgaucucaacacuau	Blast_hit	PREDICTED: *Nelumbo nucifera* uncharacterized LOC104591445 (LOC104591445),

mirc_83	ucaucaaggguguuuugccacu	Blast_hit	TPA_asm: *Oryzias latipes* strain Hd-rR, complete genome assembly,
mirc_84	cuaccagaucgaauagccucgug	Blast_hit	*Pseudoplusia includens* SNPV IE, complete genome
mirc_85	aaugggcggucuccuaacacc	Blast_hit	*Rasamsonia emersonii* CBS 393.64 MAP kinase kinase kinase (Bck1)
mirc_86	gugguguuaguagaacuugua	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 1 sequence
mirc_87	aagcgacuucgaagaaaaaccc	Blast_hit	*Bacillus* sp. FJAT-18017 genome
mirc_88	cgcacuggcccucccucugaccu	Blast_hit	PREDICTED: *Astyanax mexicanus* paired related homeobox 1 (prrx1),
mirc_89	cggaaaaagauucacucccau	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome X sequence
mirc_90	uuguguacucugcuuuguuaca	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 10 sequence
mirc_91	guaguaguaguaguaguaguagua	Blast_hit	*Oryza sativa* Japonica Group DNA, chromosome 4, cultivar: Nipponbare,
mirc_92	uuucagucauuuguccgcggac	Blast_hit	*Echinostoma caproni* genome assembly E_caproni_Egypt ,scaffold
mirc_93	cuaccuaaucaucaccccuauc	Blast_hit	*Stereum hirsutum* FP-91666 SS1 hypothetical protein (STEHIDRAFT_128450),
mirc_94	ucgaguaaauggcgguuauaugu	Blast_hit	*Apteryx australis mantelli* genome assembly AptMant0, scaffold
mirc_96	uguggcaugguacugacucacu	Blast_hit	*Moraxella catarrhalis* strain 25239, complete genome
mirc_97	cucccccuuccuacgacccaucc	Blast_hit	PREDICTED: *Apteryx australis mantelli* nuclear factor I/B (NFIB),
mirc_98	ucgcuccggcuucccccaugg	Blast_hit	*Apteryx australis mantelli* genome assembly AptMant0, scaffold
mirc_99	guaguaguaguaguaguagua	Blast_hit	*Oryza sativa* Japonica Group DNA, chromosome 4, cultivar: Nipponbare,
mirc_100	uacuuacaacacccuuaggcuc	Blast_hit	PREDICTED: *Xenopus* (Silurana) *tropicalis* acyl-CoA synthetase
mirc_101	uacccaaggaugcucuagaacu	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 9 sequence
mirc_102	uguguacuguacccauaucgac	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 11 sequence
mirc_103	ucgcuccucguccuugguucac	Blast_hit	*Setosphaeria turcica* Et28A hypothetical protein mRNA
mirc_104	ccacuagauggaucaugcuug	Blast_hit	*Spirometra erinaceieuropaei* genome assembly S_erinaceieuropaei
mirc_105	cuaccagaucgagcagccuug	Blast_hit	*Sulfuricella denitrificans* skB26 DNA, complete genome
mirc_106	aaguuuggacuuacauucguag	Blast_hit	*Ovis canadensis canadensis* isolate 43U

			chromosome 3 sequence
mirc_107	ucaagcauccuuauguucucgcu	Blast_hit	*Synechococcus* sp. CC9902, complete genome
mirc_108	ccgugcuaggcggucguaagug	Blast_hit	*Oryza sativa* Japonica Group DNA, chromosome 4, cultivar: Nipponbare,
mirc_109	aucuugaucguuuugcaaaaug	Blast_hit	*Nitrosomonas europaea* ATCC 19718, complete genome
mirc_110	ugaauggauguguugaaaaacg	Blast_hit	*Trifolium pratense* genome assembly redclover, chromosome : chr4
mirc_111	uucccgcccgcccaugcccccu	Blast_hit	*Rhodosporidium toruloides* strain CECT1137, genomic scaffold,
mirc_112	cccguggagucgcuuuaaacug	Blast_hit	*Oryza sativa* Japonica Group DNA, chromosome 1, cultivar: Nipponbare,
mirc_113	uuauugcuaucguugguaccu	Blast_hit	*Lactobacillus helveticus* strain CAUH18, complete genome
mirc_114	uaaaauucgcguuacguaagac	Blast_hit	*Spirometra erinaceieuropaei* genome assembly S_erinaceieuropaei
mirc_115	ugcuucaggaacuaugcccu	Blast_hit	*Batrachochytrium dendrobatidis* JAM81 hypothetical protein (BATDEDRAFT_28984),
mirc_116	cucccccuccccuaccuccucc	Blast_hit	PREDICTED: *Chrysochloris asiatica* deltex 3, E3 ubiquitin ligase
mirc_117	uggacgggacgaucucaacacu	Blast_hit	PREDICTED: *Nelumbo nucifera* uncharacterized LOC104591445 (LOC104591445),
mirc_118	agcaugcaccugucgggcgccc	Blast_hit	*Bifidobacterium bifidum* strain BF3, complete genome
mirc_119	uaccuaacaucgaaaggcaccg	Blast_hit	PREDICTED: *Salmo salar* synaptic vesicle glycoprotein 2C-like
mirc_121	gcuaccagaucgagcagucuug	Blast_hit	*Singulisphaera acidiphila* DSM 18658, complete genome
mirc_122	cuaccagaucgaacagccuugu	Blast_hit	*Pseudomonas cremoricolorata* strain ND07, complete genome
mirc_123	uucuccaguagccuguuaggua	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 3 sequence
mirc_124	ucacaaccuccuugagugagu	Blast_hit	*Drosophila busckii* chromosome 2R sequence
mirc_126	uugguccccuucaaccagcugu	Blast_hit	*Drosophila busckii* chromosome 2L sequence
mirc_127	aggacucuagaugaaugacacac	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 24 sequence
mirc_128	uugcuaccagaucgagaagccu	Blast_hit	*Bordetella pertussis* strain B3621, complete genome
mirc_129	caacugccucgcacucugccgu	Blast_hit	PREDICTED: *Zea mays* uncharacterized LOC100193635 (LOC100193635),
mirc_130	ugccggagguggaaggcgccg	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 3 sequence
mirc_131	ggcagcgacgucgugaggg	Blast_hit	*Stenotrophomonas maltophilia* D457 complete genome

mirc_132	ugucuuuuucugcuuugcug	Blast_hit	*Burkholderia* sp. HB1 chromosome 1, complete sequence
mirc_133	cacugucauggaagaaguccu	Blast_hit	*Apteryx australis mantelli* genome assembly AptMant0, scaffold
mirc_134	gaggacagguguuaaggggacu	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 9 sequence
mirc_135	acagaggagccuggcucucgg	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome X sequence
mirc_136	ccgacgagucuucaguugggu	Blast_hit	*Spirometra erinaceieuropaei* genome assembly S_erinaceieuropaei
mirc_137	uugcauagucacaaaagugaug	Blast_hit	*Onchocerca flexuosa* genome assembly O_flexuosa_Cordoba ,scaffold
mirc_138	uaaucucacaagguaaagcug	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 3 sequence
mirc_139	ucugcacugucaagacaacuug	Blast_hit	PREDICTED*: Zonotrichia albicollis* LanC lantibiotic synthetase
mirc_140	ucagaagugagaacacaaagcu	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 16 sequence
mirc_141	acgcacugagcugagcacaccu	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 18 sequence
mirc_142	gaggucugcaggcuuugcugug	Blast_hit	*Pan troglodytes* chromosome 22 clone:PTB-017A17, map 22, complete
mirc_143	cuauacaccaaaagauaugcccu	Blast_hit	*Apteryx australis mantelli* genome assembly AptMant0, scaffold
mirc_144	ugcaugugaggucugcagacu	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 5 sequence
mirc_145	uagcacucaggacuguuuucuc	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 16 sequence
mirc_147	uuuguucgccuggcucagucg	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 1 sequence
mirc_148	ccagacagccauagagau	Blast_hit	*Drosophila busckii* chromosome 3L sequence
mirc_149	ugacuagauccaacacucaucug	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome X sequence
mirc_150	acccggaucagcuuugcccu	Blast_hit	*Scylla paramamosain* map kinase-interacting serine/threonine mRNA,
mirc_151	augaggcagucgcggcacg	Blast_hit	*Zea mays* LOC100284192 (umc1302), mRNA
mirc_152	ugggaacuugcagacagucucc	Blast_hit	*Endocarpon pusillum* Z07020 hypothetical protein mRNA
mirc_153	cccuggagaagugaauaucugagg	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 2 sequence
mirc_154	uaaaugcauugucugguaugucau	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 1 sequence
mirc_156	ucgaauccuguccgcagcgca	Blast_hit	*Drosophila busckii* chromosome 2R sequence
mirc_157	uauagaucugauauggcguguau	Blast_hit	*Apteryx australis mantelli* genome assembly

			AptMant0, scaffold
mirc_158	guggcaccuguuagagccuugguac	Blast_hit	*Homo sapiens* solute carrier family 25 (mitochondrial carrier;
mirc_159	auguuuuuaguguucguccauu	Blast_hit	*Oryza sativa* Japonica Group DNA, chromosome 1, cultivar: Nipponbare,
mirc_160	uagguccaaagaguuaagagu	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 1 sequence
mirc_161	uggcaagaauuccugagcacaa	Blast_hit	*Shewanella putrefaciens* 200, complete genome
mirc_162	uagcaccauuugaaaucagu	Blast_hit	PREDICTED: *Thamnophis sirtalis* uncharacterized LOC106543447 (LOC106543447),
mirc_163	uuaauccaauguuguggucugga	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 3 sequence
mirc_165	ucaucaccgucgccauuggca	Blast_hit	*Mesocestoides corti* genome assembly M_corti_Specht_Voge ,scaffold
mirc_166	uuuguucguuuuacuugggau	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 5 sequence
mirc_167	gagguguucggcgauugc	Blast_hit	*Neisseria gonorrhoeae* strain 35/02, complete genome
mirc_168	guagcacaccuguagacccaaccu	Blast_hit	*Chlorocebus aethiops* BAC clone CH252-485N20 from chromosome 16,
mirc_169	aagacgacugccguuugcucgu	Blast_hit	*Heligmosomoides polygyrus* genome assembly H_bakeri_Edinburgh
mirc_170	aguuggcgcgcauaaagccgug	Blast_hit	*Deinococcus swuensis* strain DY59, complete genome
mirc_171	auuguuaaguuugguguagaguu	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome X sequence
mirc_172	uuagguuagguuagguuag	Blast_hit	*Drosophila busckii* chromosome 2R sequence
mirc_173	cggaugccacuggucagcug	Blast_hit	*Drosophila busckii* chromosome 2R sequence
mirc_174	caaggcuggagauuucgu	Blast_hit	*Coccidioides immitis* RS NACHT and WD repeat protein partial mRNA
mirc_175	ugaaacagugcauucucuccu	Blast_hit	*Cyprinus carpio* genome assembly common carp genome, scaffold
mirc_176	cuugaggacaauuuucacugaac	Blast_hit	*Cyprinus carpio* genome assembly common carp genome, scaffold
mirc_178	cuucggcaauagaaaacgguua	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 13 sequence
mirc_179	ugggcgguuugguucauu	Blast_hit	*Confluentimicrobium* sp. EMB200-NS6, complete genome
mirc_180	auuugagaaaggcugucc	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 1 sequence
mirc_181	uuacuaagcaucauggucuggaca	Blast_hit	Rat DNA sequence from clone bRB-337A12, complete sequence
mirc_182	agagaucucggguucgaucccc	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome 7 sequence

mirc_183	gaauggcagugaggcugu	Blast_hit	*Ovis canadensis canadensis* isolate 43U chromosome X sequence
mirc_184	ucaguuggcagagcgacg	Blast_hit	*Olsenella* sp. oral taxon 807 strain F0089, complete genome
mirc_185	uaaucgaaucggacuaaccccc	Blast_hit	*Drosophila melanogaster* chromosome X

4.9 References

1. **Global Invasive Species Database** [http://www.issg.org/database]

2. Leignel V, Stillman JH, Baringou S, Thabet R, Metais I: **Overview on the European green crab *Carcinus* spp. (Portunidae, Decapoda), one of the most famous marine invaders and ecotoxicological models**. *Environmental Science and Pollution Research* 2014, **21**(15):9129-9144.

3. Lafferty KD, Kuris AM: **Biological Control of marine Pests**. *Ecology* 1996, **77**(7):1989-2000.

4. Pacific Coast Shellfish Growers Association: **Shellfish economy: treasures of the tidelands.** *Shellfish economy* 2003.

5. Rodrigues ET, Pardal MA: **The crab *Carcinus maenas* as a suitable experimental model in ecotoxicology**. *Environment International* 2014, **70C**:158-182.

6. Ghedira J, Chicano-Gálvez E, Fernández-Cisnal R, Jebali J, Banni M, Chouba L, Boussetta H, López-Barea J, Alhama J: **Using environmental proteomics to assess pollutant response of *Carcinus maenas* along the Tunisian coast**. *Science of The Total Environment* 2016, **541**:109-118.

7. Keane RM, Crawley MJ: **Exotic plant invasions and the enemy release hypothesis**. *Trends in Ecology & Evolution* 2002, **17**(4):164-170.

8. Hamilton KM, Shaw PW, Morritt D: **Physiological responses of three crustacean species to infection by the dinoflagellate-like protist *Hematodinium* (Alveolata: Syndinea)**. *Journal of Invertebrate Pathology* 2010, **105**(2):194-196.

9. Bateman KS, Tew I, French C, Hicks RJ, Martin P, Munro J, Stentiford GD: **Susceptibility to infection and pathogenicity of White Spot Disease (WSD) in non-**

model crustacean host taxa from temperate regions. *Journal of Invertebrate Pathology* 2012, **110**(3):340-351.

10. Stentiford GD, Neil DM, Peeler EJ, Shields JD, Small HJ, Flegel TW, Vlak JM, Jones B, Morado F, Moss S *et al*: **Disease will limit future food supply from the global crustacean fishery and aquaculture sectors**. *Journal of Invertebrate Pathology* 2012, **110**(2):141-157.

11. Verbruggen B, Bickley L, van Aerle R, Bateman K, Stentiford G, Santos E, Tyler C: **Molecular Mechanisms of White Spot Syndrome Virus Infection and Perspectives on Treatments**. *Viruses* 2016, **8**(1):23.

12. Verbruggen B, Bickley LK, Santos EM, Tyler CR, Stentiford GD, Bateman KS, van Aerle R: **De novo assembly of the *Carcinus maenas* transcriptome and characterization of innate immune system pathways**. *BMC Genomics* 2015, **16**:458.

13. Kingsolver MB, Huang Z, Hardy RW: **Insect antiviral innate immunity: pathways, effectors, and connections**. *Journal of Molecular Biology* 2013, **425**(24):4921-4936.

14. Armitage SA, Peuss R, Kurtz J: ***Dscam* and pancrustacean immune memory - A review of the evidence**. *Developmental and Comparative Immunology* 2014.

15. **Fastqc. a quality control tool for high throughput sequence data** [http://www.bioinformatics.babraham.ac.uk/projects/fastqc]

16. Bolger AM, Lohse M, Usadel B: **Trimmomatic: A flexible trimmer for Illumina Sequence Data**. *Bioinformatics* 2014.

17. Leggett RM, Clavijo BJ, Clissold L, Clark MD, Caccamo M: **NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries**. *Bioinformatics* 2014, **30**(4):566-568.

18. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S *et al*: **High-quality draft assemblies of mammalian genomes from massively parallel sequence data**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**(4):1513-1518.

19. Ribeiro FJ, Przybylski D, Yin S, Sharpe T, Gnerre S, Abouelleil A, Berlin AM, Montmayeur A, Shea TP, Walker BJ *et al*: **Finished bacterial genomes from shotgun sequence data**. *Genome research* 2012, **22**(11):2270-2277.

20. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H *et al*: **Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads**. *Genome Research* 2014, **24**(8):1384-1395.

21. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD *et al*: **SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing.** *Journal of Computational Biology* 2012, **19**(5):455-477.

22. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y *et al*: **SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler**. *GigaScience* 2012, **1**:18-18.

23. Sahlin K, Vezzi F, Nystedt B, Lundeberg J, Arvestad L: **BESST - Efficient scaffolding of large fragmented assemblies**. *BMC Bioinformatics* 2014, **15**(1):281.

24. Gurevich A, Saveliev V, Vyahhi N, Tesler G: **QUAST: quality assessment tool for genome assemblies**. *Bioinformatics* 2013, **29**(8):1072-1075.

25. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nature Methods* 2012, **9**(4):357-359.

26. Parra G, Bradnam K, Korf I: **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes**. *Bioinformatics* 2007, **23**(9):1061-1067.

27. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs**. *Bioinformatics* 2015, **31**(19):3210-3212.

28. Smit A, Hubley R, Green P: **RepeatMasker Open-4.0.** 2013-2015.

29. Birney, E., Clamp, M. and Durbin, R., 2004**. GeneWise and genomewise**. *Genome Research*, 14(5):988-995.

30. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M: **Gene identification in novel eukaryotic genomes by self-training algorithm**. *Nucleic Acids Research* 2005, **33**(20):6494-6506.

31. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith Jr RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD *et al*: **Improving the Arabidopsis genome annotation**

using maximal transcript alignment assemblies. *Nucleic Acids Research* 2003, **31**(19):5654-5666.

32. Stanke M, Waack S: **Gene prediction with a hidden Markov model and a new intron submodel**. *Bioinformatics* 2003, **19**(suppl 2):ii215-ii225.

33. Greifswald B: **Incorporating Illumina RNAseq into AUGUSTUS with Tophat**. 2014.

34. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR: **Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments**. *Genome Biology* 2008, **9**(1):R7.

35. Fromm B, Worren MM, Hahn C, Hovig E, Bachmann L: **Substantial loss of conserved and gain of novel microRNA families in flatworms**. *Molecular Biology and Evolution* 2013, 30(12):2619-2628

36. Sandve GK, Gundersen S, Johansen M, Glad IK, Gunathasan K, Holden L, Holden M, Liestøl K, Nygård S, Nygaard V *et al*: **The Genomic HyperBrowser: an analysis web server for genome-scale data**. *Nucleic Acids Research* 2013, **41**(Web Server issue):W133-W141.

37. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS: **MicroRNA targets in Drosophila**. *Genome Biology* 2004, **5**(1):R1-R1.

38. Thadani R, Tammi MT: **MicroTar: predicting microRNA targets from RNA duplexes**. *BMC Bioinformatics* 2006, **7 Suppl 5**:S20.

39. Bethesda (MD): National Library of Medicine (US) NCfBI: **Chapter 18, The Reference Sequence (RefSeq) Project.** . *The NCBI handbook* 2002.

40. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics* 2010, **26**(6):841-842.

41. Buchfink B, Xie C, Huson DH: **Fast and sensitive protein alignment using DIAMOND**. *Nature Methods* 2015, **12**(1):59-60.

42. Li H, Durbin R: **Fast and accurate short read alignment with Burrows–Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.

43. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer**. *Nature Biotechnology* 2011, **29**(1):24-26.

44. Kent WJ: **BLAT—The BLAST-Like Alignment Tool**. *Genome Research* 2002, **12**(4):656-664.

45. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J *et al*: **Pfam: the protein families database**. *Nucleic Acids Research* 2014, **42**(Database issue):D222-230.

46. Bonnivard E, Catrice O, Ravaux J, Brown SC, Higuet D: **Survey of genome size in 28 hydrothermal vent species covering 10 families**. *Genome* 2009, **52**(6):524-536.

47. Gregory TR: **Animal Genome Size Database.** 2016. http://www.genomesize.com/

48. Song L, Bian C, Luo Y, Wang L, You X, Li J, Qiu Y, Ma X, Zhu Z, Ma L *et al*: **Draft genome of the Chinese mitten crab, *Eriocheir sinensis***. *Gigascience* 2016, **5**:5.

49. Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK *et al*: **The Ecoresponsive Genome of *Daphnia pulex***. *Science (New York, NY)* 2011, **331**(6017):555-561.

50. Wang S, Lorenzen MD, Beeman RW, Brown SJ: **Analysis of repetitive DNA distribution patterns in the *Tribolium castaneum* genome**. *Genome Biology* 2008, **9**(3):R61.

51. Pozzoli U, Menozzi G, Fumagalli M, Cereda M, Comi GP, Cagliani R, Bresolin N, Sironi M: **Both selective and neutral processes drive GC content evolution in the human genome**. *BMC Evolutionary Biology* 2008, **8**:99.

52. Kenny NJ, Sin YW, Shen X, Zhe Q, Wang W, Chan TF, Tobe SS, Shimeld SM, Chu KH, Hui JH: **Genomic sequence and experimental tractability of a new decapod shrimp model, *Neocaridina denticulata***. *Marine Drugs* 2014, **12**(3):1419-1437.

53. Zhang JZ: **Protein-length distributions for the three domains of life**. *Trends in Genetics* 2000, **16**(3):107-109.

54. Chen Y-C, Liu T, Yu C-H, Chiang T-Y, Hwang C-C: **Effects of GC Bias in Next-Generation-Sequencing Data on *De Novo* Genome Assembly**. *PloS One* 2013, **8**(4):e62856.

55. Treangen TJ, Salzberg SL: **Repetitive DNA and next-generation sequencing: computational challenges and solutions**. *Nature Reviews Genetics* 2012, **13**(1):36-46.

56. Rhoads A, Au KF: **PacBio Sequencing and Its Applications**. *Genomics, Proteomics & Boinformatics* 2015, **13**(5):278-289.

57. Piednoel M, Donnart T, Esnault C, Graca P, Higuet D, Bonnivard E: **LTR-retrotransposons in *R. exoculata* and other crustaceans: the outstanding success of GalEa-like copia elements**. *PloS One* 2013, **8**(3):e57675.

58. Ambros V: **The functions of animal microRNAs**. *Nature* 2004, **431**(7006):350-355.

59. Ritchie W, Rasko JEJ: **Refining microRNA target predictions: Sorting the wheat from the chaff**. *Biochemical and Biophysical Research Communications* 2014, **445**(4):780-784.

60. Aiewsakun P, Katzourakis A: **Endogenous viruses: Connecting recent and ancient viral evolution**. *Virology* 2015, **479–480**:26-37.

61. Katzourakis A, Gifford RJ: **Endogenous Viral Elements in Animal Genomes**. *PLoS Genet* 2010, **6**(11):e1001191.

62. Rusaini, La Fauce KA, Elliman J, Bowater RO, Owens L: **Endogenous Brevidensovirus-like elements in *Cherax quadricarinatus*: Friend or foe?** *Aquaculture* 2013, **396-399**:136-145.

63. Rozenberg A, Brand P, Rivera N, Leese F, Schubart CD: **Characterization of fossilized relatives of the White Spot Syndrome Virus in genomes of decapod crustaceans**. *BMC Evolutionary Biology* 2015, **15**:142.

64. Ng TH, Chiang YA, Yeh YC, Wang HC: **Review of Dscam-mediated immunity in shrimp and other arthropods**. *Developmental and Comparative Immunology* 2014.46(2):129-138

65. Powell A, Pope EC, Eddy FE, Roberts EC, Shields RJ, Francis MJ, Smith P, Topps S, Reid J, Rowley AF: **Enhanced immune defences in Pacific white shrimp (*Litopenaeus vannamei*) post-exposure to a vibrio vaccine**. *Journal of Invertebrate Pathology* 2011, **107**(2):95-99.

66. Johnson KN, van Hulten MCW, Barnes AC: **“Vaccination” of shrimp against viral pathogens: Phenomenology and underlying mechanisms**. *Vaccine* 2008, **26**(38):4885-4892.

67. Sadd BM, Schmid-Hempel P: **Insect Immunity Shows Specificity in Protection upon Secondary Pathogen Exposure**. *Current Biology* 2006, **16**(12):1206-1210.

68. Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE, Zipursky SL: ***Drosophila Dscam*** **Is an Axon Guidance Receptor Exhibiting Extraordinary Molecular Diversity**. *Cell* 2000, **101**(6):671-684.

69. Schmucker D, Chen B: ***Dscam*** **and** ***DSCAM*****: complex genes in simple animals, complex animals yet simple genes**. *Genes & Development* 2009, **23**(2):147-156.

70. Smith PH, Mwangi JM, Afrane YA, Yan G, Obbard DJ, Ranford-Cartwright LC, Little TJ: **Alternative splicing of the** ***Anopheles gambiae Dscam*** **gene in diverse** ***Plasmodium falciparum*** **infections**. *Malaria Journal* 2011, **10**(1):1-7.

Chapter 5

Sequencing and *De novo* assembly of the *Homarus gammarus* transcriptome and characterization of its immune system.

Supplementary material available in the 'Chapter 5' folder on the DVD

Chapter 5: Sequencing and *De novo* assembly of the *Homarus gammarus* transcriptome and characterization of its immune system.

5.1 Abstract

The European lobster (*Homarus gammarus*) is a prominent aquatic crustacean of economic importance across European oceans and seas. Its meat is a highly valued commodity and therefore the species is commercially fished, particularly in the United Kingdom. Due to its commercial and ecological value, research efforts have been directed to support stocking and culture programs for this species. Understanding the critical developmental stages where mortality in culture is highest and host-pathogen interactions for this species are examples. Despite these initiatives, molecular information of *H. gammarus* is sparse, limiting our ability to conduct experiments at the genomic level. To address this need, we applied next generation sequencing technology to produce a de novo transcriptome for *H. gammarus*. A *de novo* transcriptome assembly was produced using sequence data from nine different tissues and the Trinity pipeline, resulting in 106,498 transcripts. The transcripts were filtered and subsequently annotated using a variety of tools (including BLAST, MEGAN and RSEM) and databases (including NCBI, Gene Ontology and KEGG). Similar to transcriptomes produced for other aquatic crustacea, around 20-25 % of transcripts could be annotated. This relatively low proportion of annotated transcripts is associated with the lack of genomic resources for aquatic crustacea. The White Spot Syndrome Virus has the ability to infect all tested aquatic crustacea but displays varying degrees of pathogenicity. Comparisons between the immune system of a relatively resistant species (*Carcinus maenas*) and the more susceptible *H. gammarus* showed little differences, suggesting that the differences in susceptibility to WSSV between these two species may be linked with sequence variation in viral receptors, including Rab7, rather than structural differences in their immune system.

5.2 Introduction

The European Lobster (*Homarus gammarus*) is a much valued seafood commodity. The flavour of its meat is held in high regard by consumers, enabling this species to yield substantial prices on the market [1]. It is commercially fished across Europe, yielding an average of 4972.5 tonnes between 2010 and 2013 [2]. Lobster fisheries are concentrated around the United Kingdom, which accounted for 65 % of total capture (2013) [2]. Because of commercial and ecological interests there are numerous research initiatives with the objective of supporting the lobster sector, ranging from optimization of stocking programs [3] to development of land-based larvae-to-plate culture systems [4]. The production of lobsters (from juveniles for restocking to fully grown animals) is complicated by cannibalistic behaviour, operating costs and density [4]. Despite these issues successes have been achieved and commercial lobster farms are operational, but are thus far not widespread throughout the world.

A major factor that impacts population dynamics, fisheries and ecology of several lobster species is the rise in lobster diseases [5]. Lobsters are faced with several significant pathogens, both in wild populations or post-capture holding facilities [5-8]. Parasitic marine plankton species like the dinoflagellate *Hematodinium perezi* can infect lobsters and cause the meat to acquire a bitter taste, resulting and an unmarketable product, or in worst case mortality [5, 9]. The *Panulirus argus* Virus (PaV1) can cause significant mortalities in juveniles of the Caribbean spiny lobster (*Panulirus argus*) [10]. Bacterial pathogens like *Aerococcus viridians*, which causes gaffkaemia, and *Aquimarina homaria*, which potentially causes epizootic shell disease, have had major impact on American lobster (*Homarus americanus*) fisheries and there have been indications of prevalence amongst European lobsters [11]. The susceptibility of *H. gammarus* to White Spot Syndrome Virus (WSSV), the most devastating virus in crustacean aquaculture, constitutes a threat to management and aquaculture of this species [12]. Bateman *et al.* 2012 showed that *H. gammarus* can develop WSSV infections after feeding with WSSV-positive supermarket-derived shrimp [13]. Research on molecular interactions between the lobster and its pathogens can aid in management of the diseases affecting lobster fisheries and aquaculture, improving sustainability of these actives while protecting lobster populations in the marine ecosystem.

Molecular research on *H. gammarus*, and other lobster species, has been scarce. One of the major hurdles is the lack of genomic information in public databases, including DNA and protein sequences. However, with development of next-generation sequencing technologies it is now possible to generate such data at affordable cost. We used Illumina HiSeq 2500 RNA-sequencing to sequence, assemble and quantify the lobster transcripts expressed across nine tissue types. This dataset will be an excellent platform for future studies on *H. gammarus* facilitating the molecular analysis of essential processes including development, growth and response to disease. We have investigated the immune system of *H. gammarus* in order to infer its potential responses to pathogens. In previous work we applied Illumina HiSeq 2500 RNA-sequencing to assemble a transcriptome for the shore crab *Carcinus maenas* [14]. Bateman *et al.* 2012 showed that *C. maenas* is relatively resistant to WSSV injection compared to other crustaceans, including *H. gammarus* [12]. We aimed to investigate whether the variation in disease susceptibility could be attributed to differences in the immune system of both *H. gammarus* and *C. maenas*. Therefore, we characterized and compared the components of the immune system and known WSSV interacting molecules for both species. Such a description of immune system transcripts should also aid identification of host-pathogen interactions between *H. gammarus* and other pathogens.

5.3 Methods

5.3.1 Animals and tissues – Work performed by colleagues at Cefas

Ten individual lobsters (*Homarus gammarus*) were obtained from a lobster vivier in Weymouth, UK (landed and caught from the surrounding area) and placed on ice prior to dissecting tissues and organs of interest (including heart, muscle, hepatopancreas, nerve, eye, gut, gill, testis and ovary). All tissues and organs were immediately snap-frozen in liquid nitrogen prior to sample preparation and analysis.

All tissues were disrupted by grinding frozen tissue fragments with liquid nitrogen before homogenisation with a rotor stator homogeniser in lysis reagent. RNA was extracted using Qiagen’s mRNeasy mini kit, with on column DNase digestion, according to the manufacturer’s instructions. RNA quality was measured using an Agilent 2100 Bioanalyzer with RNA 6000 nano kit (Agilent Technologies, CA, USA). cDNA libraries for each individual were constructed using 1.0 µg of RNA. Four replicates were used for each tissue, including two males and two females (except for gonad tissue, where four same sex samples were sequenced). ERCC Spike-In control mixes (Ambion via Life Technologies, Paisley, UK) were added to control for technical variation during sample preparation and sequencing, and analysed using manufacturer’s guidelines. mRNA purification was performed via poly (A) enrichment using Tru-Seq Low Throughput protocol and reagents (Illumina, CA, USA). Finally, cDNA libraries were constructed using Illumina’s TruSeq Stranded mRNA Sample Preparation kit. Each sample was labelled with a unique barcode sequence to enable multiplexing of samples across three lanes. All libraries were diluted to 10nM and sequenced on an Illumina HiSeq 2500 with the 2 × 100 bp paired-end read module.

5.3.2 Data pre-processing and de novo transcriptome assembly

The quality of the sequencing data was assessed with FastQC v 0.10.1 [15]. After this assessment the reads were pre-processed in order to remove any remaining adapter sequences and reduce low confidence information. The paired-end libraries were trimmed by Trimmomatic v 0.32 [16] with the following settings: removal of Illumina adapters, removal of the first 12 bases, removal of the last 3 bases, removing bases with low quality based on a sliding window of 4 and minimal Phred quality of 20 and applying a minimal remaining length of 25. *De novo* assembly was performed with paired-end reads where both mates in a pair passed the quality thresholds. Sequencing reads from all tissues were combined and used for *de novo* assembly with the Trinity assembler v 2.0.6 [17]. Trinity settings were optimized based on contig count and length distribution statistics. Ultimately Trinity was run with a k-mer of 25 and requiring minimal k-mer coverage of 10. After assembly, transcripts were clustered with CD-HIT-EST v 4.6 using default parameters, to reduce redundancy [18, 19]. The number of contigs, contig lengths, N50 and other assembly statistics on the resulting assembly were derived through QUAST v 2.2 [20].

5.3.3 Transcriptome annotation

Annotation of the *de novo* transcriptome was achieved through the Trinotate suite v. 2.0.1. Within the Trinotate suite, BLASTx 2.2.28+ [21] was used to identify sequence similarities between *H. gammarus* transcripts and the UniProt/SwissProt database. The proteins encoded by the transcripts were identified through TransDecoder v.2.0.1 [22] and their sequences were compared to the UniProt/SwissProt data with BLASTp 2.2.28+ [21]. Conserved domains in the protein sequences were analysed through HMMER 3.1.b2 [23] in combination with the Pfam domain database (version 28.0) [24]. Transmembrane regions in protein sequences were predicted with TMHMM-2.0c [25] and identification of potential signalling peptides was performed through SignalP v.4.1 [26]. The output of these programs was combined into a Trinotate annotation report, applying a BLAST e-value threshold of 1e^{-5}. Outside Trinotate, the transcripts, and their corresponding predicted protein sequences, were compared to the NCBI non-redundant (nr) database using BLASTx 2.2.28+ and BLASTp 2.2.28+ with an e-value threshold of 1e^{-5}, retaining the top 10 hits. The presence of highly conserved core eukaryotic genes was tested with CEGMA v2.5 [27]. Transcripts were annotated with Gene Ontology functional annotations based on Blast2GO PRO (Sept. 2015)

[28] using the results of BLASTx to nr. Finally, taxonomic classifications of the transcripts were determined and visualized using MEGAN 5.10.4 [29], and transcripts that did not annotate to the metazoan taxon were removed from the transcriptome assembly.

5.3.4 Transcriptome expression and differential expression analysis

The reads from each individual sample were mapped to the transcriptome, including the non-metazoan transcripts in order to not alter the alignment, using bowtie2 v. 2.2.6 [30]. From this alignment the expression of transcripts was derived by RSEM 1.2.21 [31], obtaining values in transcripts per million (TPM) and Fragments Per Kilobase of transcript per Million mapped reads (FPKM). Tissue-specific differentially-expressed genes and transcripts were derived by comparing a single tissue to all other tissues, the latter were treated as biological replicates. Differential expression calculation was performed with the edgeR package v. 3.4.2 [32]. Only transcripts/genes with a minimal count of 10 TPM across all samples were considered in DE calculations. Genes/transcripts with an FDR below 0.05 were considered differentially expressed. Principal component analysis was performed on transcripts counts on $\log_{10}$(FPKM+1) values, including all transcripts accepted for DE calculations. The lists of differentially-expressed genes for each tissue were analysed for enrichment of Gene Ontology categories using Blast2GO® PRO (Sept. 2015), terms with an FDR below 0.05 were considered significantly enriched.

5.3.5 Pathway analysis and comparative genomics

KEGG ontology groups were assigned to assembled transcripts using the KEGG Automatic Annotation Server (KAAS) web service, using bi-direction best hit [33]. The same server was employed to produce the pathway visualizations to show coverage of components by *C. maenas*, *H. gammarus* or both species (*C. maenas* data from Verbruggen et al. 2015 [14]). The components of the immune system pathways were identified through a custom R script (Supplementary file S1). The script included an NCBI protein database search, a tBLASTn search to the transcriptome and subsequent result filtering based on taxonomic distance. The results of the script were examined to reduce redundancies and remove unintended results from the NCBI protein database search. Phylogenetic analysis was performed in MEGA 6.06. The protein sequences for the trees were derived from the NCBI protein database and the protein sequences of the open reading frames in *H. gammarus* and *C. maenas* (identified by TransDecoder). Multiple-sequence alignments were produced using ClustalW in MEGA6.06. Phylogenetic trees were generated using the maximum likelihood algorithm and tested with 250 bootstrap replications, a Jones-Taylor-Thornton substitution model and Nearest-Neighbor-Interchange Method for tree inference as was default in MEGA 6.06 defaults.

5.4 Results and discussion

5.4.1 Sequencing and pre-processing

To generate the *H. gammarus* transcriptome, isolated RNA from nine tissues, with four individuals representing each tissue, was sequenced on the Illumina HiSeq 2500 platform. Sequencing delivered a total of 1,295,369,286 reads across the 36 samples. After quality filtering a total of 1,170,279,158 reads were retained with an average of 32,507,754 per sample. The details of the sequencing data generated across all the samples are presented in Supplementary File S2.

5.4.2 De novo assembly

In order to generate a high quality *de novo* transcriptome for *H. gammarus*, sequencing data from all samples were combined together during assembly to improve coverage of all transcripts in this species, including those expressed only in certain tissues. After assembly the final transcriptome encompassed 106,498 transcripts with a total length of 117,379,008 bp and a GC content of 41 %. In total 31,824 transcripts had a length over 1,000 bp, ranging up to a length of 24,324 bp for the largest transcript assembled, and the N50 was calculated to be 2,648 (see Table 1).

Table 1 *De novo* assembly statistics

Assembly statistic	**Value**
Contigs (>= 200 bp)	106,498
Contigs (>= 1000 bp)	31,824
Total length	117,379,008
GC %	41.34
N50	2,648
N75	1,377
L50	11,077
L75	24,184

The presence of strongly conserved genes in a genome or transcriptome can be a good indicator of quality, and absence of a large fraction of such genes should arouse concerns about the quality of the assembly [27]. The presence of 248 conserved eukaryotic genes in the *de novo* transcriptome was assessed in CEGMA [27]. In the *H. gammarus* transcriptome 99.6 % of the conserved genes were identified, with only a single core gene missing: KOG2948 (Predicted metal-binding protein). Additionally, near-universal single-copy orthologues from OrthoDB were compared to the transcriptome using BUSCO [34]. The orthologues from the arthropod lineage, 2,675 in total, were selected for the search. Out of these, 83 % were identified completely, 22 % of which were duplicated, and 5.4 % were fragmented, and there were 303 BUSCO groups missing from the assembly (Table 2). Together these results show that the *H. gammarus* transcriptome contains nearly all core genes, strengthening confidence in its quality.

Table 2 BUSCO results

BUSCO	**Count**
Complete Single-copy BUSCOs	1624 (61 %)
Complete Duplicated BUSCOs	602 (22 %)
Fragmented BUSCOs	146 (5.4 %)
Missing BUSCOs	303 (11 %)
Total BUSCO groups searched	2675

5.4.3 Transcript annotation

The assembled transcripts were annotated at the whole transcript and predicted peptide level. Comparing the transcript sequence to the NCBI non-redundant protein database using BLASTx resulted in 25,763 annotations. Furthermore, 12,330 were linked to at least one Gene Ontology term using BLAST2GO. In total 7,293 transcripts were annotated to KEGG Orthologies based on annotations using the KAAS server [33]. Within the *H. gammarus* transcripts, 54,916 open reading frames (ORF) were identified. Within the Trinotate suite [35], the peptide sequences were screened for the presence of conserved domains, signal peptides and transmembrane regions. We found that 18,924 protein sequences contained at least one predicted Pfam domain and 7,548 had predicted transmembrane regions. Predicted signalling peptides were identified in 3,265 of the predicted open reading frames. The annotation statistics are summarised in Table 3 and the complete annotation report can be found in Supplementary File S3.

Table 3 *H. gammarus* transcript annotation statistics

Input	Annotation method	Number of annotated transcripts
Trinity transcripts	BLASTx – NCBI nr protein	25,763 (24.2 %)
Trinity transcripts	BLAST2GO	12,330 (11.6 %)
Trinity transcripts	KEGG	7,293 (6.8 %)
Trinity transcripts	TransDecoder ORF finder	54,916 (51.6 %)
TransDecoder Peptides	BLASTp – UniProt/SwissProt	12,453 (22.7 %)
TransDecoder Peptides	Pfam	18,924 (34.5 %)
TransDecoder Peptides	SignalP	3,265 (5.9 %)
TransDecoder Peptides	TMHMM	7,548 (13.7 %)

5.4.4 Taxonomy

Next generation sequencing datasets can produce contaminating sequences existing within assembled transcriptomes. For example, some of the RNA could have been isolated from bacteria, fungi and viruses within *H. gammarus* or even on the kits used in the laboratory [36]. It is thus desirable to remove transcripts with questionable origin. The BLASTx to NCBI-nr results were uploaded in MEGAN5 in order to illustrate the distribution of likely sequence origin across various taxa. Figure 1 shows a partially collapsed taxonomic tree with transcripts counts. The tree shows that most transcripts with taxonomic annotation map along the metazoan, arthropod, Crustacea lineage. There is one artificial sequence which shows that even with pre-processing a complete removal of sequence adapters was not achieved. Furthermore, there are several sequences that have a bacterial or viral origin. A reduced transcriptome was produced that encompasses only transcripts mapping to the metazoan taxa (23,815 transcripts in total) in order to remove the potential contamination by organisms living within the lobster tissues sequenced or the sequencing kits and hardware. It has to be noted that this process would remove previously unidentified sequences from *H. gammarus* that bear no resemblance to previously annotated sequences in NCBI-nr (such transcripts would be present in the ‘No Hits’ taxon in Figure 1).

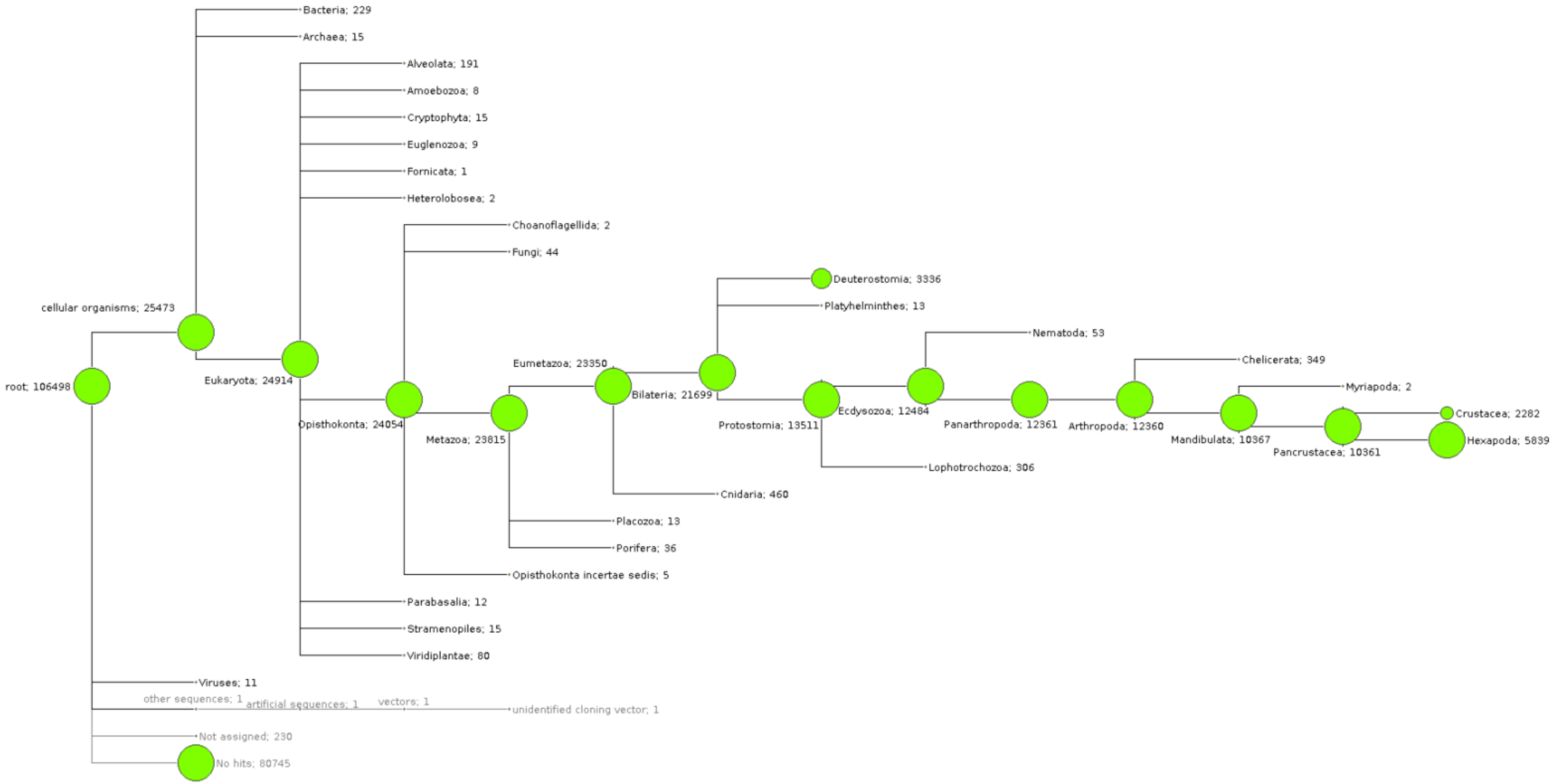


Figure 1 MEGAN5 Taxonomic tree with transcript counts. Numbers illustrate the number of transcripts representing each taxonomic group. Within the metazoan taxon, the Pancrustacea represented the largest taxonomic group

5.4.5 Tissue distribution of transcript expression

The expression of transcripts in every tissue sample was estimated through RSEM [31]. The expression of transcripts ranged between 0 to over two million FPKM for some transcripts in a tissue. Figure 2 shows the range of expression values and the squared coefficient of variation (CV^2) across expression levels. At low and high FPKM values CV^2 increases, while it is more stable at intermediate transcript expression. High variation is often observed in highly expressed transcripts and at low expressed transcripts the CV^2 calculations becomes sensitive to small variations in the mean expression. The top 10 highest expressed transcripts for every tissue are summarized in Supplementary File 3 and the highest expressed transcript in Table 4. There are several transcripts that appear amongst the highest expressed in multiple tissues. For example, the transcripts 'TR36398_c1_g1_i1' (elongation factor-1 alpha), 'TR1499_c0_g1_i1' (elongation factor 2), 'TR31500_c0_g1_i1' (clottable protein) and 'TR41301_c3_g1_i1' (actin), 'TR31683_c0_g1_i1' (polyA-binding protein) are represented in at least five tissues. The function of the proteins encoded by these transcripts can be placed in the 'housekeeping' category; therefore it is no surprise that they are found in several tissues. Transcripts that are only found in the top10 of a single tissue would be of more interest and related to the tissue function. In the gut tissue chitin binding protein is highly expressed, which is related to the presence of the peritrophic membrane, a protective structure in invertebrate midguts [37]. In the muscle tissue transcripts coding for myosin, 'TR41401_c4_g1_i2' myosin heavy chain type 1, and other muscle related proteins are predominant. However, the highest expressed transcript in this tissue type is arginine kinase, which is related to the high demand for energy. Interestingly, the arginine kinase transcript is not represented in the transcriptome filtered for metazoan sequences. The transcript 'TR4625_c0_g1_i1' has the highest sequence homology to arginine kinase of *H. gammarus* (Genbank P14208.4), but on the MEGAN5 tree it is represented on the 'Eukaryota' node and therefore not included in the metazoan transcriptome. This indicates that filtering for metazoan sequences through MEGAN5 is not optimal and *H. gammarus* transcripts can be undesirably excluded. Due to similarities in tissue function the highest expressed transcripts in the heart tissue resemble the muscle tissue. Again arginine kinase shows the highest expression, followed by other energy related transcripts like NADH dehydrogenase and muscle fibre proteins like tropomyosin.

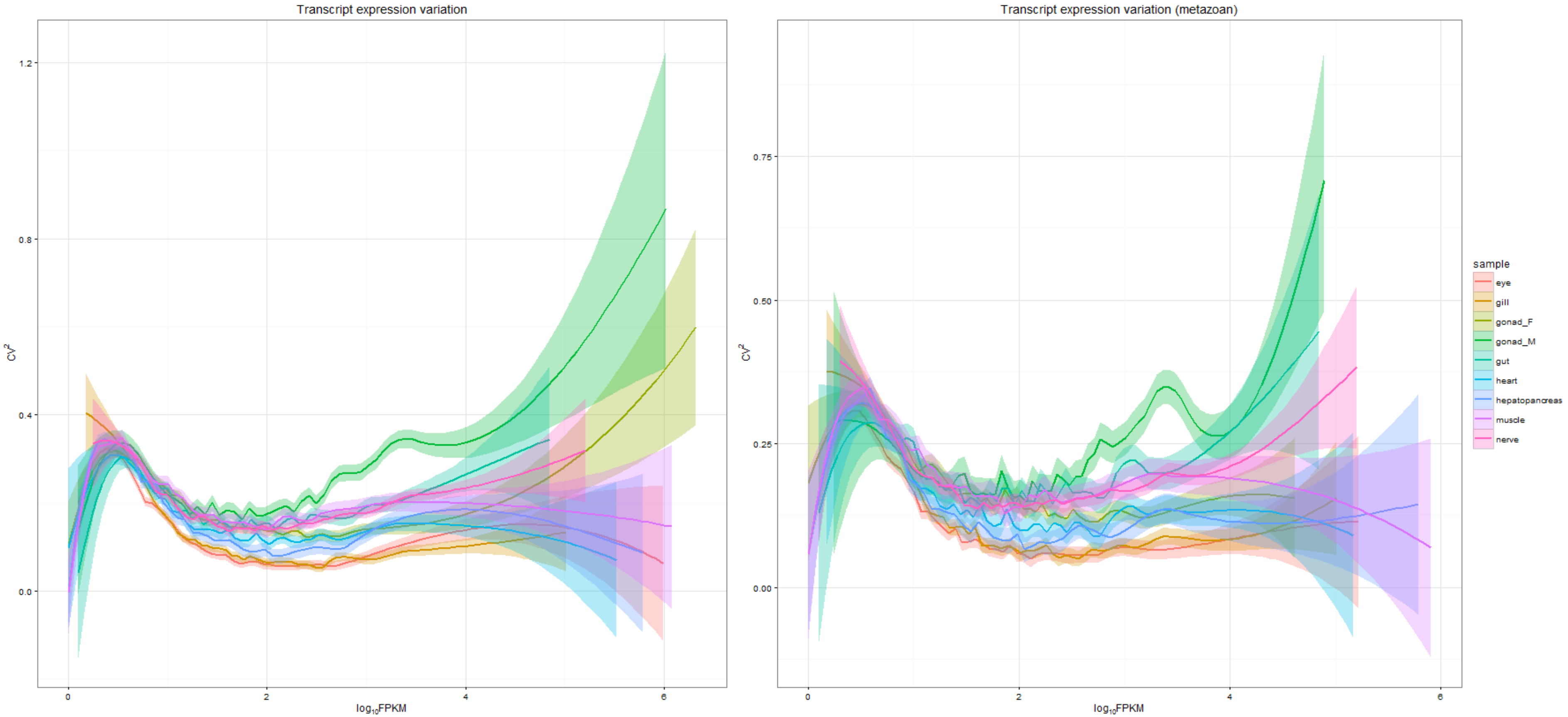


Figure 2 Transcript expression variations. Squared coefficient of variation at various levels of expression. Left: includes transcripts that are expressed in every tissue and have at least a summarized FPKM over 10 across all tissues. Right: includes transcripts of likely metazoan origin.

Table 4 Highest expressed transcript per tissue

All			
Tissue	**Transcript**	**Avg. FPKM**	**Annotation**
eye	TR18259_c0_g1_i1	984712.0	No significant similarity
gill	TR31500_c0_g1_i1	102903	Hemolymph clottable protein
gonad_F	TR20798_c2_g1_i1	2077837.25	Hypothetical protein
gonad_M	TR20798_c2_g1_i1	1046796.75	Hypothetical protein
gut	TR13506_c0_g1_i1	70281.5	Peritrophic membrane chitin binding protein
heart	TR4625_c0_g1_i1	331119.75	Arginine kinase
hepatopancreas	TR39501_c0_g1_i2	618815	Hemocyanin
muscle	TR4625_c0_g1_i1	1196705.5	Arginine kinase
nerve	TR31500_c0_g1_i1	161152.5	Hemolymph clottable protein
Metazoan			
Tissues	**Transcript**	**Avg. FPKM**	**Annotation**
eye	TR29785_c0_g1_i1	166819.75	hypothetical protein [*Macrobrachium nipponense*]
gill	TR31500_c0_g1_i1	102903	Hemolymph clottable protein
gonad_F	TR36398_c1_g1_i1	41172	Elongation factor 1A
gonad_M	TR36398_c1_g1_i1	77736.25	Elongation factor 1A
gut	TR13506_c0_g1_i1	70281.5	Chitin binding
heart	TR40601_c0_g2_i1	148572.25	NADH dehydrogenase
hepatopancreas	TR39501_c0_g1_i2	618815	Hemocyanin
muscle	TR41401_c4_g1_i2	802680.25	Myosin heavy chain
nerve	TR31500_c0_g1_i1	161152.5	Hemolymph clottable protein

5.4.6 Differential gene expression

In order to identify differentially expressed transcripts for a tissue, the samples of that tissue were compared to the remaining samples. The remaining samples represent a wide range of tissues, thus function as a background to which each tissue is compared. The number of differentially expressed transcripts in each tissue is shown in Table 5, and range from 1,114 in the eye to 11,472 in the testis. The latter number is large compared to the other samples as the second largest is hepatopancreas with around half that number of DE transcripts. There was also a higher degree of variation amongst transcripts in the testis, in particular for transcripts ranging between 1,000 and 10,000 FPKM (see Figure 2). However, investigating the distribution of the average expression of differentially-expressed transcripts in each tissue showed no peculiarities in the male gonad tissue, thus this increase in variation within any FPKM region was not responsible for the inflated DE count (Figure S 1). It is possible that the relatively inflated DE transcript count is caused by outlier samples, e.g. a particular testis sample might have been infected with a parasite which impacts expression values. The first two components of a principal component analysis (PCA) on transcript expression, plotted in Figure 3, did not show clear outliers. Samples from identical tissues all clustered together, with the exception of one of the nerve samples. The PCA plot follows the DE analysis in respect to the male gonad and hepatopancreas tissues showing the largest deviations from the other samples. The expression values of transcripts do show that the male gonad tissue had the largest number of expressed transcripts and also the largest number of exclusively expressed transcripts (over 5,000; see Table S 1). The latter are probably responsible for the large difference in total differentially expressed transcripts as compared to the other tissues.

Table 5 differentially expressed genes per tissue

Tissue	DE transcripts	DE transcripts metazoan
eye	1,114	277
gill	2,034	524
gonad_F	5,120	1,528
gonad_M	11,472	3,639
gut	1,882	313
heart	2,467	643
hepatopancreas	6,373	2,068
muscle	4,679	1,661
nerve	2,058	385

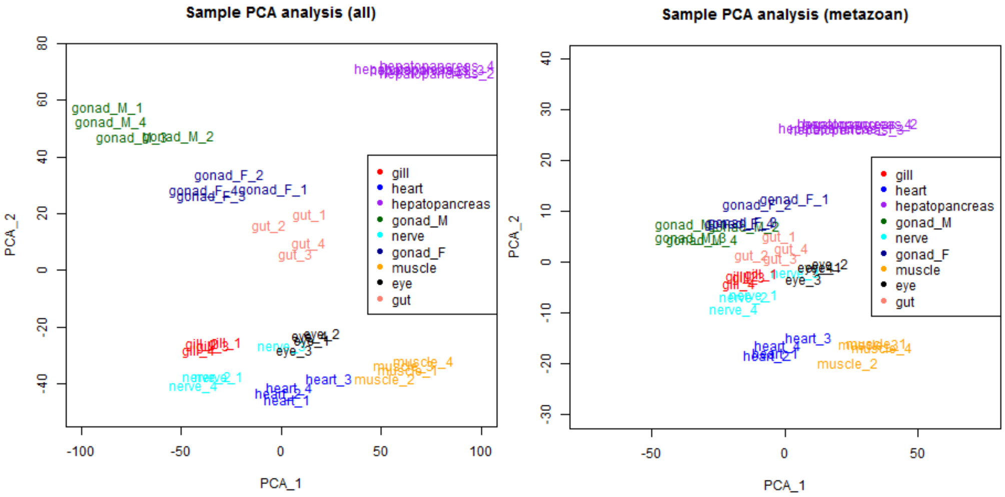


Figure 3 Principal component analysis of Lobster tissue samples. Principal component analysis was performed based on $\log_{10}$ transformed expression values. Left: all transcripts accepted for DE analysis. Right: metazoan transcripts accepted for DE analysis.

Analysis of over- or under representation of GO terms was conducted in the lists of differentially or uniquely expressed transcripts in each tissue, using BLAST2GO [28]. It was expected that the overrepresented GO terms align with the function of the tissue. The enriched GO terms are summarized in Supplementary File S5. For most tissues functionally related GO terms were identified. For example: in the eye GO terms enriched included 'phototransduction' (FDR = $3.5e^{-02}$) and 'detection of light stimulus' (FDR = $4.19e^{-02}$), as well as many terms related to programmed cell death of retinal cells. In the gill GO terms enriched related to respiration, e.g. 'respiratory system development' (FDR = $3.32e^{-06}$). Among the list of genes over-expressed or unique to female gonad tissue enriched GO terms terms included those related to metabolism and transporter activities for various compounds ('urate metabolic process', FDR $1.85e^{-02}$) which can serve to build up reserves for the embryo. In the male gonads GO terms related to cuticle (FDR $3.29e^{-10}$) and vesicular transport were identified (e.g. COPI-coated vesicle, FDR $1.72e^{-03}$) which are not related to the functions usually associated with the testis. Filtering the DE transcripts in the testis by keeping only transcripts that are also expressed in other tissues, results in enrichment of terms related to retrograde vesicle transport (COPI-coated vesicles), thus similar to the original list of DE transcripts. Vice versa, keeping only DE transcripts uniquely expressed in the male gonads showed enrichment for many terms (over 300) with trends in transporter activities, cell type differentiation/transition and production of hard tissues (cuticle, enamel and hair are listed). All of these are not related to *H. gammarus* testis function. A possible explanation could be the presence of a parasite in the male testes, which would have to be confirmed through microscopy on the tissue samples. However a repeat BLASTx of these transcripts and subsequent analysis through MEGAN5 did not indicate any particular species of parasite. GO enrichment of these unique/not-unique transcripts for the testis can also be found in Supplementary File S5. Given lack of explanation an experimental error cannot be excluded at this point. Transporter activities were particularly over-represented in the gut ('transmembrane transporter activity'; FDR = $4.22e^{-04}$). The heart is a muscle that requires large quantities of energy, and thus many terms related to energy generation ('energy derivation by oxidation of organic compounds', FDR = $9.17e^{-21}$) and muscle cells ('myosin complex', FDR = $2.32e^{-04}$) were found to be enriched. For the hepatopancreas enrichment of GO terms related to catabolism, including hydrolase and polysaccharide catabolic activity were over-represented which relates to the digestive function of this organ. Similarly to the heart, differentially expressed transcripts in the muscle samples were functionally enriched for cellular components that are part of muscle cells (sarcomere, FDR = $1.66e^{-20}$ and contractile fiber FDR = 1.09e^{-19}). Finally, in the nerve enriched GO terms included 'voltage-gated sodium channel activity' (FDR = 7.27e^{-05}), related to generation of action potential. Thus, as expected, over-representation of GO terms allowed for the identification of functional pathways associated with tissue function for most of the tissues analysed, further informing on the molecular mechanisms associated with those functions.

5.4.7 Characterisation of the molecular pathways associated with the immune system

Diseases constitute a risk for the management of both aquaculture and wild populations of *H. gammarus*. To study diseases in a species it is fundamental to understand the interactions between the host and its pathogens, and this requires knowledge of the immune system. The characterization of the immune system further develops the available genomics resources for *H. gammarus*, facilitating disease studies in this species. Of particular interest to this thesis are the interactions between the immune system and WSSV. Given observations that WSSV can infect a broad range of aquatic crustaceans with different levels of pathogenicity and susceptibility, ranging from susceptible *H. gammarus* to relatively resistant *C. maenas*, it was investigated whether such differences are explained by significant differences in the immune systems of these species [12]. The immune system of *C. maenas* was studied in my previous work, allowing for comparisons between the molecular pathways related to immune defence in the lobster compared to the crab[14]. In Verbruggen *et al.* 2015 [14] KEGG pathways were used to identify the presence and absence of immune system components of immune system pathways including the toll-like receptor pathway, the immune deficiency (IMD) pathway, the JAK-STAT and mitogen activated protein kinase (MAPK) signalling pathways. Because viruses often use the endocytosis pathway as a means to enter the host cell, this pathway was also compared. The metazoan transcripts of *H. gammarus* were assigned KEGG orthology groups through a KEGG annotation server, resulting in 7,293 transcript-KOG relations (30.6 % of all metazoan transcripts – Supplementary File S6). Together with the information for *C. maenas* the same server was used to provide overlap comparisons for each pathway. Additionally, components of the invertebrate immune system not represented in KEGG pathways were identified through filtered BLAST searches.

Pattern Recognition Proteins

Hosts produce (receptor) proteins that are capable of binding pathogen associated molecular patterns (PAMP). There are several groups of recognition proteins, each specializing in different types of PAMP, including peptidoglycan and lipopolysaccharides [38]. A possible exception is Down syndrome cell adhesion molecule (Dscam), a molecule that displays highly variable regions and is hypothesized to play a similar role to antibodies in members of the Pancrustacea clade [39]. Once activated through binding its target, the pattern recognition proteins initiate signalling or proteolytic cascades that mediate the host immune response [38, 40]. Table 6 shows which recognition proteins were identified in the *H. gammarus* transcriptome and whether a counterpart was found in the *C. maenas* transcriptome. Not all significant similarities are shown in Table 6, since there are redundancies and unrelated queries. The complete list can be found in Supplementary File S7. Whereas the *C. maenas* transcriptome did not appear to contain peptidoglycan recognition proteins (PGRP), the *H. gammarus* transcriptome might contain such transcripts. Transcript TR44312_c0_g1_i1 had a high level of similarity to FreD (Genbank: AIE45535.1), a shrimp protein that is hypothesized to act as a PGRP [41]. Additionally, transcript TR21885_c0_g1_i1 was similar to the bumblebee PGRP. This difference in receptor presence could be indicative of a difference in how both species deal with pathogens that display a peptidoglycan molecular pattern. Since this is a bacterial PAMP it does not explain difference in WSSV susceptibility, but may contribute to interpret putative differences in susceptibility to bacterial pathogens.

Table 6 Pathogen recognizing proteins in *H. gammarus*

Protein	*H. gammarus*	Query	e-value	Ancestor	Found in *C. maenas*
PGRP	TR44312_c0_g1_i1	gi|662179288 FreD [*Penaeus monodon*]	6.00E-110	Decapoda	No
	TR21885_c0_g1_i1	gi|289583706 peptidoglycan recognition protein [*Bombus ignitus*]	4.00E-48	Pancrustacea	No
C-type lectin	TR45055_c0_g1_i1	gi|558701366 C-type lectin 4 [*Marsupenaeus japonicus*]	2.00E-111	Decapoda	Yes
	TR39758_c0_g1_i1	gi|83595279 C-type lectin [*Glossina morsitans morsitans*]	1.00E-71	Pancrustacea	Yes
Dscam	TR38654_c0_g1_i1	gi|331031260 down syndrome cell adhesion molecule [*Pacifastacus leniusculus*]	0	Astacidea	Yes
GNBP	TR39963_c0_g1_i1	gi|939841624 Gram-negative binding protein gnbp [*Daphnia magna*]	2.00E-80	Crustacea	Yes
TECP	TR41443_c13_g1_i1	gi|331031262 TEP isoform 1 [*Pacifastacus leniusculus*]	0	Astacidea	Yes

After a pathogen has been recognized, the host immune system will typically illicit a response. The response can occur through different mechanisms, e.g. a proteolytic cascade leading to melanisation or activation of transcription factors that lead to the production of antimicrobial or antiviral proteins. The pathways that mediate these mechanisms are central in the innate immune system: the Toll-like receptor pathway, the immunodeficiency pathway, the JAK-STAT pathway and the melanization pathway.

Toll-like receptor pathway

Upon activation of the initial toll-like receptor, the toll-like receptor signalling pathway will activate its associated transcription factor, Nf-κb. The activated transcription factor subsequently moves to the nucleus where it enhances the expression of genes involved in the immune response [38]. The toll-like receptor pathway was represented in KEGG and therefore a comparison between *H. gammarus* and *C. maenas* was made using this pathway as reference (note that KEGG contains vertebrate pathways, differences are expected when investigating aquatic invertebrates such as crab and lobster). Figure 4 shows the presence and absence of components of the toll-like receptor pathway as indicated through KEGG. Most of the components of this pathway are shared amongst both species. However, according to the KEGG annotation there are notable absences in the *H. gammarus* transcriptome including the toll-like receptors and NF-κb, which are both instrumental in innate immunity. Given that KEGG is constructed using predominantly information from vertebrate species, its components are annotated according to similarity to sequences from those vertebrate species. Thus in addition to the KEGG pathway a manual search for invertebrate TLR-pathway components was performed as was done for the recognition proteins. The results in Table 7 show that the proteins of the *Drosophila melanogaster* TLR-pathway, as reported in Kingsolver *et al*. [38] and Li *et al.* [42], are all present. This includes toll-like receptors and Dorsal, the NF-κb homologue. Thus it can be concluded that the TLR-pathway is represented in both the lobster and crab transcriptomes, which is not surprising when considering the central role this pathway plays in the innate immune system.

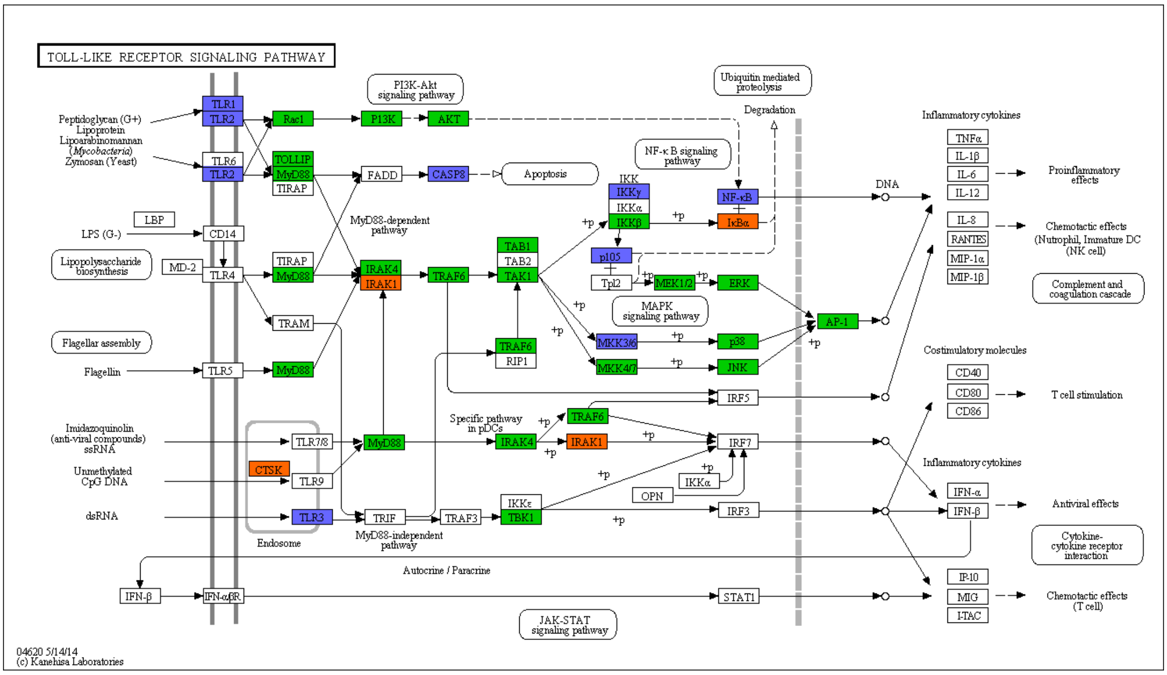


Figure 4 KEGG Toll-like receptor signalling pathway. KEGG reference pathway for Toll-like receptor signalling pathway (map04620). Proteins are indicated by boxes. Shades indicate presence in *H. gammarus* and *C. maenas* transcriptomes: present in both transcriptomes (green), present in only *H. gammarus* (orange), present only in *C. maenas* (purple) and not present in either (white).

Table 7 TLR pathway proteins in *H. gammarus*

Protein	*H. gammarus*	Query	e-value	Ancestor	Found in *C. maenas*
spatzle	TR38162_c0_g1_i1	gi|341650466 Spz1 [*Litopenaeus vannamei*]	7.00E-153	Decapoda	Yes
	TR35249_c0_g1_i2	gi|729056743 spatzle [*Penaeus monodon*]	7.00E-102	Decapoda	Yes
Toll-like receptor	TR19562_c0_g1_i1	gi|914344131 toll-like receptor [*Portunus trituberculatus*]	6.00E-57	Pleocyemata	Yes
	TR2891_c0_g1_i1	gi|914344131 toll-like receptor [*Portunus trituberculatus*]	5.00E-134	Pleocyemata	Yes
	TR30593_c0_g1_i1	gi|914344133 toll-like receptor [*Portunus trituberculatus*]	0	Pleocyemata	Yes
pelle	TR23620_c0_g1_i3	gi|826135738 pelle [*Eriocheir sinensis*]	0	Pleocyemata	Yes
myd88	TR38896_c2_g1_i1	gi|530341204 myeloid differentiation factor [*Eriocheir sinensis*]	0	Pleocyemata	Yes
cactus	TR27620_c0_g1_i1	gi|408366913 cactus [*Fenneropenaeus chinensis*]	5.00E-169	Decapoda	Yes
dorsal	TR20019_c0_g1_i2	gi|575498409 dorsal [*Eriocheir sinensis*]	0	Pleocyemata	Yes

Immunodeficiency (IMD) pathway

Where the toll-like receptor pathway mainly responds to the presence of Gram positive bacteria, the IMD pathway focuses on the recognition and response to Gram negative bacteria. After recognition, a signal is transmitted through IMD, FADD and DREDD which ultimately results in the activation of Relish, a transcription factor of the NF-κb family [38]. The KEGG database does not contain the IMD pathway because it is only characterized in invertebrates although the tumour necrosis factor pathway is considered to be a homologous pathway [43]. IMD pathway members in the *H. gammarus* transcriptome are shown in Table 8, additionally the KEGG mapping to the TNF-pathway (hsa04668) is shown in Figure S 2. In the *C. maenas* transcriptome the only missing member was FADD, but in the *H. gammarus* there is an additional missing gene, the FADD binding partner, Dredd. In the *Eriocheir sinensis* transcriptome FADD was also not reported which is an indication that this might a prevalent occurrence amongst aquatic crustacea as compared to hexapoda like *Drosophila* [42]. Protein binding studies based on *H. gammarus, C. maenas* or *E. sinensis* IMD could reveal whether possible alternatives to FADD exist, or whether there is a direct link between IMD and Dredd in these organisms (and other aquatic Crustacea).

Table 8 IMD pathway proteins in *H. gammarus*

Protein	***H. gammarus***	**Query**	**e-value**	**Ancestor**	**Found in *C. maenas***
IMD	TR33354_c1_g1_i3	gi|523485292 IMD [*Fenneropenaeus chinensis*]	4.00E-52	Decapoda	Yes
FADD	-	-	-	-	No
Dredd	-	-	-	-	Yes
Tab2	TR38854_c2_g2_i2	gi|914701667 TGF-beta-activated kinase 1 and MAP3K7-binding protein 2 [*Litopenaeus vannamei*]	0	Decapoda	Yes
Tak1	TR34157_c0_g1_i1	gi|167864302 tak1 [*Culex quinquefasciatus*]	1.00E-42	Pancrustacea	Yes
	TR34887_c1_g2_i1	gi|167864308 tak1 [*Culex quinquefasciatus*]	1.00E-139	Pancrustacea	Yes
IKK	TR31700_c1_g1_i1	gi|442628615 I-kappaB kinase epsilon, isoform C [*Drosophila melanogaster*]	0	Pancrustacea	Yes
	TR37566_c2_g1_i1	gi|386765870 I-kappaB kinase beta [*Drosophila melanogaster*]	1.00E-51	Pancrustacea	Yes
Relish	TR32007_c0_g1_i1	gi|846925119 relish [*Macrobrachium rosenbergii*]	0	Pleocyemata	Yes

JAK-STAT signalling pathway

The JAK-STAT signalling pathway differs from the IMD and TLR pathways in that it does not respond directly to the presence of pathogens, instead it is activated through chemical messengers like cytokines. The messenger molecules bind the receptor (Domeless in invertebrates) which results in phosphorylation of JAK and STAT. The latter translocates to the nucleus and regulates gene expression of target genes. Through phosphatases, such as SOCS, and inhibitors, like PIAS, the response can be regulated. It is known that the JAK-STAT pathway plays a role in WSSV infection [44]. It was shown during WSSV infection of *Penaeus monodon* that the virus can exploit host STAT to promote expression of its own genes (early gene *ie1*) [45]. Furthermore, it was recently shown that the viral WSSV-miR-22 can influence the translation of host STAT, thereby reducing its numbers [46]. Knockdown of JAK/STAT through siRNA resulted in increased copies of WSSV, partly due to a reduction of expression of response proteins like thioester-containing proteins [46]. Thus JAK-STAT signalling plays an apparently contradictory role through both promoting WSSV gene expression and bolstering the host immune response. It should be noted that STAT is not the only transcription factor capable of promoting *ie1* expression and furthermore its requirement could only be necessary in early stages of infection. As infection spreads a disruption of the host immune response through STAT repression might be more effective. Since the JAK-STAT signalling pathway is well conserved it is no surprise that its components are present in both *C. maenas* and *H. gammarus* (Table 9).

Table 9 JAK-STAT signalling pathway in *H. gammarus*

Protein	***H. gammarus***	**Query**	**e-value**	**Ancestor**	**Found in *C. maenas***
Domeless	TR40378_c2_g1_i6	gi|552331594 domeless [*Litopenaeus vannamei*]	0	Decapoda	Yes
JAK	TR39053_c0_g1_i1	gi|831252996 Janus kinase [*Litopenaeus vannamei*]	0	Decapoda	Yes
STAT	TR39575_c0_g1_i1	gi|576250986 signal transducer and activator of transcription [*Scylla paramamosain*]	0	Pleocyemata	Yes
PIAS	TR35494_c3_g1_i2	gi|147903229 protein inhibitor of activated STAT [*Ciona intestinalis*]	1.00E-75	Bilateria	Yes
SOCS	TR31627_c3_g1_i1	gi|646701115 Suppressor of cytokine signaling 5 [*Zootermopsis nevadensis*]	5.00E-120	Pancrustacea	Yes

Response proteins

The signalling cascade through the IMD, Toll and JAKSTAT pathways results in a transcriptional immune response mediated by transcription factors including STAT and members of the NF-κB family. One part of this immune response is to increase expression of pathogen recognizing proteins like thioester-containing proteins. Another part is to generate proteins that are capable of attacking the pathogen e.g. anti-lipopolysaccharide factor (ALF) and lysozyme [47]. Such proteins are listed in Table 10. In *H. gammarus* the same group of response proteins was found as in the set of predicted proteins of *C. maenas*. Proteins identified in shrimp were also investigated; these are stylicins (from *Litopenaeus stylirostris*) [48] and penaeidins (from *Litopenaeus vannamei*) [49, 50]. However, there were no transcripts with significant sequence similarity to these proteins, either in *H. gammarus* or in *C. maenas*. Thus it appears that both stylicins and penaeidins are limited to shrimp species.

Table 10 Response proteins in *H. gammarus*

Protein	***H. gammarus***	**Query**	**e-value**	**Ancestor**	**Found in *C. maenas***
ALF	TR31265_c0_g1_i1	gi|349502999 anti-lipopolysaccharide factor [*Macrobrachium rosenbergii*]	6.00E-44	Pleocyemata	Yes
Crustin	TR25234_c1_g1_i1	gi|452056154 crustin-like protein [*Macrobrachium rosenbergii*]	3.00E-22	Pleocyemata	Yes
Carcinin	TR22441_c0_g1_i1	gi|187450114 carcinin-like protein [*Fenneropenaeus chinensis*]	9.00E-33	Decapoda	Yes
Lysozyme	TR34979_c1_g1_i1	gi|262385514 lysozyme [*Procambarus clarkii*]	8.00E-40	Astacidea	Yes
iNOS	TR34333_c0_g1_i2	gi|959093568 inducible nitric oxide synthase [*Tigriopus japonicus*]	4.00E-147	Crustacea	Yes

Melanization pathway

The pathways discussed above are involved in the mediation of gene expression. In contrast, the melanization pathway acts through proteolytic cascades and its components are readily available as zymogens. The activation of PRRs results in the processing of the zymogens into active enzymes which eventually lead to functionally active phenol oxidase (PO). PO generates melanin which can encapsulate pathogens. Furthermore, production of melanin from phenols and quinones generates reactive oxygen species that can damage the pathogen. Melanization has significant impact on WSSV infection in shrimp (Sutthangkul *et al.* 2015 [51]), with inhibition of PPO activity resulting in higher susceptibility of *P. monodon* to WSSV. Through yeast two-hybrid experiments it was shown that WSSV453 can inhibit the activity of PPO activating enzyme (PPAE), thereby decreasing the effectiveness of this immune system pathway. As in *P. monodon*, all of the players in the melanization pathway have counterparts in *H. gammarus* and *C. maenas* (Table 11). This shows that, like other immune pathways, melanization is well conserved within invertebrates.

Table 11 Melanization pathway in *H. gammarus*

Protein	*H. gammarus*	Query	e-value	Ancestor	Found in *C. maenas*
lozenge	TR39957_c3_g1_i1	gi|607367383 Protein lozenge [*Cerapachys biroi*]	2.00E-90	Pancrustacea	Yes
mp1	TR39920_c0_g1_i4	gi|28571479 melanization protease 1, isoform A [*Drosophila melanogaster*]	3.00E-61	Pancrustacea	Yes
peroxynectin	TR33034_c0_g1_i1	peroxinectin [*Hydra vulgaris*]	4.00E-41	Eumetazoa	Yes
PPAE	TR39910_c2_g1_i1	gi|414151636 prophenoloxidase activating enzyme [*Litopenaeus vannamei*]	5.00E-150	Decapoda	Yes
PPO	TR37035_c0_g1_i1	gi|448918007 prophenoloxidase [*Callinectes sapidus*]	0	Pleocyemata	Yes
serpent	TR36556_c2_g2_i7	gi|270008081 serpent [*Tribolium castaneum*]	3.00E-30	Pancrustacea	Yes
serpin	TR29171_c0_g1_i1	gi|764399009 serine proteinase inhibitor-2 [*Eriocheir sinensis*]	0	Pleocyemata	Yes
sp7	TR39920_c0_g1_i4	gi|665393450 serine protease 7, isoform F [*Drosophila melanogaster*]	8.00E-46	Pancrustacea	Yes

RNAi pathway

RNA interference (RNAi) is one of the major antiviral pathways within the invertebrate innate immune system [52]. Small RNA molecules can regulate the translation of proteins and this strategy can be employed by the host, for example for the fine tuning of the activity of the immune system or through directly targeting mRNA of viral origin. Similarly, a virus can influence expression of the host genes, e.g. inhibit the immune system or maintaining a cellular environment conducive to viral replication. The role of miRNAs in WSSV infection has been reviewed in Verbruggen et al. 2016 [53]. WSSV expresses several miRNAs that regulate either its own gene expression or that of the host. In both cases the virus requires host proteins from the RNAi pathway for synthesis of its own miRNAs, in particular Drosha and Dicer-1 [54]. Thus for successful WSSV infection in *H. gammarus,* the presence of at least these two proteins would be a requirement. Similar to the other pathways of the immune system, most known members of the RNAi pathway are indeed found in the *H. gammarus* transcriptome. A notable absentee is Dicer-2, a protein that is involved in the siRNA route in RNAi. This could indicate that Dicer-1 plays the role of Dicer-2 in *H. gammarus* as occurs in *Homo sapiens* [55]. This is not the case in *C. maenas* and *D. melanogaster* where Dicer-2 transcripts were found, and in the latter case shown to have separate function [55]. Since these species are closely related to *H. gammarus* more research is required to establish the roles of Dicer in the lobster before firm conclusions can be drawn.

Table 12 RNAi pathway components

Protein	*H. gammarus*	Query	e-value	Ancestor	Found in *C. maenas*
TRBP	TR26624_c0_g1_i2	gi|444174849 TAR RNA-binding protein 1 [*Penaeus monodon*]	0	Decapoda	Yes
DGCR8	TR38870_c0_g1_i1	gi|941117436 Microprocessor complex subunit DGCR8 [*Daphnia magna*]	4.00E-147	Crustacea	Yes
Drosha	TR26624_c0_g1_i3	gi|954619683 Ribonuclease 3 [Trichinella sp. T8]	7.00E-85	Ecdysozoa	Yes
AGO1	TR40750_c4_g1_i1	gi|283827858 argonaute 1 [*Marsupenaeus japonicus*]	0	Decapoda	Yes
AGO2	TR41530_c2_g2_i1	gi|563729913 argonaute2 [*Penaeus monodon*]	3.00E-120	Decapoda	Yes
Dicer1	TR27435_c0_g1_i1	gi|17738129 Dicer-1 [*Drosophila melanogaster*]	0	Pancrustacea	Yes
Dicer2	-	-	-	-	Yes

Endocytosis pathway

Whilst not directly associated with the immune system, the endocytosis pathway is nevertheless of significant importance to viral infections because most viruses employ the endocytosis pathway as their gateway into the host cells [56]. This is generally accomplished through binding of proteins on the host cell surface to the virus, followed by endocytic uptake (usually Clathrin- or caveolar-mediated endocytosis) [56]. Once inside the vesicles, the virus is transported closer to its destination, the host nucleus. WSSV has been shown to use this strategy in successful infections and several interactions between host receptors and viral proteins have been identified, for example interactions between integrin and VP26 in *L. vannamei* (see Verbruggen *et al.* 2016 [53] for more a detailed discussion on these interactions). Further experiments involving endocytosis inhibiting chemicals have shown that these interactions can result in uptake through Clathrin-mediated endocytosis for cellular entry [57]. Lastly, an interaction between VP28 and Rab7, an important regulator in the endocytosis pathway, has been identified. This could be a trigger that aids WSSV in escaping from the endosomes in order to avoid degradation in lysosomes and to continue toward the nucleus [53].

The different endocytic pathways and the subsequent vesicle transport and maturation systems are shown in Figure 5. Both *H. gammarus* and *C. maenas* have a good coverage of the components of this system according to KEGG annotation. As compared to *C. maenas,* more components were identified for *H. gammarus*, most notably, caveolin was identified in *H. gammarus* but not in *C. maenas* (although a transcript with some similarity a caveolin protein sequence was identified [14]). This pathway can be used in gene expression studies to identify whether the host cell transport mechanisms are adjusted in response to viral presence, either under direction of the host or the virus.

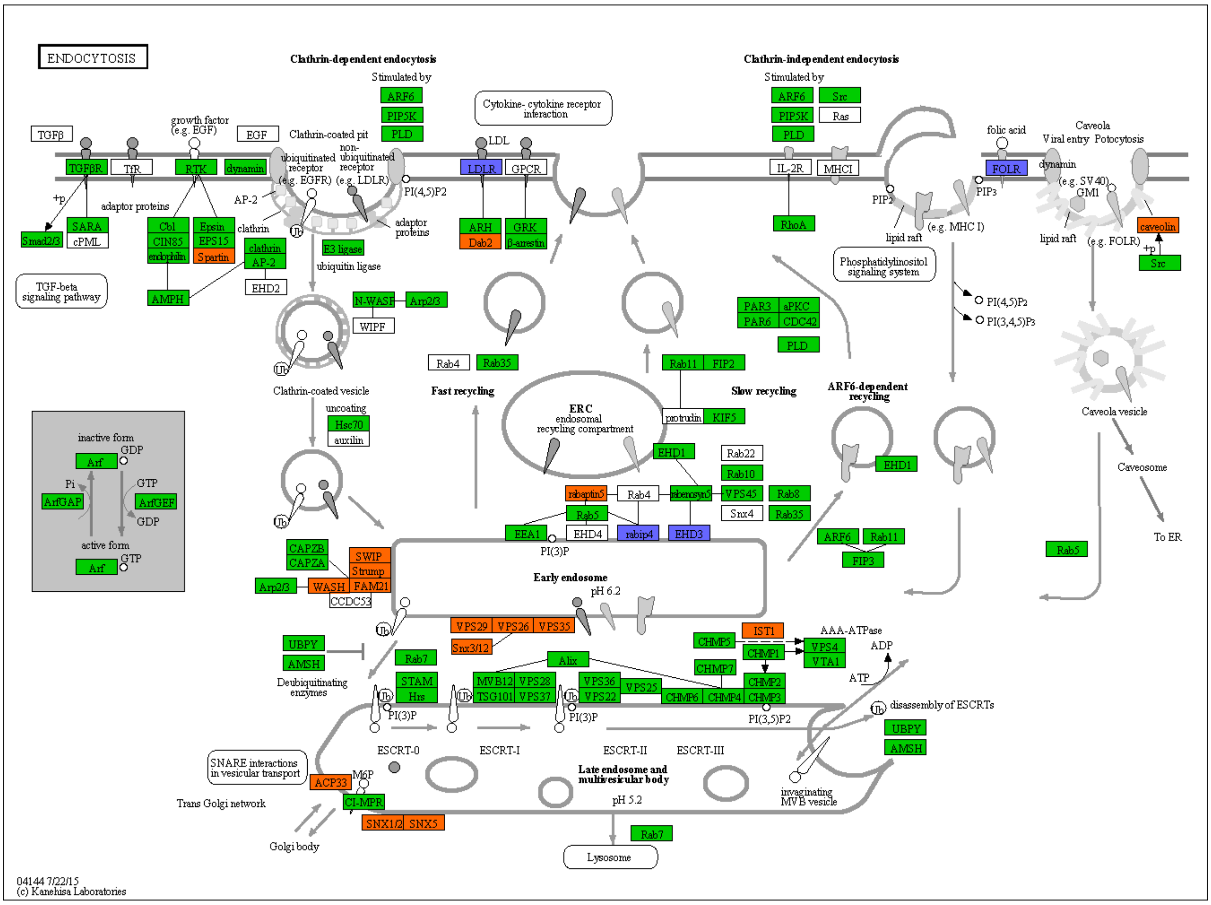


Figure 5 KEGG endocytosis pathway. KEGG reference pathway for endocytosis (ko04144). Proteins are indicated by boxes. Shades indicate presence in *H. gammarus* and *C. maenas* transcriptomes: present in both transcriptomes (green), present in only *H. gammarus* (orange), present only in *C. maenas* (purple) and not present in either (white).

5.4.8 Comparative genomics of WSSV receptors

The immune systems of *H. gammarus* and *C. maenas* are similar and share the majority of their components. Therefore, it is difficult to relate differences in the sequence or presence of specific genes to susceptibility to infection by WSSV. The factors that can influence disease susceptibility are multiple and complex. These include the interaction between WSSV and host proteins, particularly those at the initial stages of infection. As was mentioned before, WSSV requires binding to cell surface receptors in order to initiate Clathrin-mediated endocytosis. Experiments in shrimp have shown that the host integrin and lectin proteins can serve as the receptor for WSSV [58, 59]. Once taken up the virus has to escape the endosomes in order to avoid degradation in lysosomes [56]. At this stage another well documented interaction between WSSV and host proteins can take place. Rab7, an important regulator on late endosomes (Figure 5), has been shown to interact with VP28 [60]. This interaction could be related to the virus leaving the late endosomes, attaching itself to late endosomes for perinuclear transport or small amounts of VP28 could be present on the host cell surface acting as a viral receptor [60]. Thus there are three host proteins that are important for the initial infection stages of WSSV. The sequence and configuration of these receptors determines whether the viral proteins can bind and differences in these can affect binding either beneficially or detrimentally, explaining some of the differences in susceptibility observed between hosts [61, 62]. In line with this the sequences of the WSSV receptors for *H. gammarus*, *C. maenas* and other crustaceans were aligned and compared in a phylogenetic tree.

Prior to comparing the viral receptors, it is necessary to understand the phylogeny of crustaceans. It can be expected that crustacean species that are further away on the phylogenetic tree would have a larger difference in sequence. Koenemann *et al.* 2010 studied arthropod phylogeny based on sequences for 16S rDNA, cytochrome C oxidase subunit I and the nuclear ribosomal gene 18S rDNA of 88 arthropod species [63]. Their work showed differences according to the method chosen for the analysis but the Decapoda order appears relatively stable across all trees (see Figure 6), and the phylogenetic distance between crabs (*Carcinus*) and lobsters (*Jasus*, *Hommarus* is not shown) is smaller than crabs compared to shrimps (*Penaeus*). A phylogenetic analysis of cytochrome C oxidase subunit I based on the *H. gammarus* and *C. maenas* transcriptromes (Chapter 3) is shown in the Supplement to this chapter.

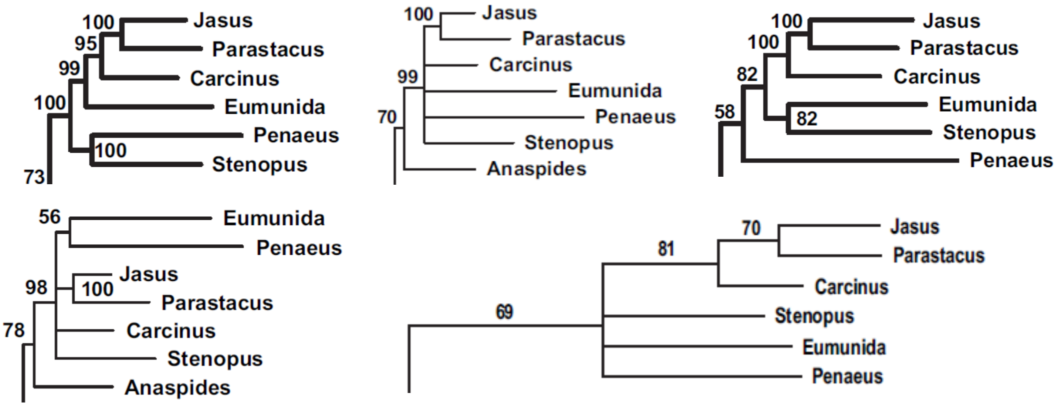


Figure 6 Decapod phylogenic trees. Cropped phylogenetic trees from Koenemann et al. 2010 [63]. Trees were produced from five separate runs with different software tools and settings. Shrimps are represented by *Penaeus*, crabs by *Carcinus* and lobsters by *Jasus*.

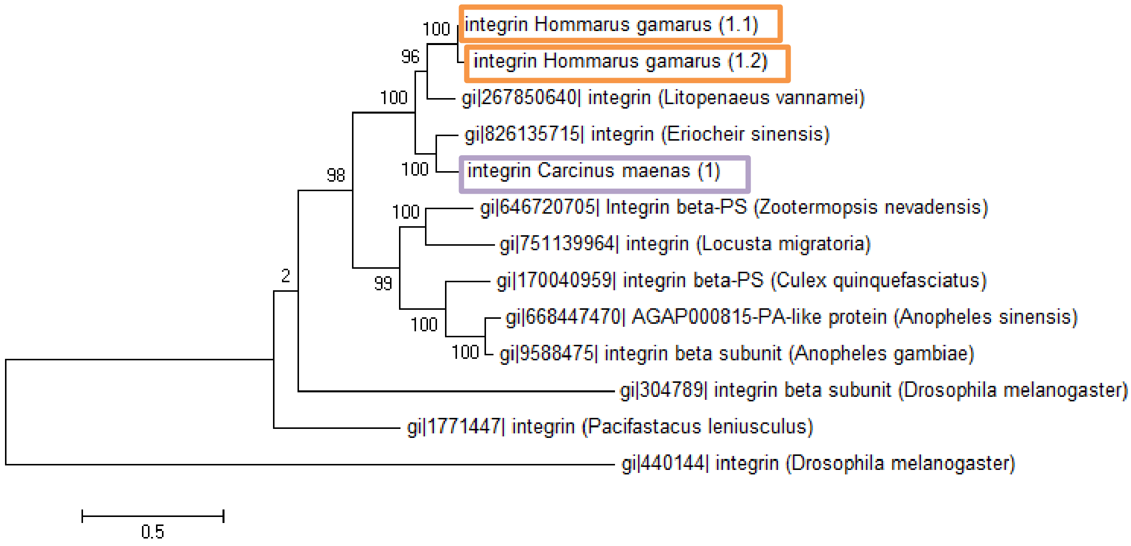

100
96
100
100
98
100
99
100
100
2
integrin Hommarus gamarus (1.1)
integrin Hommarus gamarus (1.2)
gi|267850640| integrin (Litopenaeus vannamei)
gi|826135715| integrin (Eriocheir sinensis)
integrin Carcinus maenas (1)
gi|646720705| Integrin beta-PS (Zootermopsis nevadensis)
gi|751139964| integrin (Locusta migratoria)
gi|170040959| integrin beta-PS (Culex quinquefasciatus)
gi|668447470| AGAP000815-PA-like protein (Anopheles sinensis)
gi|9588475| integrin beta subunit (Anopheles gambiae)
gi|304789| integrin beta subunit (Drosophila melanogaster)
gi|1771447| integrin (Pacifastacus leniusculus)
gi|440144| integrin (Drosophila melanogaster)
0.5

Figure 7 Maximum Likelihood phylogenetic tree for crustacean intregrin. A phylogenetic tree was generated for integrin sequences from *H. gammarus* (orange), *C. maenas* (purple) and available crustacean sequences from NCBI, hexapoda were added as outgroup. The tree was generated in MEGA 6.06. There were two possible peptide sequences for *H. gammarus*, both derived from transcript ‘TR40213_c2_g2_i3’. The *C. maenas* peptide sequence is based on transcript ‘comp89839_c0_seq1’.

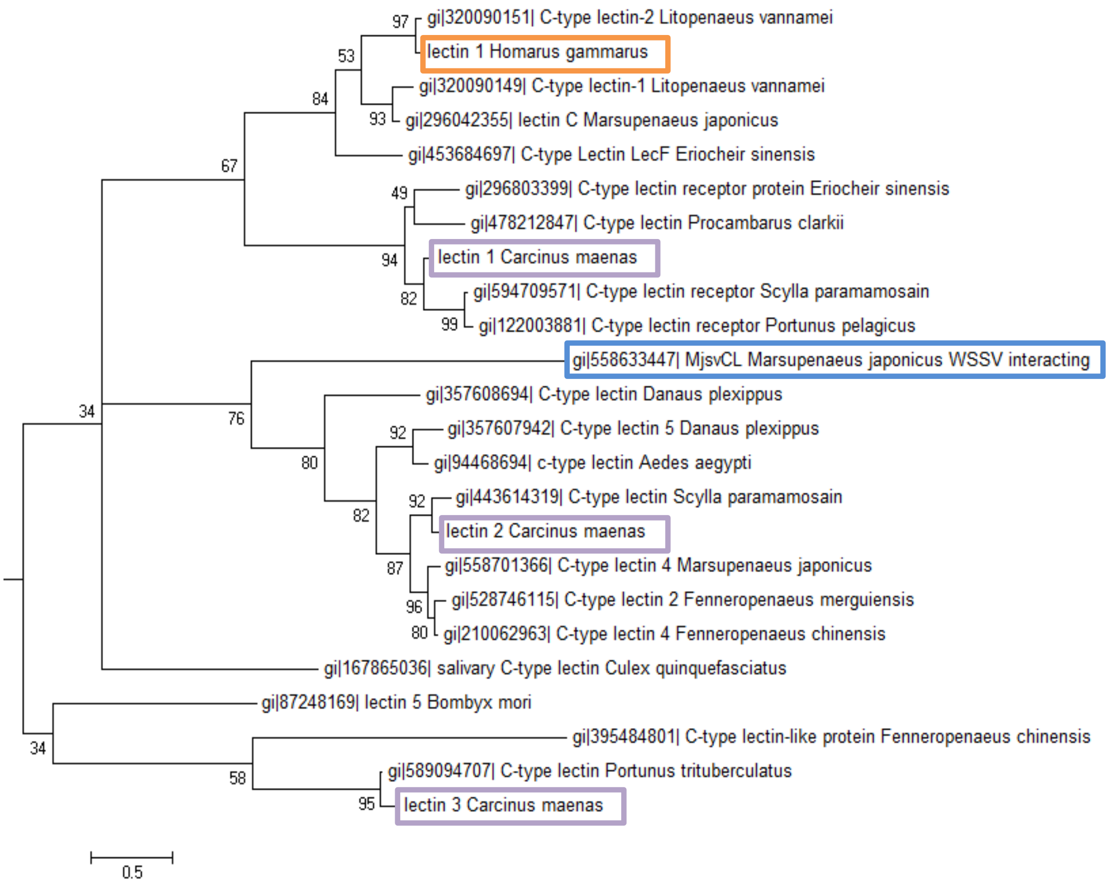

97
gi|320090151| C-type lectin-2 Litopenaeus vannamei
53
lectin 1 Homarus gammarus
84
gi|320090149| C-type lectin-1 Litopenaeus vannamei
93
gi|296042355| lectin C Marsupenaeus japonicus
67
gi|453684697| C-type Lectin LecF Eriocheir sinensis
49
gi|296803399| C-type lectin receptor protein Eriocheir sinensis
gi|478212847| C-type lectin Procambarus clarkii
94
lectin 1 Carcinus maenas
82
gi|594709571| C-type lectin receptor Scylla paramamosain
99
gi|122003881| C-type lectin receptor Portunus pelagicus
gi|558633447| MjsvCL Marsupenaeus japonicus WSSV interacting
34
76
gi|357608694| C-type lectin Danaus plexippus
92
gi|357607942| C-type lectin 5 Danaus plexippus
80
gi|94468694| c-type lectin Aedes aegypti
92
gi|443614319| C-type lectin Scylla paramamosain
82
lectin 2 Carcinus maenas
87
gi|558701366| C-type lectin 4 Marsupenaeus japonicus
96
gi|528746115| C-type lectin 2 Fenneropenaeus merguiensis
80
gi|210062963| C-type lectin 4 Fenneropenaeus chinensis
gi|167865036| salivary C-type lectin Culex quinquefasciatus
gi|87248169| lectin 5 Bombyx mori
34
gi|395484801| C-type lectin-like protein Fenneropenaeus chinensis
58
gi|589094707| C-type lectin Portunus trituberculatus
95
lectin 3 Carcinus maenas
0.5

Figure 8 Maximum Likelihood phylogenetic tree for crustacean lectins. A phylogenetic tree was generated for lectin sequences from *H. gammarus* (orange), *C. maenas* (purple) and available crustacean sequences from NCBI, hexapoda were added as outgroup. The tree was generated in MEGA 6.06. The *H. gammarus* peptide sequence is based on transcript 'TR33161_c0_g1_i1'. The *C. maenas* peptide sequences are based on: lectin 1 'comp82774_c0_seq1', lectin 2 by 'comp87522_c4_seq1' and lectin 3 by 'comp75697_c0_seq1'. The blue box indicates the lectin that was identified to aid in WSSV infection [59].

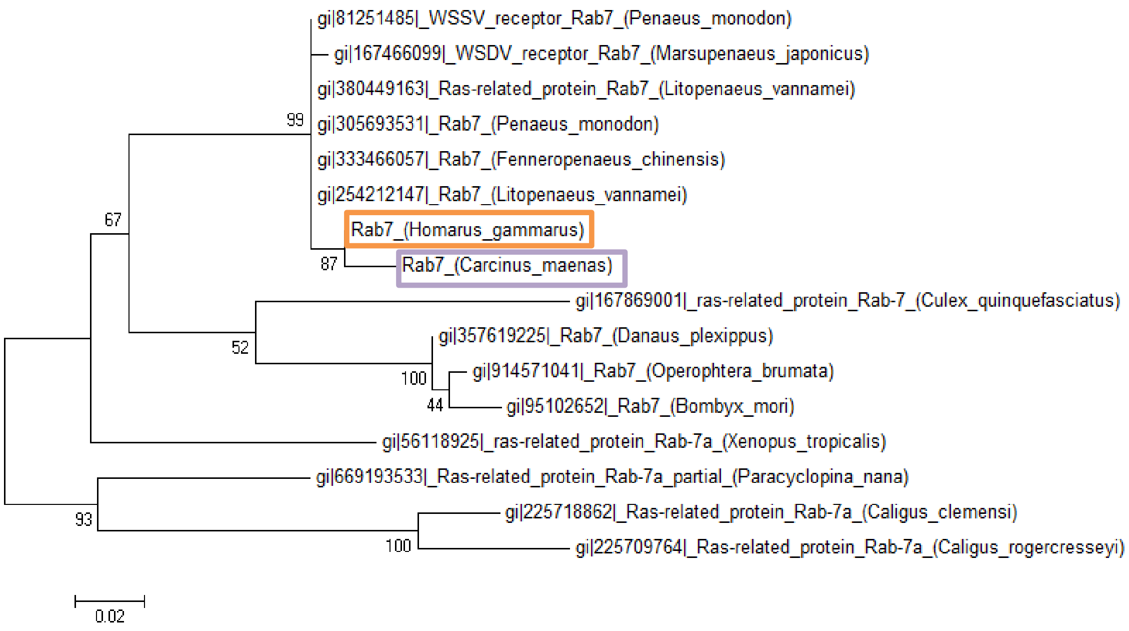

gi|81251485|_WSSV_receptor_Rab7_(Penaeus_monodon)
gi|167466099|_WSDV_receptor_Rab7_(Marsupenaeus_japonicus)
gi|380449163|_Ras-related_protein_Rab7_(Litopenaeus_vannamei)
gi|305693531|_Rab7_(Penaeus_monodon)
gi|333466057|_Rab7_(Fenneropenaeus_chinensis)
gi|254212147|_Rab7_(Litopenaeus_vannamei)
Rab7_(Homarus_gammarus)
Rab7_(Carcinus_maenas)
gi|167869001|_ras-related_protein_Rab-7_(Culex_quinquefasciatus)
gi|357619225|_Rab7_(Danaus_plexippus)
gi|914571041|_Rab7_(Operophtera_brumata)
gi|95102652|_Rab7_(Bombyx_mori)
gi|56118925|_ras-related_protein_Rab-7a_(Xenopus_tropicalis)
gi|669193533|_Ras-related_protein_Rab-7a_partial_(Paracyclopina_nana)
gi|225718862|_Ras-related_protein_Rab-7a_(Caligus_clemensi)
gi|225709764|_Ras-related_protein_Rab-7a_(Caligus_rogercresseyi)
99
67
87
52
100
44
93
100
0.02

Figure 9 Maximum Likelihood phylogenetic tree for crustacean rab7. A phylogenetic tree was generated for rab7 sequences from *H. gammarus* (orange), *C. maenas* (purple) and available crustacean sequences from NCBI, hexapoda were added as outgroup. The tree was generated in MEGA 6.06. The *H. gammarus* peptide sequence is based on transcript 'TR37571_c0_g1_i1'. The *C. maenas* peptide sequence is based on transcript 'comp81392_c0_seq1'.

For the viral receptors, transcripts with significant tBLASTn homology to these receptors were selected. The peptide sequences of the open reading frames that corresponded to the viral interaction proteins were used in MEGA to generate the phylogenetic trees. In the phylogenetic tree for the integrin receptor (Figure 7), instead of grouping with the crab, the *H. gammarus* transcripts show a closer relationship to *L. vannamei*. This contrasts to the trees in Figure 6. It is indicative of a difference in evolutionary rate for integrin compared to 16SrDNA, cytochrome C oxidase subunit I and 18S rDNA.

There are many members of C-type lectin family across the invertebrates. The phylogenetic tree for the lectins identified in *H. gammarus* and *C. maenas* illustrate this diversity (Figure 8). For *H. gammarus* only a single lectin was identified whereas three were found in *C. maenas*. The *H. gammarus* lectin is once again grouped alongside the lectins from *L. vannamei* in the group that also contains *C. maenas* lectin 1. The other *C. maenas* lectins each appear to be members of separate families, each in turn showing homology to other crab/crayfish/shrimp lectins. The lectin that was identified to have an interaction with the virus and promote its infection groups with *C. maenas* lectin 2 family, but does boast large difference in sequence [59].

The Rab7 tree (Figure 9) shows that this gene is particularly well conserved amongst invertebrates, indicating its importance to the fundamental endocytosis pathway. The aquatic crustaceans group together with nearly all shrimp species showing identical sequences. The *H. gammarus* and *C. maenas* rab7 have slight differences in sequence, six amino acids across the protein sequence. It is noteworthy that this is the only three wherein the *H. gammarus* and *C. maenas* sequences group together, indicating that more stronger conserved sequences align better with the consensus from Koenemann *et al.* 2010 [63].

It is interesting to see that the WSSV receptors of *H. gammarus* are very similar to those identified in *L. vannamei*. Given the trees presented in Koenemann *et al.* 2010 [63] one would expect a larger distance between these species. This might be interpreted as a possible indication of different evolutionary rates, and thus a selective pressure, for WSSV receptors in crustaceans. Based on current data one can only speculate on the nature of this selective pressure but there are hypotheses that WSSV is originally derived from a virus that infected crab [64]. Given current observations it could be that the disease caused by this precursor of WSSV is the basis of this pressure, but this requires further investigation.

5.5 Conclusion

The European lobster, *H. gammarus*, is an important economic species with limited genomic resources available. In this body of work a *de novo* assembled transcriptome for this species based on RNA-sequencing data from nine tissues was produced. The transcriptome consisted of 106,498 transcripts out of which 25,763 (24.2 %) showed similarity to known sequences. Nearly all well-conserved genes were identified in the transcriptome and differentially-expressed genes related to tissue function. In order to enable studies on host pathogen interactions and disease, including viral pathogens like WSSV, we characterized the immune system and showed the presence of most known components in the toll-like receptor, IMD, JAK-STAT, RNAi and endocytosis pathways. We compared the coverage of the immune system to that of the shore crab, *C. maenas*, in order to derive whether susceptibility to WSSV infection could be attributed to significant differences therein. The immune system pathways appeared very similar in both species, often containing the same components. The most significant difference is that a PGRP was identified in *H.*

gammarus but lacking in the *C. maenas* transcriptome. This could be relevant in the case of bacterial infections, but is insufficient in explaining the difference in WSSV susceptibility since the virus does not contain peptidoglycan. Another difference is the absence of Dredd in the IMD pathway of *H. gammarus*, which is another factor whose principal task involves antibacterial defence. Perhaps more relevant to the WSSV pathogen are differences in interacting proteins between virus and host. We generated phylogenetic trees for three known WSSV-interacting proteins containing sequences of *H. gammarus*, *C. maenas* and proteins from susceptible shrimp species. The tree for Rab7, a highly conserved protein involved in regulation of endocytosis which interacts with WSSV VP28, showed that the *H. gammarus* Rab7 is more similar to those of the susceptible species. I hypothesise that the differences in WSSV pathogenicity could thus related to the Rab7 receptor. However, verifying this hypothesis would require more detailed knowledge of the interaction between Rab7 and VP28 in order to derive the important domains/amino acids involved, and how binding affinity between WSSV and RAB7 related to pathogenicity.

The current dataset points to differences in the Rab7 receptor as possible explanation for diverse pathogenicity of WSSV in aquatic crustacean species. However it should to be noted that the current dataset is limited in its ability to explain differences in pathogenicity. For example expression changes upon pathogen exposure were not determined as part of this study and information on miRNAs that often play an important role in crustacean responses to viruses was not measured. Infection studies of *H. gammarus* exposed to WSSV and comparative analysis of the responses to WSSV with other susceptible and resistant species would allow for a more in-depth understanding of the molecular pathways determining susceptibility to this important pathogen in the future.

5.6 Supplementary Tables, Figures and Topics

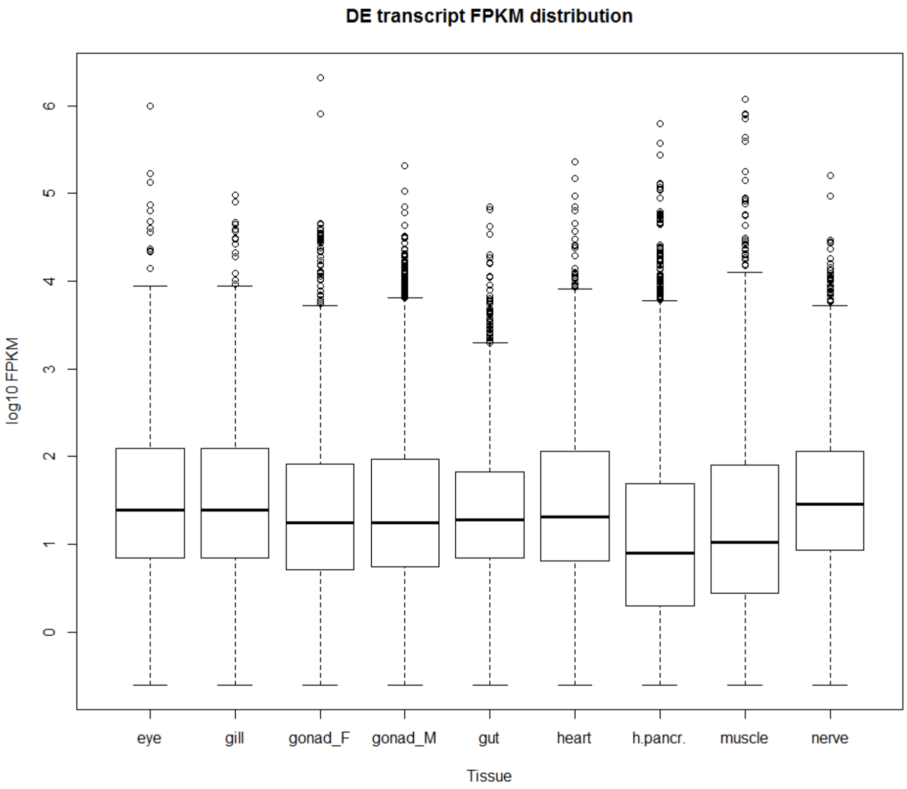


Figure S 1 DE transcript FPKM distribution. Distribution of gene expression values, in FPKM, per tissue.

Table S 1 Expression of transcripts in tissues.

Tissue	Total number of transcripts expressed in at least 1 sample	Number of transcripts only expressed in this tissue	Total number of transcripts expressed in at least 1 sample (metazoan)	Number of transcripts only expressed in this tissue (metazoan)
All tissues	106,498		23,815	
eye	53,382	593	14,625	132
gill	57,640	958	14,947	185
gonad_F	55,741	1,133	14,981	221
gonad_M	68,793	5,792	18,556	1,288
gut	57,034	853	14,355	121
heart	54,502	620	14,133	98
hepatopancreas	46,691	633	12,844	123
muscle	46,988	604	13,319	100
nerve	58,104	901	15,054	155

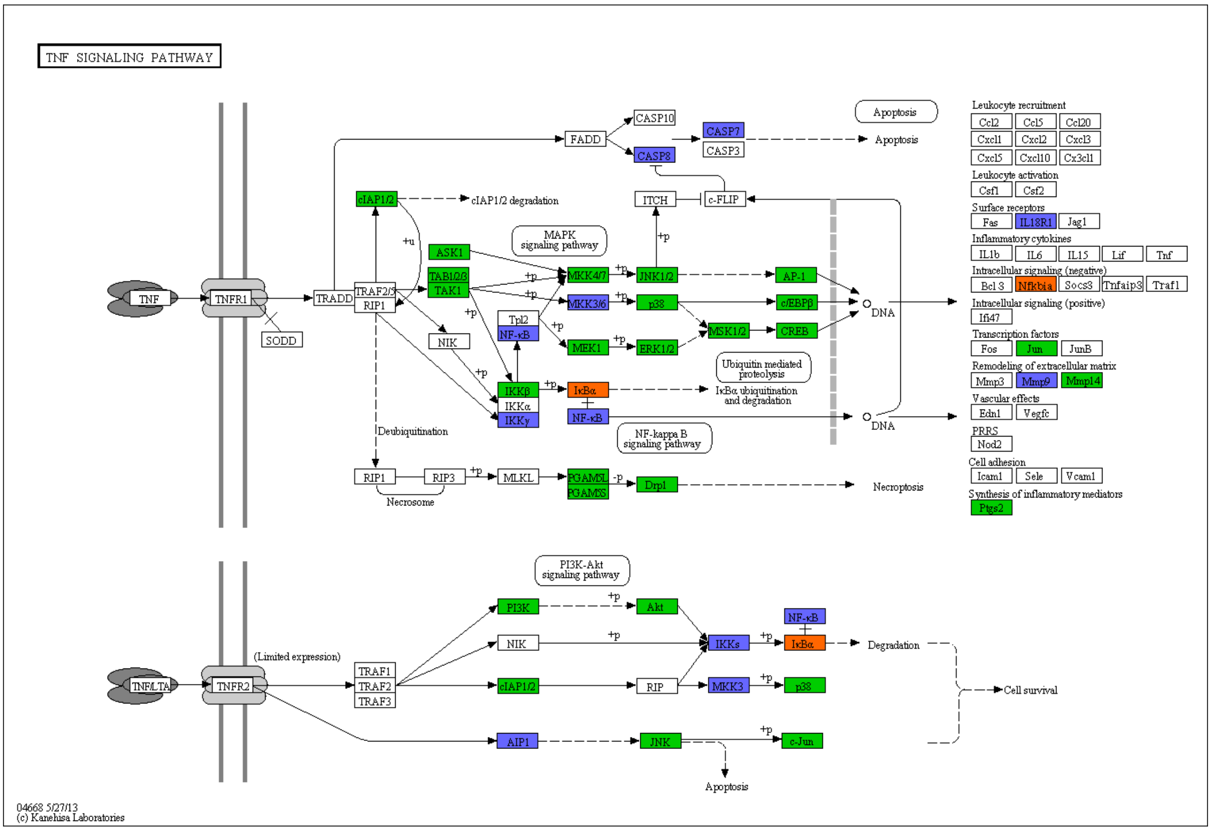


Figure S 2 KEGG TNF pathway. KEGG reference pathway for the TNF pathway (hsa04668). Proteins are indicated by boxes. Shades indicate presence in *H. gammarus* and *C. maenas* transcriptomes: present in both transcriptomes (green), present in only *H. gammarus* (orange), present only in *C. maenas* (purple) and not present in either (white).

Phylogenetic analysis based on transcriptome sequences for H. gammarus and C. maenas.

Koenemann *et al.* 2010 [63] presented an extensive analysis of arthropod phylogeny based on 18SrDNA, 16S rDNA, and cytochrome c oxidase I (COI). Their data was derived from PCR amplification and sequencing of a wide range of arthropod species. To see whether similar results could be obtained through analysis of sequences in the transcriptoms from Chapter 3 (*C. maenas*) and Chapter 5 (*H. gammarus*) we performed a phylogenetic analysis for COI. 18SrDNA and 16SrDNA could not be identified since ribosomal RNA is depleted during RNA isolation (see methods).

C. maenas COI and *H. gammarus* COI were identified using the sequences obtained by Koenemann *et al.* 2010 [63] as query for a BLASTN search to the transcriptomes. Searches identified one potential COI transcripts in *C. maenas* (comp90149_c3_seq1) and one in *H. gammarus* (TR37622_c1_g1_i1). Since the alignments in Koenemann *et al.* 2010 were performed on partial sequences, the same partial sequences were retrieved for *C. maenas* and *H. gammarus*. For example, in the *C. maenas* transcript only bases 3936 – 3456 covered the *C.maenas* sequence. A maximum likelihood phylogenetic tree (Nearest-Neighbour-Interchange withTamura-Nei model and 500 bootstrap iterations), was generated in MEGA 6.06.

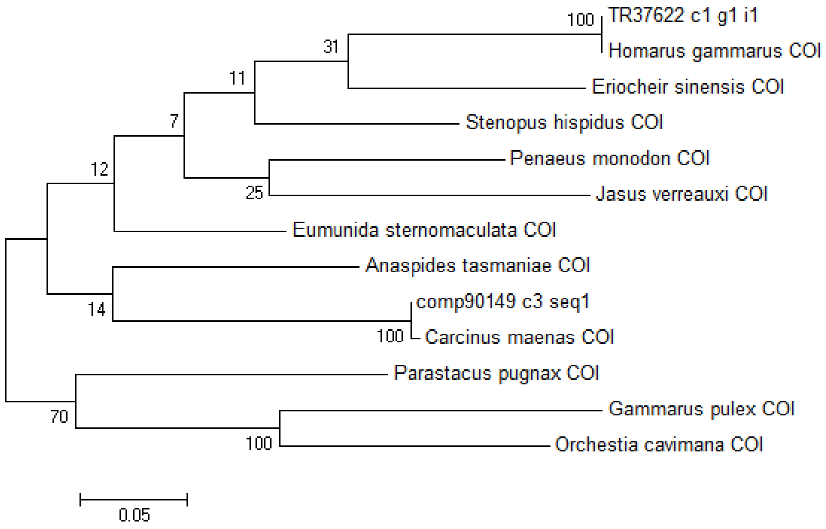


Figure S 3 Crustacean phylogenetic tree of Cytochrome Oxidase Subunit I. Phylogenetic tree generated by MEGA 6.06. Sequences derived from Koenemann *et al.* 2010 [63] and Chapter 3 (*C. maenas*: comp90149 c3 seq1) and Chapter 5 (*H. gammarus*: TR37622 c1 g1 i1).

The sequence alignment showed a sequence identity > 99% for both the *C. maenas* and *H. gammarus* transcripts as compared to the sequences from Koenemann *et al.* 2010. Therefore it can be expected that any differences in the tree are the result of choices made in the multiple sequence alignment and tree assembly process. The tree in Figure S 3 shows differences to those presented in Figure 6. For example, *C. maenas* is closes to *Anaspides tasmaniae* instead of the expected *E. sinensis*. It is however noteworthy that the bootstrap shows that support for the branches is not very high. Given that sequences are nearly identical and the analysis by Koenemann *et al.* 2010 [63] includes additional sequence data it is advisable to follow their analysis.

5.7 Supplementary Files

Supplementary file 1: S1_Pathway_annotation.R – R script for immune system component identification.

Supplementary file 2: S2_Hgammarus_readcounts.txt – post quality filtering readcounts for every sample.

Supplementary file 3: S3_Hgammarus_trinotate_annotation.xls – annotation of transcripts in the *H. gammarus* transcriptome

Supplementary file 4: S4_Hgammarus_top10_expressed.xlsx – Top10 highest expressed transcripts in every tissue, along with annotation.

Supplementary file 5: S5_Hgammarus_GO_enrichment.xlsx – Tissue specific GO enrichment as determined by BLAST2GO®

Supplementary file 6: S6_Hgammarus_transcripts_to_KOG.txt – KEGG orthology groups associated with *H. gammarus* transcripts based on bi-direction best hit.

Supplementary file 7: S7_Hgammarus_immune_pathways.zip – *H. gammarus* transcripts with significant similarity to immune system proteins. The archive contains excel files for every immune pathway described in the text.

5.8 References

1. FAO: **Species Fact Sheets: *Homarus gammarus***. *FAO Species Catalogue, **13*** 2015.
2. FAO: **Fishery Statistical Collections: Global Capture Production**. 2015.
3. Daniels CL, Wills B, Ruiz-Perez M, Miles E, Wilson RW, Boothroyd D: **Development of sea based container culture for rearing European lobster (*Homarus gammarus*) around South West England**. *Aquaculture* 2015, **448**:186-195.
4. Drengstig A, Bergheim A: **Commercial land-based farming of European lobster (*Homarus gammarus L.*) in recirculating aquaculture system (RAS) using a single cage approach**. *Aquacultural Engineering* 2013, **53**:14-18.
5. Behringer DC, Butler MJ, Stentiford GD: **Disease effects on lobster fisheries, ecology, and culture: overview of DAO Special 6**. *Diseases of Aquatic Organisms* 2012, **100**(2):89-93.
6. Shields JD: **Diseases of spiny lobsters: a review**. *Journal of Invertebrate Pathology* 2011, **106**(1):79-91.
7. Stentiford GD, Neil DM: **Diseases of *Nephrops* and *Metanephrops*: A review**. *Journal of Invertebrate Pathology* 2011, **106**(1):92-109.
8. Cawthorn RJ: **Diseases of American lobsters (*Homarus americanus*): A review**. *Journal of Invertebrate Pathology* 2011, **106**(1):71-78.
9. Small HJ: **Advances in our understanding of the global diversity and distribution of *Hematodinium spp.* – Significant pathogens of commercially exploited crustaceans**. *Journal of Invertebrate Pathology* 2012, **110**(2):234-246.

10. Shields JD, Behringer DC: **A new pathogenic virus in the Caribbean spiny lobster *Panulirus argus* from the Florida Keys**. *Diseases of Aquatic Organisms* 2004, **59**(2):109-118.

11. Stebbing PD, Pond MJ, Peeler E, Small HJ, Greenwood SJ, Verner-Jeffreys D: **Limited prevalence of gaffkaemia (*Aerococcus viridans var. homari*) isolated from wild-caught European lobsters *Homarus gammarus* in England and Wales**. *Diseases of Aquatic Organisms* 2012, **100**(2):159-167.

12. Bateman KS, Tew I, French C, Hicks RJ, Martin P, Munro J, Stentiford GD: **Susceptibility to infection and pathogenicity of White Spot Disease (WSD) in non-model crustacean host taxa from temperate regions**. *Journal of Invertebrate Pathology* 2012, **110**(3):340-351.

13. Bateman KS, Munro J, Uglow B, Small HJ, Stentiford GD: **Susceptibility of juvenile European lobster *Homarus gammarus* to shrimp products infected with high and low doses of white spot syndrome virus**. *Diseases of aquatic organisms* 2012, **100**(2):169-184.

14. Verbruggen B, Bickley LK, Santos EM, Tyler CR, Stentiford GD, Bateman KS, van Aerle R: ***De novo* assembly of the *Carcinus maenas* transcriptome and characterization of innate immune system pathways**. *BMC Genomics* 2015, **16**:458.

15. **Fastqc. a quality control tool for high throughput sequence data** [http://www.bioinformatics.babraham.ac.uk/projects/fastqc]

16. Bolger AM, Lohse M, Usadel B: **Trimmomatic: A flexible trimmer for Illumina Sequence Data**. *Bioinformatics* 2014.

17. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al*: **Full-length transcriptome assembly from RNA-Seq data without a reference genome**. *Nature Biotechnology* 2011, **29**(7):644-652.

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

19. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences**. *Bioinformatics* 2006, **22**(13):1658-1659.

20. Gurevich A, Saveliev V, Vyahhi N, Tesler G: **QUAST: quality assessment tool for genome assemblies**. *Bioinformatics* 2013, **29**(8):1072-1075.

21. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications**. *BMC Bioinformatics* 2009, **10**:421-421.

22. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M *et al*: ***De novo* transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity**. *Nature Protocols* 2013, **8**(8):10.1038/nprot.2013.1084.

23. Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching**. *Nucleic Acids Research* 2011, **39** (Web Server issue):W29-37.

24. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J *et al*: **Pfam: the protein families database**. *Nucleic Acids Research* 2014, **42** (Database issue):D222-230.

25. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL: **Predicting transmembrane protein topology with a hidden markov model: application to complete genomes**. *Journal of Molecular Biology* 2001, **305**(3):567-580.

26. Petersen TN, Brunak S, von Heijne G, Nielsen H: **SignalP 4.0: discriminating signal peptides from transmembrane regions**. *Nature Methods* 2011, **8**(10):785-786.

27. Parra G, Bradnam K, Korf I: **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes**. *Bioinformatics* 2007, **23**(9):1061-1067.

28. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research**. *Bioinformatics* 2005, **21**(18):3674-3676.

29. Huson DH, Mitra S, Ruscheweyh HJ, Weber N, Schuster SC: **Integrative analysis of environmental sequences using MEGAN4**. *Genome Research* 2011, **21**(9):1552-1560.

30. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nature Methods* 2012, **9**(4):357-359.

31. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome**. *BMC Bioinformatics* 2011, **12**:323.

32. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics* 2010, **26**(1):139-140.

33. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: **KAAS: an automatic genome annotation and pathway reconstruction server**. *Nucleic Acids Research* 2007, **35**(Web Server issue):W182-185.

34. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs**. *Bioinformatics* 2015, **31**(19):3210-3212.

35. **Trinotate** [http://trinotate.sourceforge.net/]

36. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW: **Reagent and laboratory contamination can critically impact sequence-based microbiome analyses**. *BMC Biology* 2014, **12**.

37. Lehane MJ: **Peritrophic matrix structure and function.** *Annual Review of Entomology* 1997, **42**:525-550.

38. Kingsolver MB, Huang Z, Hardy RW: **Insect antiviral innate immunity: pathways, effectors, and connections**. *Journal of Molecular Biology* 2013, **425**(24):4921-4936.

39. Armitage SA, Peuss R, Kurtz J: **Dscam and pancrustacean immune memory - A review of the evidence**. *Developmental and Comparative Immunology* 2014.**48**(2): 315-323.

40. Lemaitre B, Hoffmann J: **The host defense of *Drosophila melanogaster***. *Annual Review of Immunology* 2007, **25**:697-743.

41. Udompetcharaporn A, Junkunlo K, Senapin S, Roytrakul S, Flegel TW, Sritunyalucksana K: **Identification and characterization of a QM protein as a possible peptidoglycan recognition protein (PGRP) from the giant tiger shrimp *Penaeus monodon***. *Developmental & Comparative Immunology* 2014, **46**(2):146-154.

42. Li X, Cui Z, Liu Y, Song C, Shi G: **Transcriptome Analysis and Discovery of Genes Involved in Immune Pathways from Hepatopancreas of Microbial Challenged Mitten Crab *Eriocheir sinensis***. *PloS One* 2013, **8**(7):e68233.

43. Myllymäki H, Valanne S, Rämet M: **The Drosophila Imd Signaling Pathway**. *The Journal of Immunology* 2014, **192**(8):3455-3462.

44. Chen WY, Ho KC, Leu JH, Liu KF, Wang HC, Kou GH, Lo CF: **WSSV infection activates STAT in shrimp**. *Developmental and Comparative Immunology* 2008, **32**(10):1142-1150.

45. Liu WJ, Chang YS, Wang AH, Kou GH, Lo CF: **White spot syndrome virus annexes a shrimp STAT to enhance expression of the immediate-early gene ie1**. *Journal of Virology* 2007, **81**(3):1461-1471.

46. Ren Q, Huang Y, He Y, Wang W, Zhang X: **A white spot syndrome virus microRNA promotes the virus infection by targeting the host STAT**. *Scientific Reports* 2015, **5**:18384.

47. Tassanakajon A, Somboonwiwat K, Supungul P, Tang S: **Discovery of immune molecules and their crucial functions in shrimp immunity**. *Fish & Shellfish Immunology* 2013, **34**(4):954-967.

48. Rolland JL, Abdelouahab M, Dupont J, Lefevre F, Bachere E, Romestand B: **Stylicins, a new family of antimicrobial peptides from the Pacific blue shrimp *Litopenaeus stylirostris***. *Molecular Immunology* 2010, **47**(6):1269-1277.

49. Destoumieux D, Bulet P, Loew D, Van Dorsselaer A, Rodriguez J, Bachere E: **Penaeidins, a New Family of Antimicrobial Peptides Isolated from the Shrimp *Penaeus vannamei* (Decapoda)**. *Journal of Biological Chemistry* 1997, **272**(45):28398-28406.

50. Destoumieux D, Munoz M, Cosseau C, Rodriguez J, Bulet P, Comps M, Bachere E: **Penaeidins, antimicrobial peptides with chitin-binding activity, are produced and stored in shrimp granulocytes and released after microbial challenge**. *Journal of Cell Science* 2000, **113**(3):461-469.

51. Sutthangkul J, Amparyup P, Charoensapsri W, Senapin S, Phiwsaiya K, Tassanakajon A: **Suppression of Shrimp Melanization during White Spot Syndrome Virus Infection**. *Journal of Biological Chemistry* 2015, **290**(10):6470-6481.

52. Wang PH, Huang T, Zhang XB, He JG: **Antiviral defense in shrimp: From innate immunity to viral infection**. *Antiviral Research* 2014.

53. Verbruggen B, Bickley L, van Aerle R, Bateman K, Stentiford G, Santos E, Tyler C: **Molecular Mechanisms of White Spot Syndrome Virus Infection and Perspectives on Treatments**. *Viruses* 2016, **8**(1):23.

54. He Y, Zhang X: **Comprehensive characterization of viral miRNAs involved in white spot syndrome virus (WSSV) infection**. *RNA Biology* 2012, **9**(7):1019-1029.

55. Lee YS, Nakahara K, Pham JW, Kim K, He Z, Sontheimer EJ, Carthew RW: **Distinct Roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA Silencing Pathways**. *Cell* 2004, **117**(1):69-81.

56. Mercer J, Schelhaas M, Helenius A: **Virus entry by endocytosis**. *Annual Review of Biochemistry* 2010, **79**:803-833.

57. Huang J, Li F, Wu J, Yang F: **White spot syndrome virus enters crayfish hematopoietic tissue cells via clathrin-mediated endocytosis**. *Virology* 2015, **486**:35-43.

58. Li DF, Zhang MC, Yang HJ, Zhu YB, Xu X: **Beta-integrin mediates WSSV infection.** *Virology* 2007, **368**:122-132.

59. Wang XW, Xu YH, Xu JD, Zhao XF, Wang JX: **Collaboration between a soluble C-type lectin and calreticulin facilitates white spot syndrome virus infection in shrimp**. *Journal of Immunology* 2014, **193**(5):2106-2117.

60. Sritunyalucksana K, Wannapapho W, Lo CF, Flegel TW: **PmRab7 is a VP28-binding protein involved in white spot syndrome virus infection in shrimp**. *Journal of Virology* 2006, **80**(21):10734-10742.

61. Longdon B, Brockhurst MA, Russell CA, Welch JJ, Jiggins FM: **The Evolution and Genetics of Virus Host Shifts**. *PLoS Pathogens* 2014, **10**(11):e1004395.

62. Woolhouse M, Scott F, Hudson Z, Howey R, Chase-Topping M: **Human viruses: discovery and emergence**. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2012, **367**(1604):2864-2871.

63. Koenemann S, Jenner RA, Hoenemann M, Stemme T, von Reumont BM: **Arthropod phylogeny revisited, with a focus on crustacean relationships**. *Arthropod structure & development* 2010, **39**(2-3):88-110.

64. Stentiford GD: **European Union Research Laboratory meeting 2015**. Personal communication 2015.

Chapter 6

RNA sequencing of the shore crab (*Carcinus maenas*) after White Spot Syndrome Virus exposure points to endocytosis regulation as source of resistance to infection.

Supplementary material available in the ‘Chapter 6’ folder on the DVD

Chapter 6: RNA sequencing of the shore crab (*Carcinus maenas*) after White Spot Syndrome Virus exposure suggests endocytosis regulation as source of resistance to infection.

6.1 Abstract

Shrimp farming and other implementations of crustacean aquaculture have experienced large growth over the years. The sector suffers from disease epidemics that kill a large portion of production and inhibits further development. Since its emergence in the 1990s the White Spot Syndrome Virus (WSSV) has caused billions of dollars' worth of damage, making it the most devastating virus to impact the shrimp farming sector. The pathogenicity of WSSV is high, reaching 100 % mortality within 3-10 days in penaeid shrimp. However, alone amongst tested aquatic crustacea, the shore crab *Carcinus maenas* was shown to be relatively resistant WSSV. To uncover the basis of this resistance we infected *C. maenas* with WSSV and measured mRNA and miRNA over a time series of 7 time points during a period of 28 days. *C. maenas* was confirmed to be resistant to WSSV infection as no mortalities occurred during the infection period and transcription of WSSV transcripts was only observed after 7 days suggesting replication, and in only five out of 28 samples in total. A *de novo* assembly of the transcriptome was conducted resulting in 129,049 transcripts, of which 30 % could be annotated. Differentially expressed transcripts were identified in every time point: in the first 12 hours post exposure we observed decreased expression of important regulators of endocytosis. It is established that WSSV enters the host cells through endocytosis and that interactions between the viral protein VP28 and Rab7 are important in successful infection, thus changes in expression of these components can have major impact on viral entry into host cells. Additionally we observed an increased expression of transcripts involved in RNA interference across many time points, indicating a longer term response to initial viral exposure. miRNA sequencing resulted in several miRNAs that were differentially expressed, including a novel *C. maenas* miRNA that was significantly downregulated in every WSSV exposed individual, suggesting that it may play an important role in mediating the response of crabs to virus. *In silico* target prediction pointed to involvement in endocytosis regulation. Taken together we hypothesize that the origin of *C. maenas* resistance to WSSV stems from the prevention of viral entry by endocytosis, a process probably regulated through miRNAs, resulting in inefficient uptake of virions.

6.2 Introduction

The crustacean aquaculture sector, and shrimp farming in particular, is an important part of global food production. According to statistics of the Food and Agriculture Organization (FAO) of the United Nations the sector has consistantly grown over the last years and it is likely to continue to do so in the future [1]. A major obstacle to further growth of the industry is the prevalence of diseases [2]. Over the last decade the most devastating disease in crustaceans has been caused by White Spot Syndrome Virus (WSSV) infection, which first emerged in the 1990s [3]. The host range of WSSV is large and is thought to comprise most aquatic crustacean species, which includes the commercially important penaeid shrimp [4].Since its emergence White Spot Disease (WSD) has had a major impact on global shrimp aquaculture [5]. It was estimated that to date the losses due to this disease have accumulated to over US $ 10bn, with approximately 1 US $ bn in losses added every year [6-8]. WSD is caused by the White Spot Syndrome Virus (WSSV), an enveloped dsDNA virus belonging to the family of the *Nimaviridae*. Upon infection hosts display lethargic behaviour, reduced food consumption, discoloration of tissues and a loosening cuticle [9, 10]. Some hosts develop white spots, the clinical sign after which WSSV was named. Penaeid shrimps generally suffer high mortality rates as a result of WSD, often close to 100 % within 3-10 days post infection [11].

The host range of WSSV extends beyond Penaeid shrimp and includes most decapod crustaceans [12], but the susceptibility and pathogenicity of WSD varies between these different hosts [9]. Bateman *et al.* 2012 [9] tested the susceptibility of various aquatic crustacean species to WSSV infection and grouped species into three categories: the first group contained highly susceptible species including penaeid shrimp and crabs such as *Eriocheir sinensis*. The European lobster *Homarus gammarus*, edible crab *Cancer pagurus* and the Norway lobster *Nephrops norvegicus* were classified as moderately susceptible, and only one species was placed in the low susceptibility group: the shore crab *Carcinus maenas*. The observation that the disease susceptibility of *C. maenas* to WSSV is relatively high offers an opportunity to investigate what mechanisms operate in this species to facilitate this enhanced resistance. Identifying mechanism (s) for resistance to WSSV infection in *C. maenas* might provide information for developing methodologies to prevent or treat WSD in more susceptible and economically important species like penaeid shrimps.

Over the years there has been extensive research into WSSV infection, particularly in penaeid shrimp. This work has resulted in identification of biological processes and molecular interactions that are significant in the WSSV infection process. A comprehensive review on this was published by Verbruggen *et al.* 2016 [13]. Like most viruses, for WSSV it has been established that the virus enters the host cell through Clathrin-mediated endocytosis after binding to proteins on the host cell surface (e.g. integrin [14]) [15]. Once inside, expression of the WSSV genome takes place in the nucleus under action of host transcription factors [13, 16-18]. Viral proteins influence the host pathways in order to maintain a cellular environment conducive to viral replication and virion assembly [13]. *C. maenas* resistance to WSSV infection could emerge from disruption of any one of these infection stages, although often earlier stages such as viral endocytosis are more likely. In addition to the interactions involving proteins, there has been research into the role of miRNAs in viral infection and it is now well established that miRNAs are important for successful replication [13, 19, 20]. miRNAs are small, ~21 nucleotides (nt), non-coding RNA molecules that are produced from pre-miRNA and regulate gene expression at the post-transcriptional level through association with the RISC complex. This can range from fine-tuning of gene expression to complete inhibition said gene. Viral or host miRNAs can impact the infection process on several levels, including: viral miRNAs targeting viral or host proteins and host miRNAs targeting host or viral proteins. To date, WSSV has been shown to express 89 miRNAs [13, 21]. A characterization by He *et al.* 2012 [22] showed that most viral miRNAs could target viral genes (e.g. WSSV-miR144, WSSV-miR164 and WSSVmiR211) and others that might influence host genes due to lack of viral targets (e.g. WSSV-miR36-5p, WSSV-miR36-3p and WSSV-miR72) [22]. Conversely host miRNAs are an important part of antiviral defence and it has been shown that shrimp miR-7 can target WSSV early gene *wsv477*, reducing viral replication [23, 24]. It is likely that in *C. maenas* miRNAs play also an important role in the resistance to WSSV.

The experiments described in this chapter aimed to characterize the response of *C. maenas* to WSSV through the study of temporal changes in mRNA and miRNA expression in WSSV injected crabs compared to crabs injected with saline solution. In order to capture both short term and long term effects, sampling was performed over a period of 28 days. Expression of mRNA and miRNA in gill samples were measured with next generation sequencing technology. After sequencing the RNA-sequencing data was used to assemble and annotate a transcriptome. A previously published *C. maenas* transcriptome, Verbruggen *et al.* 2015 [25], was used to verify the assembly. For the small RNA sequencing data, a genome scaffold (Chapter 4) was used to identify miRNA precursors, confirming their origins. WSSV replicated successfully in several samples as verified by mapping sequencing data to the WSSV. Differentially expressed transcripts and miRNAs between WSSV injected and control samples were identified over time points. Through functional annotation and pathway analysis, relevant biological processes underlying the infection processes and *C. maenas* apparent resistance to WSD were identified. After identification of the important factors, their suitability for treatment development that could be applied to other species was discussed.

6.3 Materials and Methods

6.3.1 Animals and experimental set up

All experimental trials were performed within the biosecure exotic disease facility at the Cefas Weymouth laboratory and utilised local, filtered and UV treated seawater. Day length was set at 14 h day/ 10 h night with a 30 min transition to simulate dusk and dawn. The flow rate was set a 3-4 L /min and salinity during the experimental period remained constant at 35ppt. A temperature of 20 °C was maintained throughout the trial.

One hundred and twenty shore crabs approximately 25mm - 60mm in carapace width, were collected from the shoreline at Newton’s Cove, Weymouth, UK (50°34’ N, 02°22’ W). All animals utilised in this experiment appeared to be healthy. To prevent conflict, shore crabs were housed individually in custom-made compartments within large trough tanks, with individuals separated but sharing the same water supply, and a maximum of 15 crabs were housed per trough. All animals were acclimatised to the trial start conditions for a minimum of one week before trials commenced.

6.3.2 Preparation of viral inoculum and challenge trials

Viral inoculates of WSSV were obtained from the OIE Reference Laboratory for White Spot Syndrome Virus (WSSV) at the University of Arizona, USA. The OIE isolate of WSSV (UAZ 00-173B) was generated in *Litopenaeus vannamei* [26] from an original outbreak of WSD in *Fenneropeneaus chinensis* [26] in China in 1995. Subsequent passage of this isolate into Specific Pathogen Free (SPF) *L. vannamei* held at the Cefas Weymouth laboratory have demonstrated continued virulence of this isolate. All challenges reported utilised WSSV-infected *L. vannamei* carcasses generated within the Cefas Weymouth laboratory. As such, WSSV-infected shrimp carcasses were prepared by injection of the UAZ 00-173B isolate into SPF *L. vannamei* obtained from the Centre for Sustainable Aquaculture Research (CSAR) at the University of Swansea, United Kingdom. Individual *L. vannamei* were inoculated via intra-muscular injection of the diluted viral homogenate at a rate of 10 l g-1 shrimp weight and the initial viral loading was 2206708.6/mg. Following incubation, dead and moribund shrimp were removed from the experimental tanks and infection with WSSV was confirmed using histopathology, transmission electron microscopy (TEM) and PCR, as appropriate (see below for techniques). Remaining tissues were stored at –80 °C until required. Confirmed infected carcasses were used to prepare inoculums. Infected shrimp carcasses were macerated in isolated conditions using a sterile razor blade prior to homogenisation in 2% sterile saline (4mL of saline per gram of minced tissue) using a blender until tissues were liquefied. The homogenate was centrifuged at 5,000 x g for 20 minutes at 4°C to pellet solid debris prior to the supernatant being diluted 1:20 with sterile saline and filtered (0.2µm Minisart syringe filter, Sartorius Stedim Biotech GmbH, Germany) to form the inoculum for the injection studies.

6.3.3 WSSV Challenge

120 shore crabs (*Carcinus maenas*), were randomly allocated into two treatment groups (n=60 per treatment) and water temperature was held constant at 20 ˚C for the duration of the trial. Crabs were injected at the base of the second walking leg with saline or a single dose of the diluted WSSV homogenate at a rate of 10µl g^{-1} wet body weight on Day 0. This corresponded to an initial viral loading of 2.19 E+07virions/ gram. Crabs were sampled at 0, 6, 12, 24, 48 hours, 7 days, 14 days and 28 days post-injection. Gill, hepatopancreas, heart, gonad, connective tissues and muscle were dissected and placed into histological cassettes and fixed immediately in Davidson's seawater fixative. For molecular analyses of viral replication, gill samples were removed and placed into tubes containing 100% ethanol. For sequencing analysis, gill tissues were dissected, immediately snap frozen in liquid nitrogen and stored at -80°C until analysis. For the purpose of this thesis, I will focus exclusively on the analysis of the sequencing data. All other analyses were conducted independently as part of the collaborative programme of work with Cefas but are outside the scope of this thesis.

6.3.4 RNA extraction, library preparation and sequencing

RNA was extracted using Qiagen's miRNeasy mini kit, with on column DNase digestion, according to the manufacturer's instructions. RNA quality was measured using an Agilent 2100 Bioanalyzer, using the RNA 6000 nano kit (Agilent Technologies, CA, USA).

mRNA sequencing

cDNA libraries for each individual crab gill, 63 samples in total, were constructed using 1.25 µg of RNA. ERCC Spike-In control mixes (Ambion via Life Technologies, Paisley, UK) were added to control for technical variation during sample preparation and sequencing, and analysed using the manufacturer's guidelines. mRNA purification was performed via poly (A) enrichment using NEB Next Poly(A)mRNA magnetic isolation module (NEB, Herts, UK). cDNA libraries were constructed using Epicentre's ScriptSeq v2 RNA-seq library preparation kit (Illumina, CA, USA) with Agencourt AMPure XP system (Beckman Coulter, Bucks, UK) to purify cDNA and Epicentre's ScriptSeq Index PCR primers to produce barcoded libraries. This enabled multiplexing of all samples across four lanes. All libraries were diluted to 10nM. Sequencing was performed on an Illumina HiSeq 2500 in standard mode with 100 bp paired-end read module.

Small RNA sequencing

cDNA libraries for each individual crab gill were constructed using 1.0 µg of total RNA, using the Illumina's Tru-seq small RNA sample preparation kit (Illumina, CA, USA) with indexed adapters 1 - 36. This enabled multiplexing of samples across two lanes. All libraries were diluted to 2nM, and sequencing was performed on an Illumina HiSeq 2500 rapid run mode with 50 bp single read module.

6.3.5 Transcriptome assembly and annotation

Prior to transcriptome assembly the reads were trimmed in order to remove bases with low confidence and adapter contamination. The paired-end read libraries were trimmed using Trimmomatic with the following settings: removal of Illumina adapters, removal of the first 12 bases, removal of the last 3 bases, removing bases based on a sliding window of 4 and minimal Phred quality of 20 and finally applying a minimal remaining length of 25. Only reads where both members of the pair passed the quality trimming were kept for *de novo* transcriptome assembly. Supplementary file S1 shows the number of reads remaining after quality-trimming for every sample. *De novo* transcriptome assembly was performed with the Trinity *de novo* assembler using reads from all samples pooled together, requiring minimal kmer coverage of 10. After assembly, transcriptome redundancy was reduced with CD-HIT-EST v4.6, default settings [27]. Assembly statistics like average length and N50 were calculated with quast-2.3 [28]. In order to investigate the presence of conserved genes the cegma 2.5 tool was used [29]. Furthermore, the presence of a selection of near-universal single-copy orthologs, representative of arthropods, was investigated using BUSCO v.1.1 (tBLASTn threshold of 1e^{-10}). The resulting transcripts were annotated using Trinotate which encompasses several steps including: ORF identification with TransDecoder [30], BLAST for sequence similarity to SwissProt (November 2015, protein domain identification with HMMER [31], identification of signal peptides with signal [32] and non-coding RNA with RNAmmer [33]. The KEGG annotation server (KAAS [34]) was used to assign KEGG Orthologies (KO) to the transcripts based on bi-directional best hits, enabling mapping to pathways contained in the KEGG database. Additionally, BLASTx was used to identify sequence similarities to the NCBI non-redundant protein database (November 2015) with an e-value threshold of 1e^{-5}. These BLAST results were used as input for MEGAN5 [35] for taxonomic classification of transcripts. Transcripts that mapped to metazoan taxonomies were treated as originating from *C. maenas*. The assembled transcriptome was compared to a previously assembled transcriptome for *C. maenas* (Verbruggen *et al.* 2015; Chapter 3) through both BLASTn and BLASTp searches with an e-value threshold of 1e^{-10}. This comparison was done for both the complete and metazoan filtered transcriptomes. Functional annotation was provided through BLAST2GO® PRO (September 2015) with the BLASTx to NCBI-nr results as input and default settings. KEGG orthology identifiers were assigned through the KEGG Automatic Annotation Server [34].

6.3.6 Detection and assembly of WSSV transcriptome

In order to detect the presence of replicating WSSV, the quality-trimmed RNA-sequencing reads were aligned to the WSSV genome (Chinese isolate; WSSV-CN, Genbank AF332093.3 [36]) using bowtie2 version 2.2.3 with default settings [37]. Overall, alignment rates were used to identify samples containing replicating WSSV (> 1 % of total reads). Reads that mapped successfully were extracted using samtools version 0.1.19 [38]. The alignment of reads to the WSSV-CN genome was visualized in IGV [39]. Coverage of the WSSV-CN genome was calculated using bedtools v2.17.0 [40]. Transcriptome assemblies based on the WSSV reads were generated by Trinity v2.1.1, both *genome guided* and *de novo*, and Cufflinks v2.1.1 [41]. Trinity assemblies were performed with either default settings or with –*min_kmer_cov* and –*min_glue* changed to values ranging from 5 to 200). Transcripts in the assemblies were aligned to the WSSV-CN genome using BLAT v36 [42], not allowing for introns. BLAT alignments were visualized in IGV along with the WSSV-CN ORF annotation, based on the AF332093.3 Genbank file. Calculation of WSSV-CN ORF expression values were based the alignment of WSSV mapping reads to the WSSV-CN genome and estimated by RSEM 1.2.21 [43]. Differential expression calculations between WSSV samples were performed by using RSEM output in EBSeq v1.1.5 comparing virus-injected to control samples. Transcripts with a differential expression probability higher than 0.95 were treated as differentially expressed.

6.3.7 Differential transcript expression and gene ontology enrichment

Sequence reads from every sample were mapped to the assembled transcriptome, including non-metazoan transcripts, using bowtie2 version 2.2.3 [37]. The alignment was used to derive expression counts for transcripts in the transcriptome using RSEM 1.2.21 [43]. Expression counts were listed in transcripts per million (TPM) and Fragments Per Kilobase of transcript per Million mapped reads (FPKM). From these expression counts, differentially-expressed transcripts were identified using EBSeq 1.1.5 [44], which is coupled to RSEM 1.2.21 [43, 45]. Differential expression analysis was performed at every time point, by comparing WSSV exposed crabs to time-matched control samples. Transcripts with a differential expression probability higher than 0.95 were treated as differentially expressed. The lists of differentially-expressed genes for each time point were analysed for functional enrichment of Gene Ontology categories using Blast2GO® PRO [46]. Terms with an FDR below 0.05 were considered significantly enriched after applying Benjamini and Hochberg correction for multiple testing.

6.3.8 miRNA annotation and expression

miRNA sequencing data were processed through the mirdeep2 version 0.0.7 package [47] . The first stage involved trimming of adapters, removal of t/rRNA sequences, filtering reads less than 18 bp in length and mapping to the *C. maenas* genome (Chapter 4) and WSSV-CN genome with the mapper script. Within the mirdeep2 package this mapped output was used to predict existing miRNAs from closely related species *Marsupenaeus japonicus*, *Tribolium castaneum* and *Daphnia pulex,* as present in miRbase (accessed January 2016) [48-51]. Novel miRNA prediction was based on the *C. maenas* draft genome. Potential WSSV miRNAs identified by Huang *et al.* 2014 [21] were used to identify miRNAs of viral origin for each sample. Expression of predicted miRNAs was derived through the mirdeep2 quantifier script. A final list of expressed miRNAs was generated by clustering the identified miRNAs with CD-HIT-EST v4.6 [27] (similarity threshold 0.95), thus eliminating redundancy. Differentially-expressed miRNAs were identified for every time point by edgeR-GLM v3.4.2 [52], after filtering for miRNAs with at least 10 counts per million across half the samples in a time point to ensure that only consistently expressed miRNAs were included in the analysis. miRNAs with a FDR below 0.05 were considered differentially expressed. Target prediction of miRNAs was performed with MicroTar [53] and miRanda v3.3a [54] against the assembled *C. maenas* transcriptome. Free energy thresholds of -10 kcal/mol and -20 kcal/mol were applied to identify miRNA-mRNA associations.

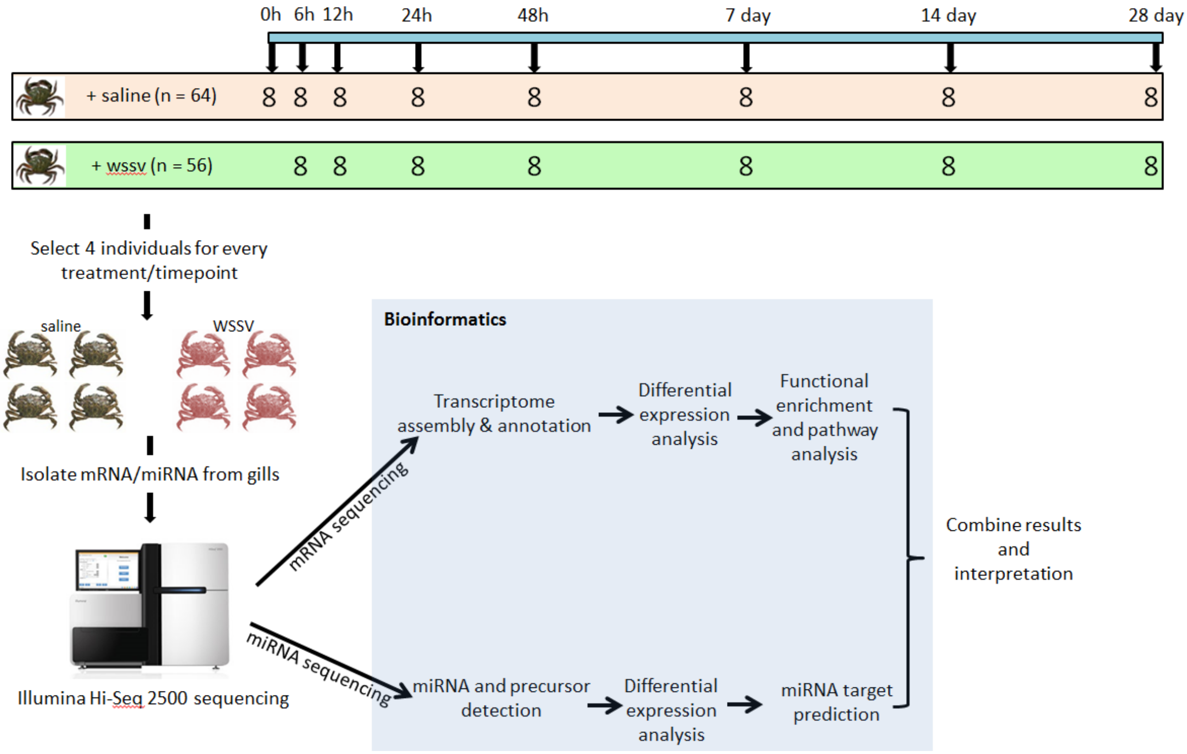

0h
6h
12h
24h
48h
7 day
14 day
28 day
+ saline (n = 64)
8 8 8 8 8 8 8 8
+ wssv (n = 56)
8 8 8 8 8 8 8
Select 4 individuals for every treatment/timepoint
saline
WSSV
Isolate mRNA/miRNA from gills
Illumina Hi-Seq 2500 sequencing
mRNA sequencing
miRNA sequencing
Bioinformatics
Transcriptome assembly & annotation
Differential expression analysis
Functional enrichment and pathway analysis
miRNA and precursor detection
Differential expression analysis
miRNA target prediction
Combine results and interpretation

Figure 1 Overview of experimental setup and analysis.

6.4 Results and discussion

Over the 28 days of the experiment there were no reported mortalities. Four individuals could be selected for every condition/time point combination for small- and mRNA isolation.

6.4.1 Sequencing and pre-processing

Illumina sequencing over four lanes yielded a total of 1,385,269,552 reads across the 63 samples (see Supplementary File S1). After quality-filtering, a total of 1,097,977,516 reads were retained with an average of 17,999,631 reads and standard deviation of 3,351,031 reads per gill library. Small RNA sequencing yielded 155,394,008 reads with an average of 2,589,900 ± 858,880 per sample. Pre-processing of the small RNA data (adapter removal, t/rRNA removal and size filtering) resulted in 73,696,599 reads being retained with an average of 1,228,276 ± 508,163 per sample. Small RNA read counts are summarised in Supplementary File S1 and overall sequencing statistics are summarized in Table 1.

Table 1 Illumina sequencing statistics

Dataset	Total	Average	Std Dev	Trimmed Total	Trimmed Average	Trimmed Std Dev
mRNA	1,385,269,552	22,680,492	4,086,594	1,097,977,516	17,999,631	3,351,031
smallRNA	155,394,008	2,589,900	858,880	73,696,599	1,228,276	508,163

6.4.2 De novo assembly and annotation

In order to generate a *de novo* transcriptome assembly representing all transcripts present in the gills of both control and infected organisms, including those only present as a response to WSSV infection, all sequences (from both control and infected samples) were combined. The assembly statistics as calculated through quast [28] are shown in Table 2. The assembled transcriptome contained 129,049 transcripts with a N50 of 1,437. In total, the transcriptome covered 39,323,954 nucleotides with a GC content of 43.18 %.

Table 2 *De novo* transcriptome assembly statistics

Assembly statistic	Value
Contigs (>= 200 bp)	129,049
Contigs (>= 1000 bp)	20,065
Median Contig length	395
Total length	39,323,954
GC %	43.18
N50	1,437
N75	839
L50	11,745
L75	23,310

The representation of a set of 248 strongly conserved genes amongst metazoan and eukaryotes, described by Parra *et al.* 2007 [29], in a transcriptome is a sign of high quality. Should a large portion of this set of transcripts be absent this can be indicative of experimental or assembly problems, the latter potentially solvable through alternative assembly strategies. Using the cegma tool on the transcriptome, 192 out of the 248 conserved genes (77.42 %) were identified completely (full length) and 237 (95.56 %) were identified at least partially (a fragment of the gene was present). Nearly all conserved eukaryotic genes thus have a counterpart in the transcriptome assembled here, although some have only been partially retrieved. In similar fashion Simão *et al.* 2015 compiled a set of benchmarking universal single copy orthologues (BUSCO) that can be expected in transcriptomes belonging to species of certain evolutionary lineages. For the *C. maenas* transcriptome the presence of the set of 2,675 arthropod BUSCOs was investigated. A total of 1,501 (56 %) BUSCO were completely covered in the transcriptome and a further 420 (16 %) were found in fragmented form. 28 % of arthropod BUSCOs remained unidentified in the *C. maenas* transcriptome (Table 3). The results from both cegma and BUSCO suggest that the assembled transcriptome is not complete and some transcripts have been fragmented during the assembly process. It is important to note that the current transcriptome is based on a single tissue, explaining why some of the orthologues were missing, as on average only about 1/3rd of all genes are expected to be expressed in any one individual tissue, while the remaining genes are likely to be functionally switched off or expressed at low levels [55].

Table 3 BUSCO group identification

BUSCO	**Count**
Complete Single-copy BUSCOs	1,201 (45 %)
Complete Duplicated BUSCOs	300 (11 %)
Fragmented BUSCOs	416 (16 %)
Missing BUSCOs	758 (28 %)
Total BUSCO groups searched	2,675

Transcript annotation

The gill transcriptome was annotated using several methods. The Trinotate annotation suite was used to identify conserved domains, signal peptides and transmembrane regions within the open reading frames identified in the *C. maenas* transcripts. Open reading frames with the potential to code for proteins/peptides were identified in 59,777 (46.3 %) of the assembled transcripts. Out of these predicted amino acid sequences, 18,508 (31.0 %) showed significant similarity to known proteins in the UniProt/Swissprot database. Furthermore, 17,005 (28.4 %) contained at least one Pfam domain. Lastly, 4,129 (6.9 %) potential signal peptides and 14,920 (25.0 %) transmembrane regions were identified (Supplementary File S2). In addition to the Trinotate suite, a BLASTx search was performed against the NCBI non-redundant protein database, yielding 30,090 (23.3 %) transcripts with a significant hit. These results were used to connect transcripts to Gene Ontology terms in BLAST2GO® PRO which resulted in 23,855 (18.5 %) connections. The KEGG annotation server (KAAS [34]) was used to assign KEGG Orthologies (KO) to the transcripts based on bi-directional best hits resulting in 5,857 annotations. KEGG is focussed on mammalian pathways and sequences and therefore the yield is expected to be lower than for other annotation methods. Overall, 45,967 (35.6 %) of transcripts could be assigned with at least one annotation using Trinotate. The overall annotation rate for *C. maenas* transcripts is low which raises the question on whether this is related to this specific dataset or rather common for sequencing work for aquatic crustaceans in general. In previous work, we produced a transcriptome for eight tissues of *C. maenas* [25] and a transcriptome for the nine tissues of the European lobster, *Homarus gammarus*. We compared these datasets to the one presented in this chapter and concluded that these rates are similar across experiments for *C. maenas* and for other aquatic crustacean species (Supplementary Information 1), likely reflecting the significant knowledge gap on sequencing data for aquatic crustaceans.

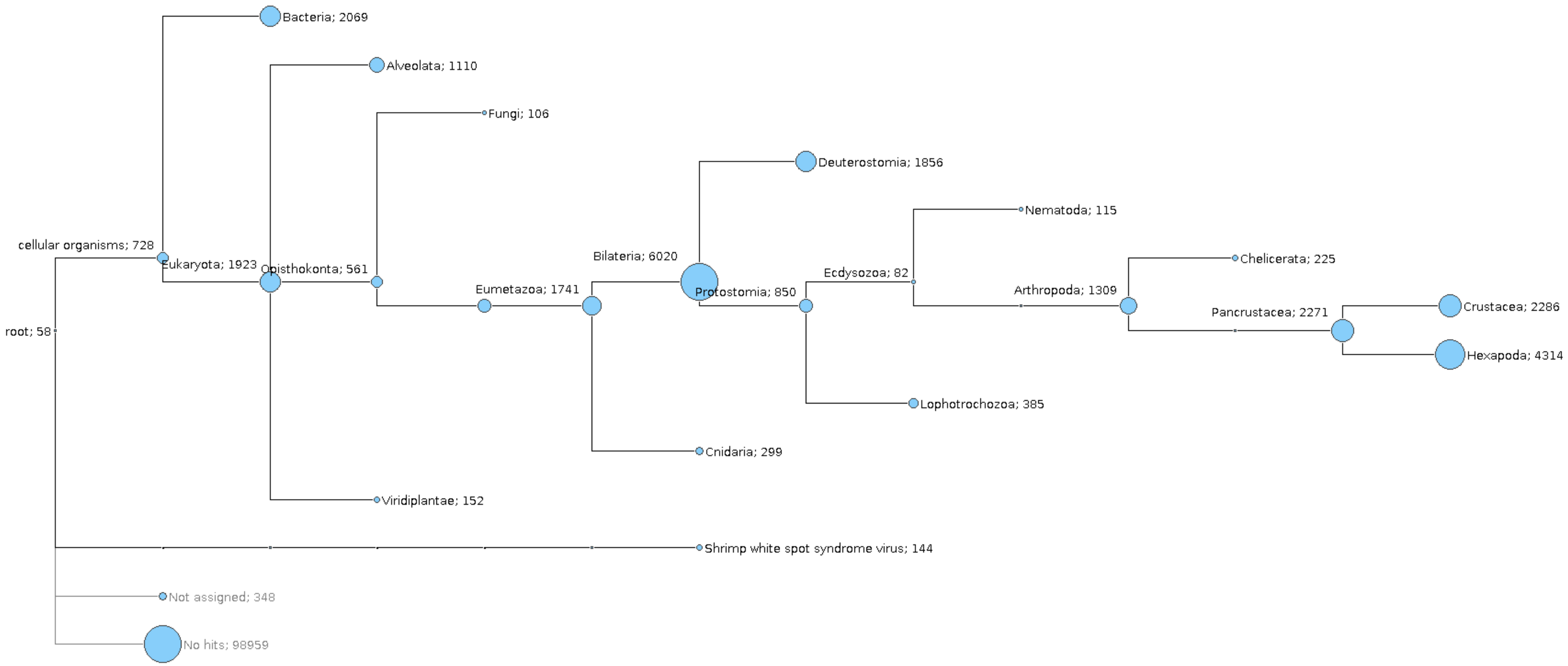


Figure 3 MEGAN5 Taxonomic tree with transcript counts. Numbers illustrate the number of transcripts representing each taxa. Transcripts were assigned to taxa depending on the mapping of the best BLASTx to NCBI-nr hit by MEGAN5. Within the metazoan taxon, the pancrustacea represented the largest taxonomic group. Minimal threshold for showing a taxon is 100 transcripts for taxon and subtaxa.

Taxonomy

It is expected that during isolation not only host/virus RNA is isolated but also RNA from other organisms present in the sample, including the microbiome located on the *C. maenas* gills. During library preparation and sequencing a distinction cannot be made between *C. maenas*, WSSV or RNA of different origin. Additionally RNA can be present in the kits used in the laboratory work [56]. A separation can be achieved bioinformatically by filtering for transcripts that show similarity to taxa that are neither host nor virus. Figure 2 shows a tree with the distribution of assembled transcripts over various taxa based on BLASTx annotation (tree generated with MEGAN5 [35]). The figure shows that WSSV transcripts are present in the final transcriptome. For multi-cellular organisms, most annotated transcripts are related to bilateria, Crustacea and hexapoda taxons. Because of the lack of genomic information within aquatic Crustacea it can be expected that large numbers of *C. maenas* transcripts map to hexapoda or bilateria taxa instead. As was stated in the last section, a large group of transcripts do not show any similarity to known sequences and therefore no biological interpretation can be given on either their function or the quality of the assembly for those transcripts. Any transcript that did not map to either WSSV or the metazoan taxon was removed from the assembly. After this filter, the 'metazoan' transcriptome encompassed 22,738 transcripts. This quantity is very close to those observed in Verbruggen *et al.* 2015, which identified 23,151 transcripts that mapped to metazoan being expressed in the gill tissue. A more detailed comparison of the 'metazoan' transcriptome in this Chapter compared to Verbruggen *et al.* 2015 is offered in the Supplementary Information 2. In comparison, only 14,947 'metazoan' transcripts appeared to be expressed in the gills of *H. gammarus* (see Chapter 5). Figure 2 also shows the presence of 144 WSSV transcripts in the transcriptome assembly, an indication that the viral genes were being expressed within the exposed crabs.

6.4.3 Detection of WSSV replication in RNA sequencing data

The MEGAN5 tree in Figure 2 illustrated the presence of WSSV transcripts in the transcriptome assembly, indicating that WSSV was successfully replicating in at least some individual gill samples. In order to detect which samples contained replicating virus, RNA sequencing reads from individual gill samples were mapped to the available WSSV-CN genome (Genbank: AF332093.3). Successful mapping of a significant number of sequencing reads to the viral genome suggested that within that individual viral mRNA was being produced and thus the virus had successfully infected the host. Only five samples contained a significant portion (> 1 %) of reads mapping to the viral genome: ws_36 (4. 87 %; 7 days), ws_37 (2.78 %; 7 days), ws_48 (8.08 %; 14 days), ws_50 (4.85 %; 28 days) and ws_51 (7.57 %; 28 days), and Sample ws_52 (0.33 %; 28 days) also had a small amount of mapped reads (Supplementary Table S 1). We used the WSSV reads to produce and characterize a WSSV transcriptome which could aid in annotation of the WSSV genome (see Supplementary information 3).

The low number of individual crabs where WSSV transcripts were found (5 out of 28) demonstrated the resistance of this species to WSSV infection. Where susceptible species like *P. monodon* and *L. vannamei* show mortality rates within 3-10 days and replication within hours of injection [57, 58], in *C. maenas* it was not until 7 days post-injection that replication became apparent. Additionally, not every crab contained replicating virus at the later time points. There was individual variation with only between 25-50 % of samples showing development of an infection.

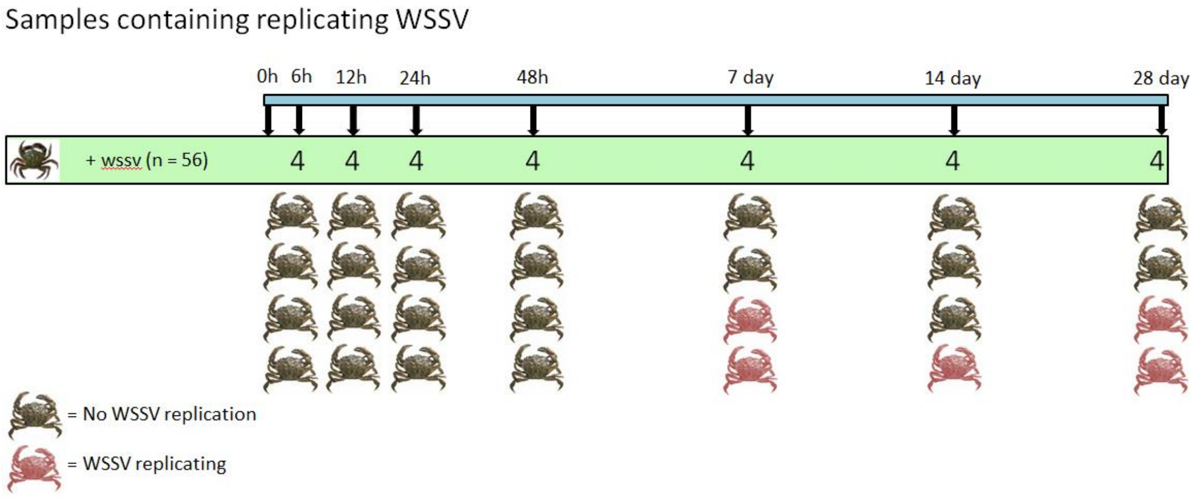
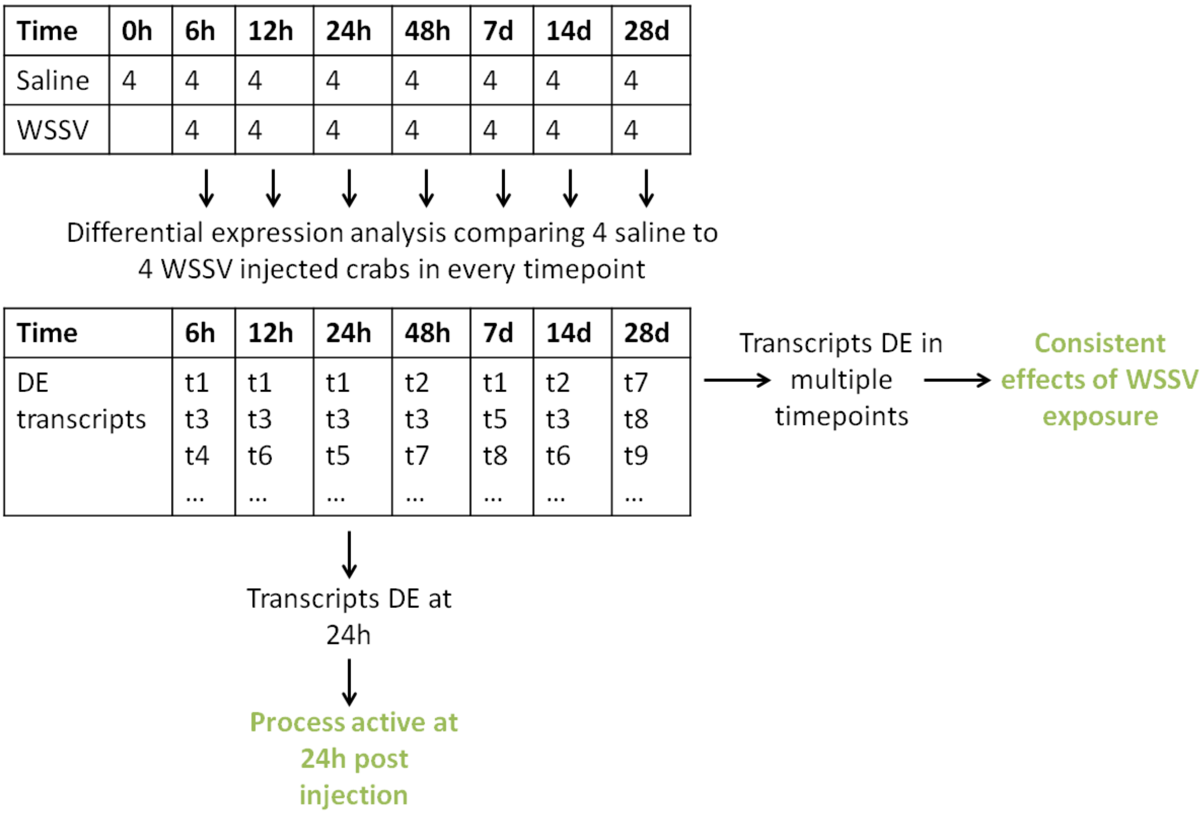


Figure 3 Individuals with replicating WSSV. RNA derived from WSSV was only detected in five individuals injected with the virus. Two individuals at 7 and 28 days post injection and one at 14 days post injection.

6.4.4 Differential gene expression

Differential expression analysis was carried out by comparing expression patterns of four saline injected crabs to four WSSV injected crabs in every time point (See Figure 4). The number of differentially expressed transcripts between the gills of WSSV injected crabs and their time-matched controls is shown in Table 4. Additionally the numbers of DE transcripts attributed to *C. maenas* (metazoan) and WSSV, as determined by MEGAN5, are shown. Differentially expressed (DE) transcripts ranged from 477 (147 metazoan) at 6 hours post injection to 1,968 (685) at 672 hours post injection. In agreement with the previous section, DE transcripts of WSSV origin are only detected in the later time points. However, the 336 hour time point only identified a single DE WSSV transcript, probably a result of there being only a single WSSV replicating sample as opposed two at 168 h and 672 h. The DE transcripts that were classified as neither WSSV nor metazoan can represent transcripts of non-host non-virus origin or novel transcripts from the host. Lists of differentially expressed transcripts at every time point are found in Supplementary File S2.



Time	0h	6h	12h	24h	48h	7d	14d	28d
Saline	4	4	4	4	4	4	4	4
WSSV		4	4	4	4	4	4	4

Differential expression analysis comparing 4 saline to 4 WSSV injected crabs in every timepoint

Time	6h	12h	24h	48h	7d	14d	28d
DE transcripts	t1 t3 t4 …	t1 t3 t6 …	t1 t3 t5 …	t2 t3 t7 …	t1 t5 t8 …	t2 t3 t6 …	t7 t8 t9 …

Transcripts DE in multiple timepoints
Consistent effects of WSSV exposure
Transcripts DE at 24h
Process active at 24h post injection

Figure 4 Differential expression analysis - single and multiple time points. Differential expression analysis starts by identification of DE transcripts in every time point. Analysis of DE transcripts in a time point can provide insight in processes that are active at that time post injection whereas transcripts that are DE in across multiple time points are indicative of changes that persist over time after WSSV exposure.

Table 4 Number of differentially expressed transcripts

Time (hours post injection)	Number of DE transcripts (downregulated/ upregulated)	Number of metazoan DE transcripts (#)	Number of WSSV DE transcripts (#)
6	477 (363 / 144)	147 (49 / 98)	0
12	1565 (944 / 621)	628 (136 / 492)	0
24	1318 (720 / 598)	519 (161 / 358)	0
48	500 (294 / 206)	181 (71 / 112)	0
168	938 (490 / 448)	265 (142 / 123)	97
336	587 (258 / 329)	217 (100 / 117)	1
672	1968 (907 / 1061)	685 (359 / 326)	76

To determine whether WSSV injected samples could be separated based on expression of DE transcripts we plotted expression values in FPKM on a heatmap and subjected the samples to a principal component analysis. The heatmap, Figure 5, shows that the WSSV replicating samples (ws_36, ws_37, ws_48, ws_50 and ws_51) cluster together. Similarly in the PCA plot, Figure 6, showed that the first principal component separates samples that contained replicating WSSV from the other samples. Less clearly, the second principal component appears to separate samples by time after injections as early time points generally have lower values. Figure 6 also illustrates that other than the WSSV replicating samples most WSSV injected samples cluster relatively close together, indicating that expression changes over time are not substantial. Analysis of the identity of DE transcripts within and across time points can reveal the biological basis of the separation of samples in the PCA plot and accordingly the responses of *C. maenas* to WSSV exposure.

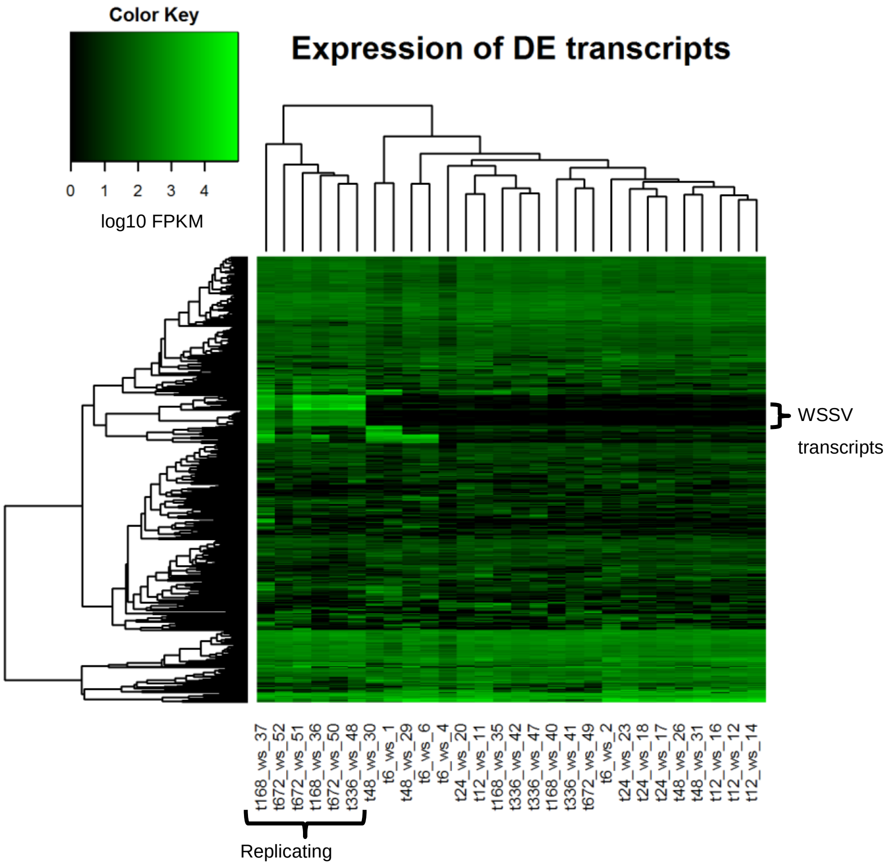


Figure 5 Heatmap of DE transcript expression values in WSSV injected samples. Expression of differentially expressed transcripts is represented by log10 FPKM values as derived by RSEM. Rows contain DE transcripts and columns WSSV injected samples. The samples that contain replicating WSSV cluster together, in addition the transcripts derived from WSSV are indicated on the heatmap.

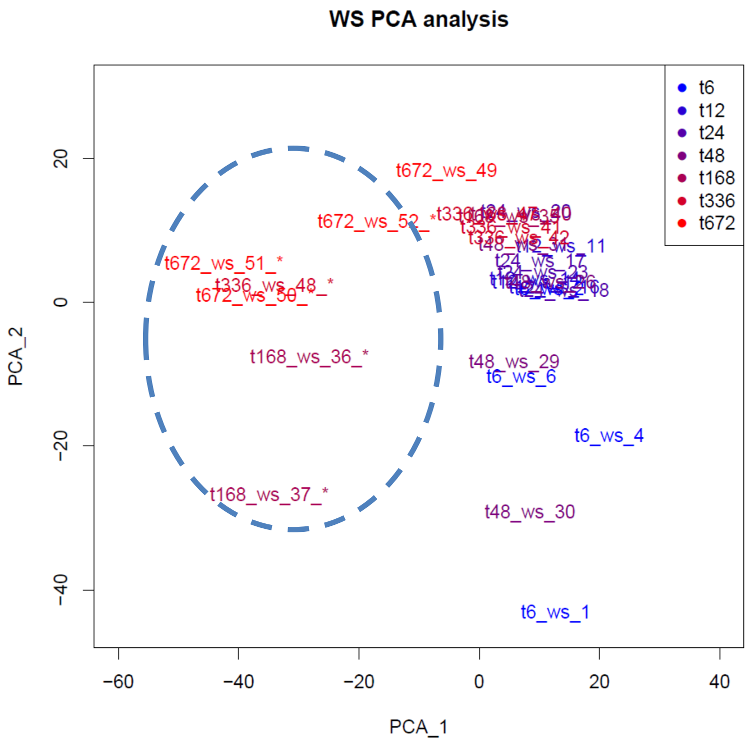


Figure 6 Principal component analysis of WSSV injected samples. Plot of the first two principal components identified in the expression of WSSV injected samples. Samples in the plot are labelled according to their time and sample ID. A colour gradient illustrates early to late time points. '*' symbols following a sample name indicate that the sample contained replicating WSSV. The ellipse illustrates the clustering of WSSV replicating samples along the first principal component.

The biological interpretation of the response of *C. maenas* to WSSV infection, as characterized by DE transcripts, can be examined either across all time points or for each time point specifically. Transcripts that are DE across all the time points allow for analysis the long term response of *C. maenas*, while analyses across individuals for any one time point can provide insight in what specifically happening at that time point. For the WSSV exposure, first the consistent effects as characterized by DE transcripts in multiple time points are discussed. This is followed by changes that occur in individual time points.

Differential expression across multiple time points

In order to characterize the response that occurs over the length of the experiment, transcripts that were DE across at least 4 of the 7 time points are listed in Table 5. The majority of the DE transcripts were annotated as predicted/putative or hypothetical proteins which complicates biological interpretation. Three transcripts have an annotation: mannose-binding protein (TR29999|c0_g1_i1), peritrophin (TR38732|c2_g1_i1) and argonaute 2 (TR48022|c1_g1_i2). Mannose-binding protein is a member of the innate immune system where it performs the role of recognizing carbohydrate patterns present on pathogens like bacteria and viruses [59]. Table 5 shows that this transcript is DE at the earliest time points (6h and 12h) and two of the later time points (168h and 336h). Mannose-binding protein expression thus changes during the initial contact with the virus and again at time points where replicating WSSV is detected, although not DE in the 672h time point. This indicates that the immune system of *C. maenas* responds to the presence of viral particles. Re-emergence of DE expression of Mannose binding-protein at later time points could be related to the release newly made virions in infected cells. Peritrophin is a structural protein that is an important part of the peritrophic matrix in insects [60], but in crustaceans it appears to be involved in oogenesis as shown by Khayat *et al.* 2001 [61]. However, a study on a perithropin-like protein in the fleshy prawn (*Fenneropenaeus chinensis*) showed expression in hemocytes, heart, stomach, intestine and gill after bacterial infection. This suggests that this protein may respond to recognize invading microorganisms and detect exposed chitin, and then trigger appropriate physiological or developmental response [62]. Since the expression of transcript TR38732|c2_g1_i1 was measured in *C. maenas* gill tissues upon pathogen exposure it is likely that the protein encoded by this transcript is functionally related to the peritrophin-like protein from *F. chinensis* rather than the structural role perithropin plays in oogenesis in the ovaries. Like mannose-binding protein, peritrophin is expressed in during both early and later time points. Argonaute2 is an important constituent of the RISC complex of the RNA interference (RNAi) pathway [63]. Argonaute2 binds to a short guide RNA (miRNA or siRNA) which directs RISC to complementary mRNAs that are targets for RISC-mediated gene silencing [64]. As discussed in Verbruggen *et al.* 2016 [13], RNAi through miRNAs is an important part in crustacean immune responses to viruses. Alterations in the expression of *C. maenas* Argonaute2 (TR48022|c1_g1_i2) appear during the first three time points (6-24h) and once more at 168h post WSSV injection. Indeed, the higher expression of *C. maenas* Argonaute 2 in WSSV injected samples is indicative of an immune response involving miRNAs. Thus overall the transcripts with annotation that are differentially expressed across multiple time points are all related to the immune response of *C. maenas* to WSSV. The results of transcripts with predicted/putative and hypothetical annotations are more difficult to interpret, just like differentially expressed *C. maenas* transcripts that did not yield any significant BLASTx results (see Supplementary File S2). However, their role in WSSV infection might be a ground for further investigation of the biological role of these transcripts.

Table 5 Metazoan transcripts DE across at least 4 time points

Transcript	6h	12h	24h	48h	168h	336h	672h	BLASTx – NCBI nr	e-value
TR45167|c3_g1_i1	2.62	-1.51		2.58		-1.78	-1.12	putative cuticle protein CPR151 [*Bombyx mori*]	2.00E-10
TR48022|c1_g1_i2	0.32	0.6	0.48		0.18			argonaute 2 [*Marsupenaeus japonicus*]	0
TR32896|c2_g2_i2			0.32	0.25	0.31		0.28	PREDICTED: terminal uridylyltransferase 7-like [*Takifugu rubripes*]	0
TR47978|c0_g1_i1		-2.05			1.38	-1.86	-2.37	hypothetical protein DAPPUDRAFT_312544 [*Daphnia pulex*]	1.00E-13
TR38732|c2_g1_i1	1.61	-2.34				-2.18	-2.26	peritrophin, partial [*Rimicaris exoculata*]	4.00E-27
TR12957|c0_g1_i1	1.24	-1.1				-1.4	-1.07	PREDICTED: uncharacterized protein LOC100159027 [*Acyrthosiphon pisum*]	2.00E-59
TR15199|c2_g1_i4	0.78		2.3	-1.84			0.83	conserved hypothetical protein [*Culex quinquefasciatus*]	4.00E-86
TR29999|c0_g1_i1	1.95	-1.39			1.25	-1.31		mannose-binding protein [*Portunus pelagicus*]	3.00E-56

The transcripts that were differentially expressed in multiple time points following WSSV infection were also interpreted along functional categories, e.g. through overrepresented GO terms. Enriched GO terms amongst transcripts that were DE in at least 3 time points are summarized in Table 6. Structural constituent of peritropic membrane (GO:0016490), is probably related to the peritrophin-like protein that plays a part in pathogen-associated molecular pattern recognition, as discussed earlier. However, the majority of enriched terms were related to RNAi, with miRNAs in particular, again confirming the importance of this pathway in the response of *C. maenas* to WSSV. The miRNA sequencing data that will be discussed below provides more detail about exactly which miRNAs were up/down regulated during this exposure experiment. Combining results from Tables 5 and 6 indicate a two-step longer term expression response: firstly, perturbations in the expression of pathogen recognizing proteins and secondly higher expression of components involved in the miRNA-mediated immune response.

Table 6 Overrepresented GO terms in transcripts DE at least 3 time points, FDR <= 0.05.

Gene Ontology - ID	Gene Ontology Term	FDR	P-Value
GO:0016490	structural constituent of peritrophic membrane	1.18E-03	9.35E-08
GO:0070922	small RNA loading onto RISC	1.11E-02	8.73E-06
GO:0090624	endoribonuclease activity, cleaving miRNA-paired mRNA	1.11E-02	8.73E-06
GO:0090625	mRNA cleavage involved in gene silencing by siRNA	1.11E-02	8.73E-06
GO:0035197	siRNA binding	1.11E-02	8.73E-06
GO:0035280	miRNA loading onto RISC involved in gene silencing by miRNA	1.11E-02	8.73E-06
GO:0035087	siRNA loading onto RISC involved in RNA interference	1.11E-02	8.73E-06
GO:0070883	pre-miRNA binding	1.11E-02	8.73E-06
GO:0070551	endoribonuclease activity, cleaving siRNA-paired mRNA	1.11E-02	8.73E-06
GO:0070578	RISC-loading complex	1.11E-02	8.73E-06
GO:0045974	regulation of translation, ncRNA-mediated	1.47E-02	1.74E-05
GO:0040033	negative regulation of translation, ncRNA-mediated	1.47E-02	1.74E-05
GO:1902555	endoribonuclease complex	1.47E-02	1.74E-05
GO:0035278	negative regulation of translation involved in gene silencing by miRNA	1.47E-02	1.74E-05
GO:0035068	micro-ribonucleoprotein complex	1.47E-02	1.74E-05
GO:0005845	mRNA cap binding complex	2.30E-02	2.90E-05
GO:0016893	endonuclease activity, active with either ribo- or deoxyribonucleic acids and producing 5'-phosphomonoesters	2.51E-02	4.11E-05
GO:0035198	miRNA binding	2.51E-02	4.35E-05
GO:0098795	mRNA cleavage involved in gene silencing	2.51E-02	4.35E-05

GO:0035279	mRNA cleavage involved in gene silencing by miRNA	2.51E-02	4.35E-05
GO:0034518	RNA cap binding complex	2.51E-02	4.35E-05
GO:0045947	negative regulation of translational initiation	2.51E-02	4.35E-05
GO:0000340	RNA 7-methylguanosine cap binding	3.35E-02	6.09E-05
GO:0031332	RNAi effector complex	3.95E-02	8.11E-05
GO:0016442	RISC complex	3.95E-02	8.11E-05
GO:0031054	pre-miRNA processing	3.95E-02	8.11E-05
GO:0008061	chitin binding	4.71E-02	1.04E-04
GO:0000993	RNA polymerase II core binding	4.71E-02	1.04E-04

Differential expression in individual time points

As mentioned, a different perception of the biological response of *C. maenas* to WSSV exposure was achieved by examining changes in transcript expression at individual time points. Table 4 and Supplementary File S2 indicate that hundreds of transcripts were DE at each time point, making it intractable to discuss each individual transcript's function and relation to WSSV exposure. Instead, interpretation of the transcriptional response was done based on processes and pathways that were over-represented within the lists of DE genes for each time point. Overrepresentation of GO terms associated with DE transcripts was not identified for every time point: data obtained at 6h , 48h and 336h post-infection did not show significant enrichment at the chosen cut off. The top 6 GO terms of the biological process (P), cellular component (C) and molecular function (F) branches for each time point are shown in Table 7, and a full table can be found in Supplementary file S3.

Table 7 Enriched GO terms in DE transcripts across time points

Time & GO branch	GO-ID	Term	FDR	P-Value
6h)				
12h) P	GO:0006418	tRNA aminoacylation for protein translation	2.84E-06	3.29E-10
P	GO:0043039	tRNA aminoacylation	2.84E-06	1.08E-09
P	GO:0043038	amino acid activation	2.84E-06	1.08E-09
P	GO:0015031	protein transport	9.91E-04	1.36E-06
P	GO:0006399	tRNA metabolic process	1.67E-03	2.90E-06
P	GO:0006886	intracellular protein transport	2.79E-03	5.37E-06
C	GO:0030662	coated vesicle membrane	7.03E-05	4.52E-08
C	GO:0030120	vesicle coat	1.10E-04	9.66E-08
C	GO:0048475	coated membrane	1.10E-04	1.00E-07
C	GO:0030117	membrane coat	1.10E-04	1.00E-07
C	GO:0030135	coated vesicle	2.51E-04	2.54E-07
C	GO:0030660	Golgi-associated vesicle membrane	5.66E-04	6.23E-07
F	GO:0004812	aminoacyl-tRNA ligase activity	2.84E-06	7.36E-10
F	GO:0016876	ligase activity, forming aminoacyl-tRNA and related compounds	2.84E-06	1.56E-09
F	GO:0016875	ligase activity, forming carbon-oxygen bonds	2.84E-06	1.56E-09
F	GO:0016490	structural constituent of peritrophic membrane	2.71E-02	8.72E-05
F	GO:0090624	endoribonuclease activity, cleaving miRNA-paired mRNA	3.13E-02	1.26E-04
F	GO:0070883	pre-miRNA binding	3.13E-02	1.26E-04
24h) P	GO:0044093	positive regulation of molecular function	2.60E-02	9.70E-06
F	GO:0016787	hydrolase activity	2.60E-02	6.06E-06
F	GO:0016788	hydrolase activity, acting on ester bonds	2.60E-02	7.33E-06
F	GO:0042578	phosphoric ester hydrolase activity	2.60E-02	9.70E-06
F	GO:0008081	phosphoric diester hydrolase activity	2.60E-02	1.19E-05
48h)				
168h) P	GO:0048016	inositol phosphate-mediated signalling	1.71E-03	3.15E-07
P	GO:0006414	translational elongation	1.07E-02	6.85E-06
P	GO:0098662	inorganic cation transmembrane transport	2.02E-02	1.48E-05
P	GO:0098660	inorganic ion transmembrane transport	2.86E-02	2.62E-05
P	GO:0098655	cation transmembrane transport	2.98E-02	3.00E-05
P	GO:0046034	ATP metabolic process	3.16E-02	3.48E-05
F	GO:0005220	inositol 1,4,5-trisphosphate-sensitive calcium-release channel activity	1.71E-03	3.15E-07
F	GO:0015278	calcium-release channel activity	2.25E-03	6.20E-07
F	GO:0005217	intracellular ligand-gated ion channel activity	3.89E-03	1.89E-06
F	GO:0022834	ligand-gated channel activity	3.89E-03	2.14E-06
F	GO:0015276	ligand-gated ion channel activity	3.89E-03	2.14E-06
F	GO:0003746	translation elongation factor activity	2.39E-02	1.98E-05

336h)				
672h) P	GO:0055114	oxidation-reduction process	2.74E-07	1.01E-10
P	GO:0044710	single-organism metabolic process	1.93E-04	2.12E-07
P	GO:0006030	chitin metabolic process	2.71E-04	3.24E-07
P	GO:1901071	glucosamine-containing compound metabolic process	9.26E-04	1.60E-06
P	GO:0006040	amino sugar metabolic process	2.75E-03	6.09E-06
P	GO:0009205	purine ribonucleoside triphosphate metabolic process	2.75E-03	6.49E-06
C	GO:0005739	Mitochondrion	2.95E-09	2.71E-13
C	GO:0044429	mitochondrial part	1.05E-07	2.88E-11
C	GO:0031966	mitochondrial membrane	5.57E-05	3.58E-08
C	GO:0098800	inner mitochondrial membrane protein complex	7.23E-05	5.31E-08
C	GO:0098798	mitochondrial protein complex	8.66E-05	7.15E-08
C	GO:0044455	mitochondrial membrane part	1.15E-04	1.05E-07
F	GO:0016491	oxidoreductase activity	8.17E-09	1.50E-12
F	GO:0008061	chitin binding	4.49E-07	2.06E-10
F	GO:0003954	NADH dehydrogenase activity	5.21E-05	2.87E-08
F	GO:0050136	NADH dehydrogenase (quinone) activity	1.47E-03	2.82E-06
F	GO:0008137	NADH dehydrogenase (ubiquinone) activity	1.47E-03	2.82E-06
F	GO:0016651	oxidoreductase activity, acting on NAD(P)H	1.67E-03	3.53E-06

In the analysis of DE transcripts across multiple time points it was observed that there were significant expression changes in pathogen recognizing proteins, the initial actors of the immune response. However at an individual time point level responses of the immune system are not illustrated through the enrichment of GO terms amongst DE transcripts (see Table 7). However given the lack of transcript annotation and subsequent linking of GO terms this is not necessarily evidence that an immune response does not occur. Especially in the later time points, where some WSSV injected samples showed evidence of replicating WSSV, one would expect changes in the expression of many components of the *C. maenas* immune system. Indeed, there have been numerous studies that show differential expression of immune system components upon WSSV challenge in susceptible species (reviewed by Shekhar and Ponniah 2014 [65]). The lack of immune response could be genuine and stem from the fact that *C. maenas* is relatively resistant to WSSV as compared to other aquatic crustaceans. In most samples the virus did not manage to replicate which removes the necessity of a significant transcription change of immune system components. However, there are additional complexities to this issue. It has been shown that WSSV hijacks immune system transcription factors to promote expression of its own genes. For example, *P. monodon* STAT (*Pm*STAT) can enhance expression of WSSV *ie1* [16, 66], and other examples are described in

Verbruggen *et al.* 2016 [13]. Therefore, oxymoronically, not responding to viral presence could be beneficial to the host, complicating establishment of successful replicating. In our experiment we only identified a significantly upregulated transcript with similarity to *Pm*STAT, TR51123|c13_g1_i2, at 672 hours post injection (See Supplementary File S2).

6 hours post injection

The fact that only five individuals developed replicating WSSV suggested that *C. maenas* might have methods of disposing the viral threat prior to infection establishment. In the time covered by this exposure experiment such mechanisms would probably occur at the earlier time points. The GO enrichment of DE transcripts in Table 7 does not point toward a clear direction for the 6 hour time point as no enrichment was identified. Nevertheless there were individual DE transcripts which could be of importance. While the whole list of annotated DE transcript cannot be discussed, several well annotated can be highlighted. There are two antimicrobial peptides: Lysozyme (TR24811|c2_g1_i1, down regulated) and crustin-1 (TR7697|c0_g1_i1, up-regulated), one of the most important antimicrobial peptides found in decapod crustaceans [67-69]. Both of these antimicrobial peptides are known for their activity against bacterial pathogens. However, Antony et al. 2011 showed that WSSV exposure caused significant upregulation of crustin-3 expression in *P. monodon* [70]. Crustins thus might act against viral pathogens like WSSV as well, given their expression changes in both *C. maenas* and *P. monodon*. The activity of the RNAi system upon WSSV exposure was mentioned in the previous section. At 6 hours post injection there is upregulation of both Argonaute 2 and Dicer-1 (TR5657|c0_g1_i1), an enzyme that produces miRNAs from their precursor molecules. This will be further discussed below, together with the analysis of the miRNA sequencing data.

12 hours post injection - endocytosis

The 12 hour time point showed differential expression of transcripts enriched for Gene Ontology terms with two main mechanisms: vesicle trafficking (e.g. GO:0030135 – coated vesicle) and RNAi (GO:0090624 - endoribonuclease activity, cleaving miRNA-paired mRNA) (Table 7). Vesicle trafficking, as part of the endocytosis pathway, is often employed by viruses to penetrate the host cells [71]. Through binding to cell surface receptors, subsequent uptake by endocytic vesicles, transport to the perinuclear area and subsequent escape from the endosomes, viruses can efficiently approach the host nucleus. Indeed it has been shown that

WSSV also employs endocytosis for viral entry [13]. WSSV enters through Clathrin-mediated endocytosis [15] and interacts with Rab7, an important regulator protein in endocytosis, through VP28 [72]. However the exact details of the WSSV entry process along with its accompanying host-pathogen interactions remain to be investigated [13]. Because of its importance to the early stage of infection, any disruption of this cellular system can significantly reduce the probability of successful infection. Indeed it has been shown for many viruses that disruption of endocytosis is detrimental to viral infection. Disruption can take place at various stages of endocytosis: from early endosome regulators to late regulators. Many examples of such cases are available in literature. For example: the infection of Infectious bursal disease virus is dependent on Rab5, an early endosome regulator [73]. Depletion of factors important in the late stages of endocytosis, e.g. Rab7, had an inhibiting effect on poxvirus infection [74]. Internalization of influenza virus is significantly affected by disruption of the COPI complex, a protein that coats vesicles that transport proteins between the Golgi apparatus and endoplasmic reticulum [75]. Similarly dsRNA-mediated silencing of COPI inhibited the early stage of vesicular stomatitis virus replication [76]. To provide an overview of the effects of WSSV exposure on the endocytosis system in *C. maenas*, expression changes were mapped onto the KEGG canonical pathway (Figure 7). Figure 7 shows that most of the components are upregulated in response to WSSV exposure indicating increased activity of this pathway, however there are several important regulators that are downregulated: Hsc70, Rab11 and Rab7. Decreased expression of these regulators could impact the flow of the vesicle transport system. Firstly, Hsc70 plays multiple roles in the formation and processing of Clathrin-coated vesicles, which are the vesicles through which WSSV virions enter the endocytosis pathway [77]. Hsc70 is required during coated-pit invagination/constriction and additionally in chaperoning Clathrin after the coat is dissociated from the vesicles [78]. Downregulation of Hsc70 would thus result in a reduction in the formation and processing of Clathrin-coated vesicles, impeding viral entry. Exemplary is the dependency of Japanese encephalitis virus entry in *Drosphila* cells on Hsc70 [79]. Rab11 is an important regulator for endocytic recycling, and its presence on a cellular compartment is defining for endocytic recycling compartments (ERC) [80]. The ERC can be employed by viruses during viral entry, e.g. the human Vaccinia virus and Kaposi's sarcoma-associated herpesvirus localize in the ERC upon which viral fusion and core uncoating might occur [80]. It is not clear whether WSSV can employ a similar route during viral entry. The last regulator that is downregulated upon WSSV exposure was Rab7. Rab7 is a well-studied regulator that is mainly involved in endosome maturation and the degradative pathway in the endocytosis system [80]. Important in regards to WSSV infection is that Sritunyalucksana *et al.* 2006 showed that the major WSSV structural protein VP28 binds to Rab7, an interaction that appears to be necessary for successful infection [72, 81, 82]. A reduction of *C maenas* Rab7 could thus significantly impact the viral entry process in this host. Taken together, the downregulation of these three regulators appear to impact the endocytosis pathway in such a way as to reduce the ability of WSSV to enter the host cell. Figure 8 summarizes the effect of downregulation of Hsc70, Rab11 and Rab7 on endocytosis. Transport of WSSV virions into early endosomes is reduced due to decreased expression of Hsc70. In the early endosome cargo is either sorted to the recycling, or degradative pathways. At 12 hours post injection decreased expression of Rab11 and Rab7 reduce transport into either the degradative or slow recycling pathways. WSSV particles that reach the early endosomes are thus likely to either accumulate in those vesicles or be transported back to the cellular exterior through the fast-recycling pathway. Given that WSSV was not able to establish a successful infection until a week after injection it can be hypothesized that passing through *C. maenas* endosomes negatively affected the virus' ability to infect *C. maenas* cells. Waikhom *et al.* 2006 [83] showed that prior passing of WSSV through an organism can affect viral pathogenicity in subsequent infections. But whether a similar effect can be achieved by only passing through *C. maenas* early endosomes, and if so by what mechanism, remains to be clarified.

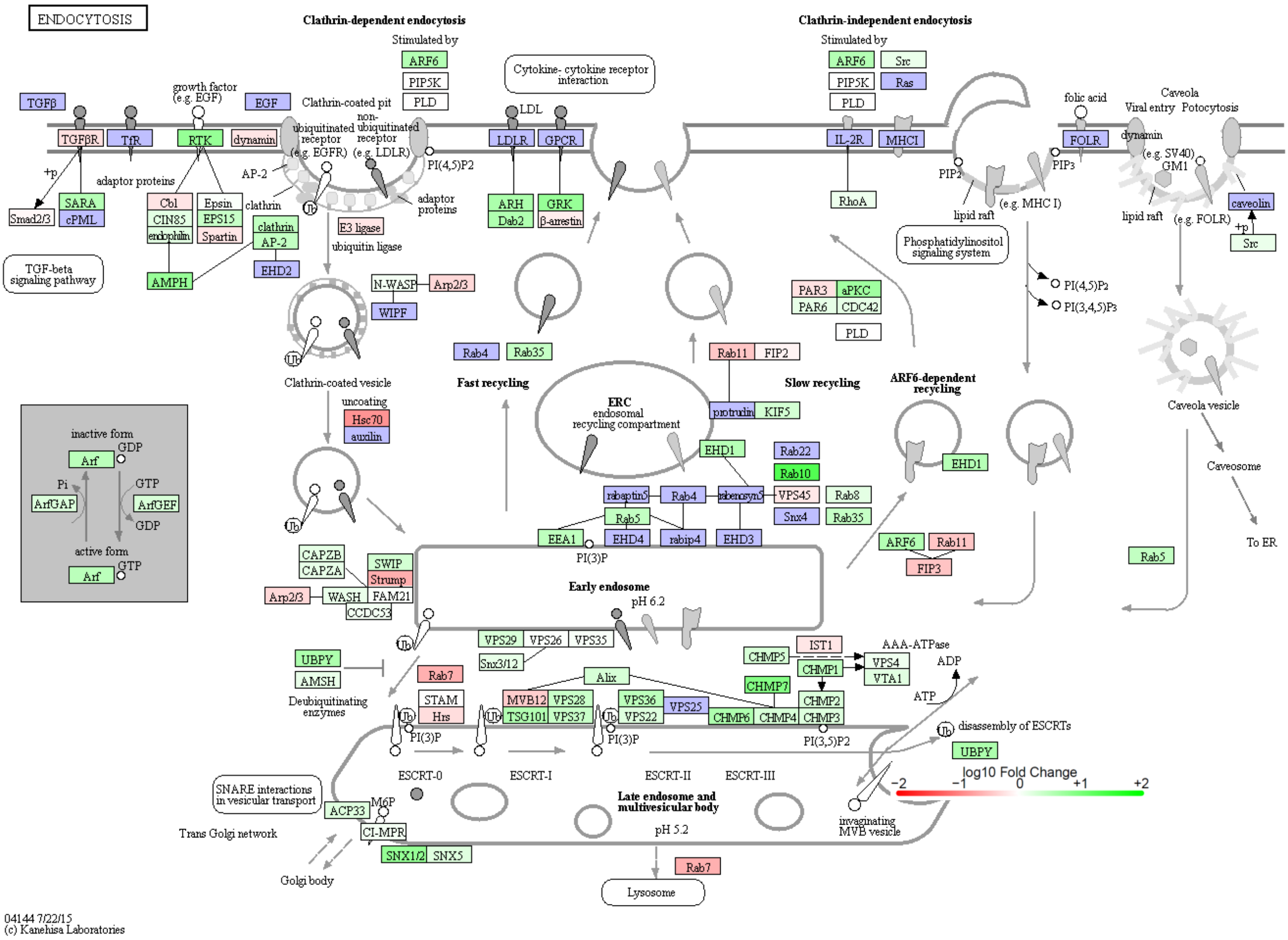


Figure 7 KEGG endocytosis pathway 12h post injection. KEGG endocytosis pathway for *C. maenas* at the 12h time point. Components are coloured according to log10 fold change as determined by EBSeq comparing WSSV exposed to control. Components not identified in the *C. maenas* transcriptome have a purple background colour. Most components have a higher expression in WSSV exposed samples, but a number of transcripts were down-regulated Rab7, Rab11 and Hsc70.

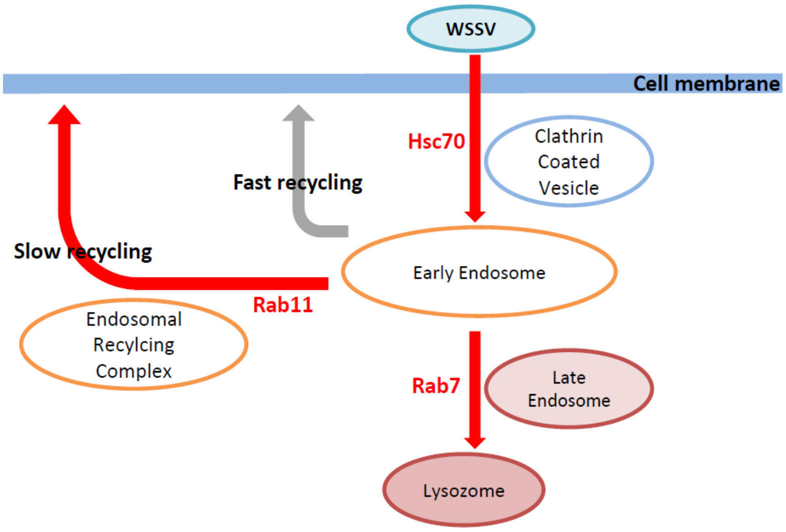


Figure 8 Overview of *C. maenas* endocytosis at 12h post injection. The downregulation of Hsc70, Rab11 and Rab7 will likely change the flows of the endocytosis pathway. Decreased expression of Hsc70 may result in a lower flow into the early endosome through Clathrin-mediated endocytosis. The flow from the early endosomes into the degrative pathway, which ends in the lysozome, may be reduced through lower expression of Rab7. Similarly, the flow from the early endosome into the ERC/slow recylcing pathway may be limited by Rab11 availability. Note: endocytosis was not explicitly measured in this experiments, conclusions are based purely on transcript expression data and probably need to be confirmed.

12 hours post injection - RNAi

The second group of enriched GO terms at the 12h time point relate to RNAi (Table 7). As stated earlier, miRNAs are an important component of the crustacean immune system [13]. Expression changes of components of the RNAi pathway at the 12h time point are listed in Table 8. Most of the components of this pathway show increased expression after viral exposure. There are components of the RNAi pathway that are represented by multiple *C. maenas* transcripts. For example: Dicer-1 is represented by transcripts TR42801|c3_g1_i4, TR42801|c3_g1_i5, TR4912|c0_g1_i1 and TR4912|c3_g1_i1. Of these TR42801|c3_g1_i5 is differentially expressed. Overall the TAR RNA-binding protein/R2D2, Dicer-1 and Argonaute 1/2 are all differentially expressed at this time point, indicating an overall increased activity of the RNAi pathway.

Table 8 *C. maenas* RNAi pathway expression changes at 12h post WSSV injection

C. maenas **transcript**	**RNAi pathway component homolog**	**NCBI gi**	**tBLASTn evalue**	**EBSeq PPDE**	**log10_FC**
TR28381|c10_g2_i1	R2D2 [*Bemisia tabaci*]	gi|619831236	2.00E-92	0.098	0.204
TR28381|c10_g2_i1	TAR RNA-binding protein 1 [*Penaeus monodon*]	gi|444174849	0	0.098	0.204
TR28381|c10_g2_i2	R2D2 [*Bemisia tabaci*]	gi|619831236	1.00E-79	1	1.969
TR28381|c10_g2_i2	TAR RNA-binding protein 1 [*Penaeus monodon*]	gi|444174849	0	1	1.969
TR28381|c10_g3_i1	R2D2 [*Bemisia tabaci*]	gi|619831236	8.00E-45	0.046	0.068
TR28381|c10_g3_i1	TAR RNA-binding protein 1 [*Penaeus monodon*]	gi|444174849	1.00E-114	0.046	0.068
TR29298|c7_g1_i1	drosha [*Marsupenaeus japonicus*]	gi|396941645	0	0.086	-0.247
TR33054|c1_g1_i1	argonaute 1 [*Marsupenaeus japonicus*]	gi|283827858	0	0.383	0.19
TR33054|c1_g1_i1	argonaute2 [*Penaeus monodon*]	gi|563729913	1.00E-108	0.383	0.19
TR33054|c3_g1_i1	argonaute 1 [*Marsupenaeus japonicus*]	gi|283827858	0	0.053	0.111
TR33054|c3_g1_i1	argonaute2 [*Penaeus monodon*]	gi|563729913	4.00E-71	0.053	0.111
TR42801|c3_g1_i4	Dicer-1 [*Drosophila melanogaster*]	gi|17738129	2.00E-43	0.073	-0.188
TR42801|c3_g1_i5	Dicer-1 [*Drosophila melanogaster*]	gi|17738129	4.00E-66	1	0.512
TR48022|c0_g1_i1	argonaute 1 [*Marsupenaeus japonicus*]	gi|283827858	1.00E-69	0.998	0.56
TR48022|c0_g1_i1	argonaute2 [*Penaeus monodon*]	gi|563729913	9.00E-83	0.998	0.56
TR48022|c0_g1_i2	argonaute 1 isoform B [*Penaeus monodon*]	gi|110294438	6.00E-74	1	0.48
TR48022|c0_g1_i2	argonaute2 [*Penaeus monodon*]	gi|563729913	4.00E-92	1	0.48
TR48022|c1_g1_i2	argonaute2 [*Penaeus monodon*]	gi|563729913	0	1	0.604
TR48022|c1_g1_i2	argonaute 1 isoform B [*Penaeus monodon*]	gi|110294438	1.00E-127	1	0.604
TR4912|c0_g1_i1	Dicer-1 [*Drosophila melanogaster*]	gi|17738129	2.00E-177	0.002	0.057
TR4912|c3_g1_i1	Dicer-1 [*Drosophila melanogaster*]	gi|17738129	7.00E-159	0	-0.151

After 24 hours the GO enrichment of the endocytosis and RNAi pathways has disappeared and been replaced with terms related to enzymatic activity, mainly hydrolase activity. These terms are more difficult to relate to the process of viral infection. Again in this time point many annotations are either derived from hypothetical or predicted sequences. In relation to the immune system it was found that TR40119|c12_g1_i8, a homologue to the *Eriocheir sinensis* NF-κB transcription factor Relish, is significantly downregulated (log10 FC = -1.49, Supplementary File S2). Huang *et al.* 2010 [84] showed that *L. vannamei* Relish can bind to the WSSV *ie1* promoter and thereby regulate expression of viral genes. Downregulation of this particular transcription factor should thus decrease the ability of WSSV to replicate. At 48 hours post WSSV injection there were no enriched GO terms and again the majority of DE transcripts were annotated to hypothetical or predicted, an indication that aquatic crustacean species are underrepresented in studies. In this time point again several transcripts related to the immune system were significantly suppressed including TR19437|c1_g1_i1, similar to *L. vannamei* Toll3 and the serine protease TR23315|c5_g1_i2. While serine proteases have many functions in invertebrates, one role they play is to cleave proteins in the proteolytic cascade of the melanization response [85]. Over all the early time points, it appears that the 12 hour time point corresponds to the timing where most significant changes in transcription occur as part of the *C. maenas* response to WSSV exposure. Whilst the other time points show differential expression of a handful of immune system related proteins, at 12 hours a clearer over-representation of RNAi and endocytosis related genes was identified. It can be hypothesized based on the response at 12 hours that cells either accumulate WSSV in early endosomes or recycle virions to the cellular surface, which could render the virions incapable of further damaging the host.

7 – 28 days post injection

During the later time points (7 – 28 days post injection) signs of replicating WSSV were detected in several crabs through mapping of sequencing reads to the WSSV-CN genome. GO enrichment analysis also revealed the presence of terms related to the virus. Enrichment calculations based on the complete transcript set, thus unfiltered for metazoan sequences only, showed enrichment of viral components: GO:0036338 viral membrane FDR 6.93E-07, GO: 0019031 viral envelope FDR 6.93E-07, GO:0044423 virion part FDR 3.70E-03 and GO:0019012 virion FDR 3.92E-03 at the 7 day time point (Supplementary File S3). Similar results were found for the samples at 28 days post injection, however not for the 14 day time point. The latter result is probably due to the fact that only a single individual sampled at that time point had developed replicating WSSV. Other than virus-related GO terms, there appears to be up-regulation of transcripts that function within transport and metabolism processes at the 7 day time point (Table 7). The latest time point, 28 days post injection, shows clear enrichment for terms related to energy generation (e.g. mitochondrion and oxidation-reduction process). When a virus is replicating the demands for energy are high due to increased protein production, and often viruses influence host metabolism to ensure enough energy is available. Indeed, WSSV has been shown to be able to induce the Warburg effect in host cells, increasing energy generation via anaerobic signalling pathways [86]. Influence of WSSV on host process is more clearly demonstrated by comparing only WSSV replicating samples to control samples in the later time points. Differential transcript expression was performed in a similar manner, but omitting the WSSV injected, but infection free samples. The number of differentially expressed transcripts identified at each time point is listed in Table 9. A selection of enriched GO terms for WSSV replicating samples are shown in Table 10. At the 168h time point, terms related to the presence of the virus are once again over-represented but in this comparison modulation of host process by the virus are also visible. The 336h time point did not yield any enriched GO terms, probably caused by there being only a single sample in the WSSV replicating group which handicaps statistical tests (represented by the lower number of DE transcripts when compared to the other WSSV replicating time points). The enriched GO terms in the latest time point reflect the ones identified in the previous comparison: terms related to energy generation and virus modulation of host processes.

There is only a limited number of DE transcripts related to the immune system, and Supplementary File S2 shows activity of processes involving biosynthesis of nitric oxide (NO). For example, *C. maenas* nitric oxide synthase (TR24850|c3_g1_i1) is upregulated at 672h. Within the immune response, inducible NO synthase can rapidly generate NO free radicals that are damaging to invading pathogens. Current knowledge of the NO response in crustaceans is limited and is believed to primarily target bacterial, fungal and parasitic pathogens instead of viruses [87]. Increased biosynthesis of NO under WSSV infection could thus be a first indicator that the invertebrate immune system also employs NO radicals against viral pathogens. However NO plays a variety of roles within organisms and thus whether the increased expression of *C. maenas* nitric oxide synthase is related to the immune response or another process required further investigation.

Table 9 Number of differentially expressed transcripts in samples with replicating WSSV

Time (hpi) WSSV replicating only	Number of DE transcripts (#)
168	3647
336	985
672	5444

Table 10 Representative enriched Gene Ontology terms in WSSV replicating samples

Time	GO-ID	Term	FDR	P-Value
168h (WSSV)	GO:0036338	viral membrane	2.22E-05	8.78E-09
	GO:0019031	viral envelope	2.22E-05	8.78E-09
	GO:0004517	nitric-oxide synthase activity	3.47E-03	7.85E-06
	GO:0006210	thymine catabolic process	3.47E-03	7.95E-06
	GO:0006208	pyrimidine nucleobase catabolic process	3.47E-03	7.95E-06
	GO:1901565	organonitrogen compound catabolic process	9.51E-03	3.44E-05
	GO:0019054	modulation by virus of host process	9.51E-03	3.76E-05
	GO:0019048	modulation by virus of host morphology or physiology	4.13E-02	2.41E-04
	GO:0046596	regulation of viral entry into host cell	4.13E-02	2.41E-04
336h (WSSV)				
672h (WSSV)	GO:0003954	NADH dehydrogenase activity	5.65E-07	7.13E-10
	GO:0005746	mitochondrial respiratory chain	9.40E-07	1.26E-09
	GO:0098803	respiratory chain complex	2.34E-06	3.33E-09
	GO:0005747	mitochondrial respiratory chain complex I	3.62E-06	6.01E-09
	GO:0031967	organelle envelope	1.27E-07	1.30E-10
	GO:0016651	oxidoreductase activity, acting on NAD(P)H	1.47E-07	1.62E-10

	GO:0015980	energy derivation by oxidation of organic compounds	1.24E-03	3.43E-06
	GO:0036338	viral membrane	7.24E-05	1.60E-07
	GO:0019031	viral envelope	7.24E-05	1.60E-07
	GO:0019054	modulation by virus of host process	2.34E-02	1.95E-04
	GO:0039648	modulation by virus of host protein ubiquitination	9.43E-03	4.17E-05

6.4.5 Analysis of miRNA expression

A consistent result of the RNA-sequencing data was the increased expression of genes involved in the production and activity of miRNAs. The importance of RNAi to *C. maenas* response to WSSV warrants investigation of differential expression of miRNAs upon exposure to this pathogen. Therefore, small RNAs were sequenced on an Illumina HiSeq-2500 platform. Known and novel miRNAs were identified based on miRbase miRNAs from related species (*M. japonicus, T. castaneum* and *D. pulex*)[48-51, 88]. Using mirdeep2 and subsequent clustering with CD-HIT-EST, a total of 322 host miRNAs were identified. Of these, 207 miRNAs were novel and subsequently named ‘cma_pmiR-XXX’, the p standing for predicted. Filtering for miRNAs with at least 10 CPM across 4 samples within a time point retained 77 miRNAs for testing, of which 24 were novel *C. maenas* miRNAs. A list of miRNAs and their expression values can be found in Supplementary File S4 and S5. For every time point differentially-expressed host miRNAs were identified (Table 11). Interestingly, the novel cma_pmiR-12 miRNA was differentially expressed in six of the seven time points. The mature sequence of cma_pmiR-12 is 5’ ATCCTGGTCACGGCACCA 3’ was identified within *C. maenas* genome scaffold_23728 (Chapter 4), and its pre-miRNA structure is shown in Figure 9. A heatmap of normalized expression values shows that this miRNA has a lower expression in WSSV-injected samples across all time points (Figure 10). Since this is a novel miRNA it would be important to determine the potential targets of cma_pmiR_12. Prediction of miRNA targets was performed by the miRanda and MicroTar software packages which resulted in 3691 target predictions using miRanda and 5592 target predictions using MicroTar with a free energy cut-off of -10 kcal/mol (Table 12). Investigating potential roles for the 321 shared targets of both prediction methods through GO-term enrichment as derived by BLAST2GO resulted in only a single enriched term: LIM domain binding (GO:0030274) with an FDR of 0.039. LIM domains are involved in cytoskeleton organization [89]. However it should be noted that only four of the 321 targets are annotated with GO:0030274 and therefore a large fraction of the targets is not represented by this ontology term. Investigating only the MicroTar results for GO term enrichment did not yield any significant results. Targets identified by miRanda at a threshold of -10 kcal/mol were enriched for 48 GO terms (Supplementary File S6). The enriched GO term with the largest group of associated transcripts was GO:0012505, endomembrane system, with FDR 0.0305. This potentially relates to the hypothesis generated in the previous section, where changes in the endocytosis pathway were identified as a major characteristic of *C. maenas* apparent resistance to WSSV infection. However, as illustrated by the small amount of overlap between the *in silico* miRNA-target prediction methods, the target predictions can be capricious and should be confirmed through further experimentation e.g. proteomics studies involving overexpression/knockdown of cma_pmiR-12. Extending such overexpression/knockdown target studies with WSSV exposure should reveal how strongly this miRNA is correlated with the resistance mechanisms operating in *C. maenas*.

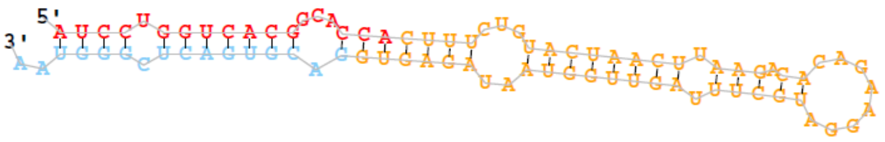
Table 11 differentially expressed *C. maenas* miRNAs.

Time	**Total expressed**	**Differentially expressed**	**logFC (Control v WSSV)**	**FDR**
6h	71	cma_pmiR-11	-8.16111282	6.35E-12
		cma_pmiR-12	-2.698356631	0.000827715
		tca-miR-2788-3p	7.693217558	0.000827715
12h	63	cma_pmiR-12	-3.812747811	1.46E-14
24h	66	cma_pmiR-12	-3.070854024	1.46E-14
48h	66	cma_pmiR-12	-2.874475441	0.000710241
		tca-miR-133-3p	-5.683121602	0.003192561
168h	70	cma_pmiR-12	-4.373954592	1.97E-07
		tca-miR-980-3p	2.164258414	0.029730543
336h	72	cma_pmiR-12	-4.142390557	0.000710241
		cma_pmiR-107	-3.987799503	0.003192561
672h	64	tca-miR-2788-3p	-9.147749945	8.75E-05
		tca-miR-263b-5p	-2.377694806	0.002527052
		tca-miR-750-3p	-3.466759432	0.002527052
		tca-miR-3884-3p	1.5331964	0.050555679

Table 12 cma_pmiR-12 target prediction

Prediction software	Number of targets (< -20 kcal/mol)	Overlap (< -20 kcal/mol)	Number of targets (< -10 kcal/mol)	Overlap (< -10 kcal/mol)
miRanda	1020	0	3691	321
MicroTar	25		5592	

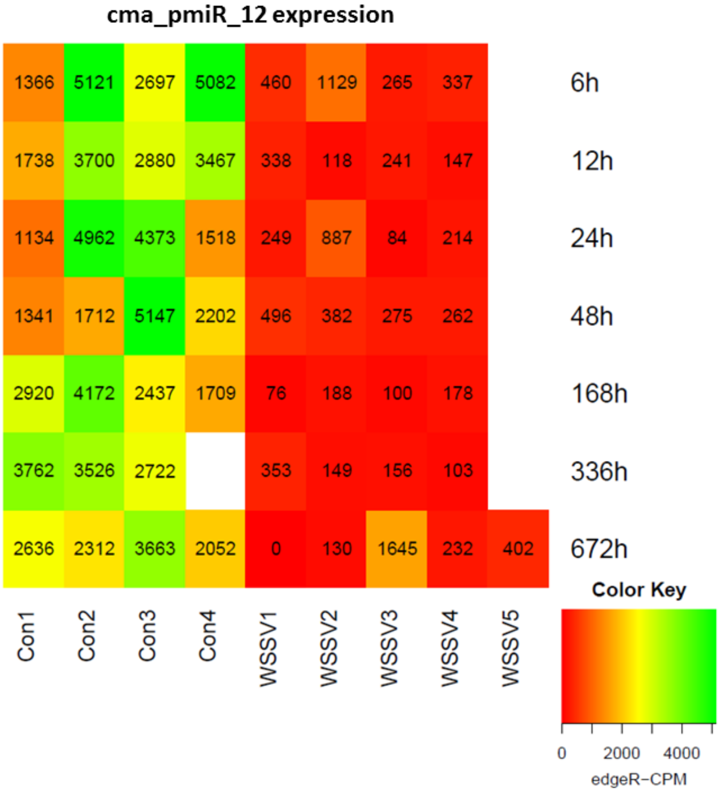
cma_pmiR-12 structure



Location: *C. maenas* scaffold_23728_2517

```
                        Mature                                                              Star
5'- cuaauuaacccaggcucguaauccuggucacggcaccacuuucuguacuaacuuaagacacagaaggaugcuuuaguugguaauagaguggacgugacucggguaaug
    ................(((.((((.(((((((...(((((((...((((((((.(((.((........))))).))))))))..))))))).))))))).)))).)))
    ...................Gauccuggucacggcacca......................................................................
    ....................auccuggucacggcacca......................................................................
```

Figure 9 Structure of cma_pmiR-12. The structure of cma_pmiR-12 as predicted by mirdeep2 and the sequence in *C. maenas* genome. The sequence is presend on *C. maenas* scaffold 23728.



	Con1	Con2	Con3	Con4	WSSV1	WSSV2	WSSV3	WSSV4	WSSV5
6h	1366	5121	2697	5082	460	1129	265	337	
12h	1738	3700	2880	3467	338	118	241	147	
24h	1134	4962	4373	1518	249	887	84	214	
48h	1341	1712	5147	2202	496	382	275	262	
168h	2920	4172	2437	1709	76	188	100	178	
336h	3762	3526	2722		353	149	156	103	
672h	2636	2312	3663	2052	0	130	1645	232	402

Figure 10 Heatmap of normalized expression of cma_pmiR_12 across time points, for individual samples. Expression values for cma_pmiR_12 were obtained using edgeR normalization. Rows depict time points, columns indicate the control/WSSV replicates for that each time point. Each box represents normalised expression values for a single *C. maenas* individual. Across all samples and time points the expression of cma_pmiR_12 is lower than those in controls.

The mirdeep2 mapper function mapped small RNA sequencing data of five samples to the WSSV-CN genome. These were the same samples found to contain replicating WSSV in the RNA-sequencing data: ws_36 (0.03 % mapped), ws_37 (0.01 % mapped), ws_48 (0.044 % mapped), ws_50 (0.045 % mapped) and ws_51 (0.009 % mapped). Using the same method as for *C. maenas*, 15 WSSV miRNAs were identified in these samples. Of these 11 were described by Huang *et al.* 2014 [21] and four were classified as novel (named 'WSSV_pmiR-[1-4]'). The sequences of the WSSV miRNAs are listed in Supplementary File S7. The study of WSSV miRNA expression by Huang *et al*. 2014 described that viral miRNA expression was tissue specific, with 18 miRNAs detected in the gill of *P. monodon*. This study confirmed the expression of 4 of these 18 miRNAs, showing there is not only a difference between tissues but also between species. The confirmed WSSV-miRNAs included WSSV-miR-N24 which was shown to target shrimp caspase 8, thereby exerting influence on the host apoptosis pathway, enabling WSSV to maintain a host environment conductive to replication [21].

Table 13 WSSV miRNAs detected in WSSV replicating samples

Time (hours post infection)	**Sample ID**	**Detected WSSV miRNA (* found in gill tissue by Huang *et al*. 2014)**	**Detected novel WSSV miRNA**
168	ws_36	WSSV-miR-N24* WSSV-miR-N38* WSSV-miR-N41	
168	ws_37	WSSV-miR-N13 WSSV-miR-N25 WSSV-miR-N29* WSSV-miR-N39 WSSV-miR-N43	WSSV-pmiR-1
336	ws_48	WSSV-miR-N47	WSSV-pmiR-2
672	ws_50	WSSV-miR-N26 WSSV-miR-N29*	WSSV-pmiR-3
672	ws_51	WSSV-miR-N41	WSSV-pmiR-4

6.5 Conclusion

While many aquatic crustaceans are very highly susceptible to WSSV infection, *C. maenas* was shown to be relatively resistant to this virus. In order to investigate the basis of this apparent resistance we performed RNA-sequencing and small RNA-sequencing experiments on *C. maenas* gill tissue after injection with either WSSV or saline solution. Our experiment confirmed that *C. maenas* is indeed recalcitrant to

WSSV infection as in our study only five individuals developed replicating virus. Furthermore, viral replication occurred only after a relatively long period of time compared to WSSV virus infections in other aquatic crustacea. The generated gene expression data suggest that the innate immune system is probably not the main mechanism for the ability of *C. maenas* to defend itself against this pathogen. Instead, it is the RNAi pathway that shows high level activity across the duration of our experiment, with several miRNAs showing differential expression. Cma_pmiR-12 in particular is of great interest both because of its differential expression patterns and potential to target the endocytosis system. Expression changes of important endocytosis regulators were identified shortly after viral exposure. We hypothesize that in *C. maenas* viral entry via endocytosis is made more complicated due to changes in endocytosis transport processes. WSSV virions are probably recycled quickly from early endosomes back to the cellular exterior. Future experiments involving WSSV exposure in combination with endocytosis inhibition or overexpression/silencing of cma_pmiR-12 (or its targets) in both *C. maenas* and susceptible shrimp species should provide deeper insight into this process, and potentially be useful in identifying management and treatment strategies able to increase resistance of penaeid shrimp species to WSSV.

6.6 Supplementary Information

6.6.1 Annotation rate comparison for C. maenas and H. gammarus

In previous work, we produced a transcriptome for eight tissues of *C. maenas* [25] and a transcriptome for the nine tissues of the European lobster, *H. gammarus*. Transcripts of both of these assemblies were annotated, enabling comparison of annotation rates across these datasets. Transcript annotation with BLASTx to the NCBI-non redundant protein database was best for the *C. maenas* transcriptome from Verbruggen *et al.* 2015 while the transcriptomes for *H. gammarus* and the assembly produced from the exposure data showed similar percentages. All three transcriptomes had an annotation rate between 20-30 %. Since ultimately annotation relies on sequence similarity, this relatively low rate of annotation is probably the result of a lack of genomic data and experimental work in aquatic crustacean species as mentioned in Verbruggen *et al.* 2015 [25]. Analysis showed that the percentage of transcripts containing an ORF was lower in the Verbruggen *et al.* 2015 transcriptome but the subsequent comparison to UniProt/SwissProt yielded a much higher rate of protein sequence identification. This is not an effect of a better assembly, but more likely a function of the similarity threshold criteria applied during the annotation, since in Verbruggen *et al.* 2015 an e-value threshold of 1e^{-3} was applied, compared to 1e^{-5} in the other assemblies. Thus the low annotation rates are not an artefact of the experiment described in this chapter specifically, but an expected characteristic of transcriptome annotations in aquatic crustacean species, for which limited annotation for related species is available.

Table 14 *C. maenas* transcriptome annotation statistics

Input	**Annotation method**	**Number of annotated transcripts**	***C. maenas* transcriptome annotation percentage**	***H. gammarus* transcriptome annotation percentage**
Trinity transcripts	BLASTx – NCBI nr protein	30,090 (23.3 %)	29.6	24.2
Trinity transcripts	BLAST2GO	23,855 (18.5 %)	3.8	11.6
Trinity transcripts	KEGG	5,857 (4.5 %)	14.3	6.8
Trinity transcripts	TransDecoder ORF finder	59,777 (46.3 %)	27.5	51.6
TransDecoder Peptides	BLASTp – UniProt/SwissProt	18,508 (31.0 %)	70.4	22.7
TransDecoder Peptides	Pfam	17,005 (28.4 %)	67.4	34.5
TransDecoder Peptides	SignalP	4,129 (6.9 %)	1.9	5.9
TransDecoder Peptides	TmHMM	14,920 (25.0 %)	0	13.7

6.6.2 Comparison of C. maenas ‘ metazoan’ transcriptomes

In previous work we characterized a *de novo* transcriptome for *C. maenas* based on RNA extracted from multiple tissue types [25] (Chapter 3). That transcriptome was larger than the one produced in the current experiment, containing 212,427 transcripts in total. The size difference can be attributed to the fact that only a single tissue was used in the current experiment compared to twelve in the assembly of Verbruggen *et al.* 2015 [25] (Chapter 3). The contents of both transcriptomes were compared through BLAST searches to check for consistency across different experiments. The GC content of both transcriptomes was similar with 43.18 % in the current assembly compared to 44.53 % in Verbruggen *et al.* 2015 [25]. Sequence similarity showed that there is a reasonable degree of agreement between both assemblies. Comparing nucleotide sequences from each assembly showed that

96,729 (75 %) of transcripts from the current assembly had significant similarity to 116,337 (55 %) transcripts in the Verbruggen *et al.* 2015 assembly. As the number show there is a group of transcripts without a counterpart in the other assembly. Filtering for metazoan sequences reduced the unrepresented group. Of the metazoan transcripts in the current assembly 20,845 (91 %) showed significant sequence similarity to 36,508 (61 %) metazoan transcripts in Verbruggen *et al.* 2015. Like the overall transcriptome size, the larger unrepresented group in the latter is likely to be due to the additional tissues represented. These results demonstrate that filtering for metazoan transcripts increases overlap between transcriptomes assembled in separate experiments. Therefore it can be recommended as an analysis strategy since it facilitates comparisons across experiments and reduces dataset size and complexity. However, it should be noted that this may also result in the loss of novel transcripts.

6.6.3 WSSV transcriptome assembly and analysis

Insight into the transcriptome of WSSV was gained through alignment of RNA-sequencing reads to the genome of WSSV-CN. The coverage of the WSSV genome by RNA-sequencing reads is shown in Figure 11. From this analysis, it was apparent that generally the whole WSSV genome was transcribed into mRNA. Across the genome, only 15 bases were not covered by the RNA-sequencing data: bases at positions 32,272-32,283, within *wsv060*, and at positions 240,187-240,189. Given that the rest of the genome was covered, it is likely that these regions indicate a deletion in the genome of the injected virus. The difference in coverage of regions is large; the median is 589x whereas the maximum is 267,400x. The region of maximal coverage lays within *wsv230*, a gene that produces ICP11, and has been described by Wang *et al.* 2007 as the most highly expressed gene in WSSV [90].

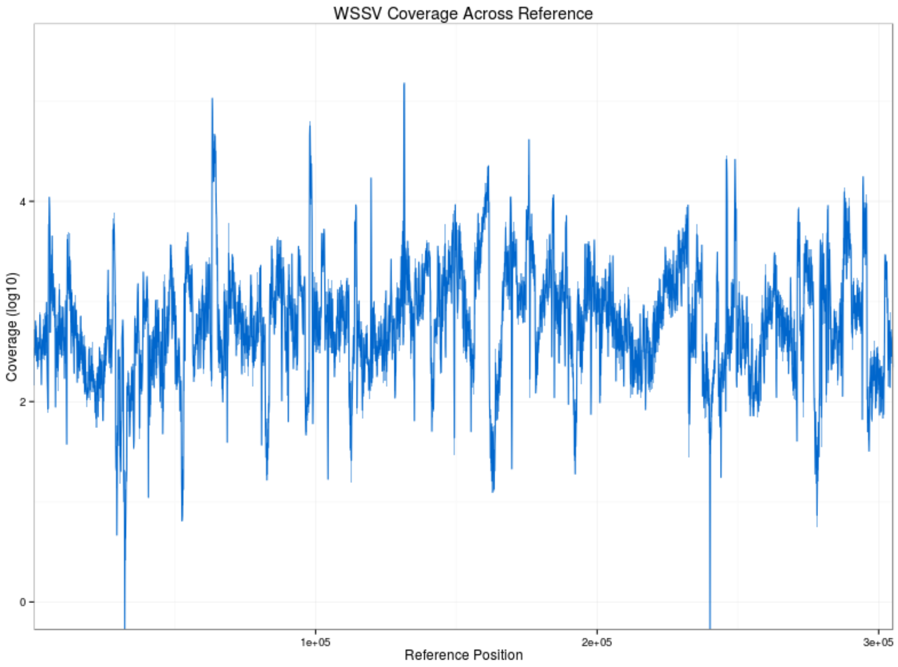


Figure 11 Coverage of WSSV genome by RNA-sequencing reads. The graph shows the $\log_{10}$ coverage of each base in the WSSV-CN genome (AF332093.3). The coverage of each position in the WSSV-CN genome represents the sum of coverage in the ws_36, ws_37, ws_48, ws_50 and ws_51 samples.

The sequencing reads of viral origin were used to generate a transcriptome for WSSV. Both genome-guided and *de novo* methods were applied with various settings. Assembly statistics are shown in Table 15 and the transcripts were aligned to the WSSV-CN genome and visualized in Figure 12. The statistics show that most assemblies were similar in size with a large N50 values. This indicates that the assembled transcriptomes consist of few, but long, transcripts. Of note, the Cufflinks assembly consisted of a single transcript the cause of which lies in the coverage of the WSSV genome and likely corresponds to a technical error of the assembler. The visualization in Figure 12 also illustrates this point. Comparing the assembled transcripts to the existing WSSV-CN annotation shows that many assembled transcripts cover multiple open reading frames (ORF).

Table 15 WSSV transcriptome assembly statistics

Quast	**Trinity (genome guided)**		**Trinity (*de novo*)**		**Cufflinks**
	Default	Min_cov/glue 100	Default	Min_cov/glue 100	
Contigs	76	242	75	65	1
Contigs (> 500 bp)	59	131	53	49	
Total length	337328	246749	412159	378973	305119
Total length	332474	212628	405258	373453	

(> 500 bp)					
Largest contig	18920	10652	35072	35063	305119
GC %	40.56	40.56	40.72	40.77	
N50	9768	2089	16673	13504	

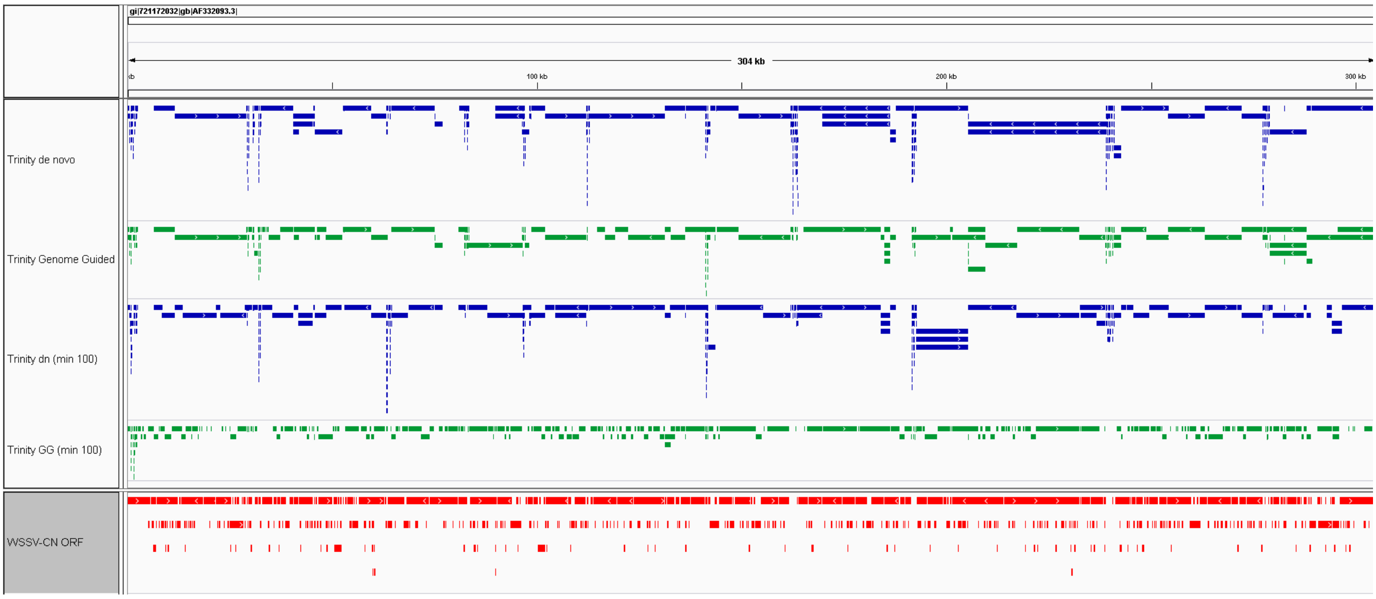


Figure 4 IGV visualization of WSSV transcriptome assemblies. Trinity *de novo* (blue) and genome guided (green) assemblies aligned to the WSSV-CN genome. Both assemblies have a version with default parameters and one with *–min_glue* / *--min_kmer_cov* of 100. The Genbank annotation track is shown in Red.

Additionally, the alignment of assembled transcripts to the WSSV genome shows that both *de novo* and genome guided assemblies are sensitive to the settings chosen. For example: Trinity genome guided and *de novo* assemblies (each at different settings) around WSSV-CN 50kb generate different results. Both the combining of multiple ORFs in single transcripts and the capriciousness of the assemblies can be ascribed to polycistronic mRNAs which are widespread in eukaryotic viruses [91]. Indeed, it has been shown that WSSV employs many polycistronic mRNAs to produce its proteins [92]. Many open reading frames identified in the WSSV genome do not contain a polyadenylation (polyA) signal. Such open reading frames lay within clusters of ORFs where only the last ORFs contains the polyA signal e.g. the *vp31/vp39b/vp11* cluster [93]. An additional complication is that such clusters can give rise to multiple polycistronic mRNAs. Kang *at al.* 2009, for example, showed that the *VP60/wsv419/wsv420/VP28* in WSSV-CN (*vp60b/wssv478/wssv479/vp28* in WSSV-TW AF440570) cluster transcribed: a ~3.4 kb polycistronic mRNA that encoded the VP60b and VP28 proteins; a ~1.3 kb bicistronic mRNA that included wssv479 and vp28; and a ~1.0 kb monocistronic mRNA containing only VP28. Thus within this cluster VP28 can be transcribed as three separate transcripts. The differences in coverage of the *VP60/wsv419/wsv420/VP28* cluster in the RNA sequencing data illustrates that it is likely that this cluster is indeed represented by different mRNAs (Figure 13). The coverage of the *VP28* is higher compared to the other ORFs in the cluster, indicating that there is indeed high expression of a *VP29* monocistronic mRNA.

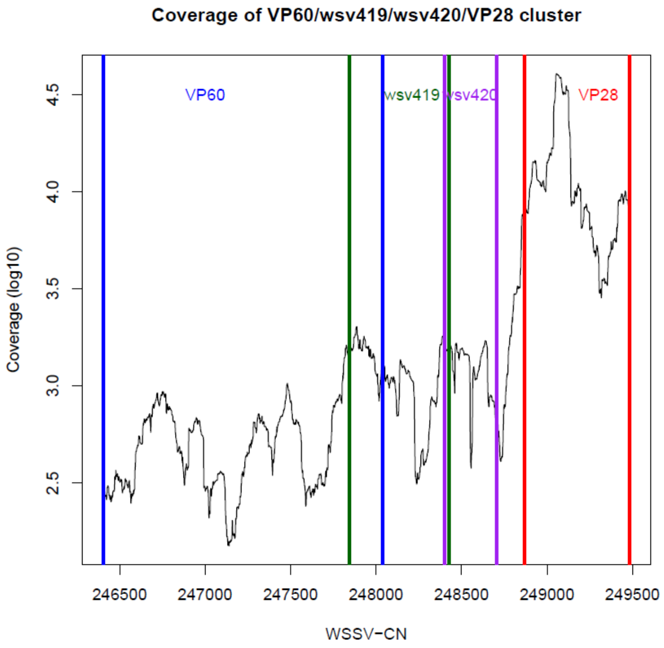


Figure 13 RNA sequencing coverage of WSSV ORF cluster. The coverage of the genomic region of WSSV-CN containing the *VP60/wsv419/wsv420/VP28* cluster. ORF boundaries are indicated by coloured vertical lines. Blue: *VP60*. Green: *wsv419*. Purple: *wsv420*. Red: *VP28*. The cluster has been shown to code for multiple poly-/monocistronic mRNAs. The coverage of the *VP28* is higher than other ORFs in the cluster, indicating high expression of the *VP28* monocistronic mRNA.

Both genome-guided and *de novo* assemblers will combine the reads in this cluster into a single transcript because there are many reads supporting it (i.e. the reads of the polycistronic mRNA). Thus the monocistronic *VP28* mRNA will be lost from the assembly. In addition in follow up gene expression calculations the reads of *VP60/wsv419/wsv420/VP28* are combined into a single FPKM value. It is possible to change the parameters of the assembler to improve results, e.g. a minimal threshold for joining contigs (--*min_glue* and --*min_kmer_cov* in Trinity). However given the variation in expression it is unlikely that a single parameter value can be found to resolve poly-/monocistronic situations throughout the transcriptome (see Figure 12). Assemblies produced with a reduced dataset, e.g. only aligning reads from the ws_52 sample, contained more transcripts, thus bringing out more monocistronic mRNAs but again were not consistent across the transcriptome (Table S 2). The assembly of this viral transcriptome is not tractable with currently available assemblers and would require a specialized assembler.

Since transcriptome assembly in WSSV proved complex, instead expression of WSSV ORFs was derived through mapping RNA sequencing reads directly to the ORF sequences as found in Genbank AF332093.3. Mapping was performed by bowtie2 and subsequent expression estimation by RSEM. Results showed that out of the 524 ORFs in the WSSV-CN genome, 361 had expression higher than 0 FPKM and 330 an expression higher than 100 FPKM. Thus, according to RSEM estimation, 163 ORFs were not expressed (Table S 3). Like the genomic coverages the expression values for most WSSV-CN ORFs lie between 100 – 10000 FPKM, see Figure 14. Measuring expression in FPKM confirms that ICP11 is indeed the most highly expressed WSSV gene. Two important structural proteins of WSSV, VP28 and VP26, are also expressed at high levels. Because of the time resolution and individual variability between samples the expression of WSSV ORFs cannot be compared in a meaningful way. However, the ws_52 sample data contained a much smaller amount of WSSV RNA sequencing reads which might be indicative of an earlier stage in infection. Therefore comparison of this sample to the other WSSV replicating samples (samples (ws_36, ws_37, ws_48, ws_50 and ws_51) might indicate which viral genes are expressed at earlier stages of infection. Analysis with EBSeq identified five DE genes between ws_52 and the other WSSV replicating: *wsv159*, *wsv430*, *wsv456*, *wsv103* and *wsv452*. *Wsv103* is an immediate early gene [17], the other have not been extensively studied.

Because of a lack of literature with similar analysis, it is difficult to place the expression of WSSV genes in *C. maenas* into context. It would be interesting to identify whether there are significant differences in WSSV gene expression between viral infections in susceptible penaeid shrimp species and *C. maenas*. These analyses could address the questions of whether WSSV change its gene expression patterns depending on the host organism and/or whether *C. maenas* inhibit expression of specific viral genes, thereby delaying replication and lowering infection success rate.

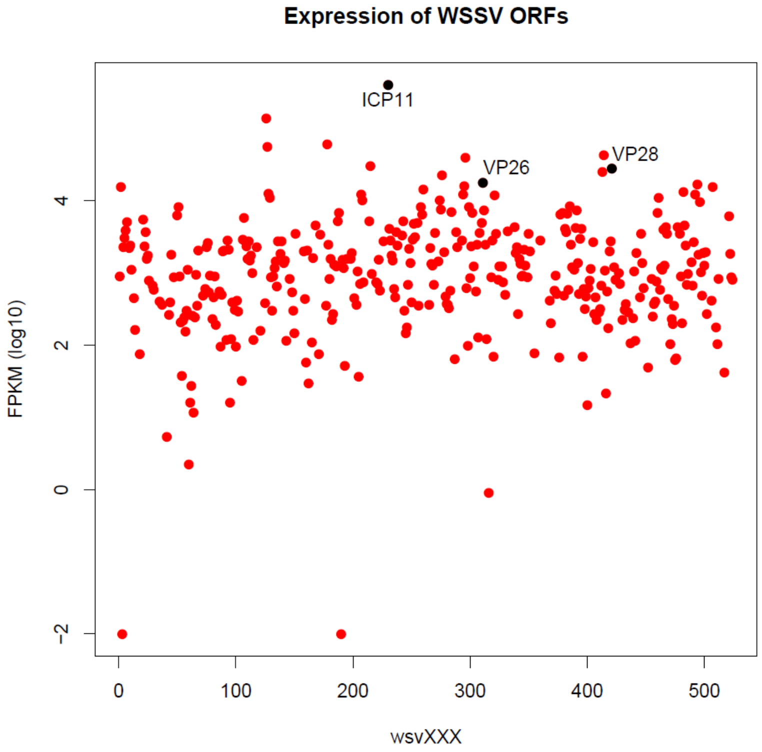


Figure 14 Expression of WSSV ORFs. The plot shows the expression values of WSSV ORFs in log$_{10}$ FPKM. ORFs are ordered according to their annotation in Genbank AF332093.3, e.g. *wsv230* is at position 230 on the x-axis. ICP11, the most expressed protein, and VP28/VP26 which are important structural proteins are highlighted.

6.7 Supplementary Tables

Table S 1 Percentage of reads aligning to WSSV genome

Sample_id	Time	WSSV alignment
con_02	0.01%	Time 0
con_04	0.01%	Time 0
con_06	0.02%	Time 0
con_07	0.01%	Time 0
con_10	0.01%	6 hours
con_12	0.01%	6 hours
con_15	0.01%	6 hours
con_16	0.01%	6 hours
con_19	0.01%	12 hours
con_22	0.01%	12 hours
con_23	0.01%	12 hours
con_24	0.01%	12 hours
con_25	0.01%	24 hours

con_28	0.01%	24 hours
con_30	0.01%	24 hours
con_31	0.01%	24 hours
con_34	0.01%	48 hours
con_36	0.01%	48 hours
con_37	0.02%	48 hours
con_40	0.00%	48 hours
con_41	0.01%	Day 7
con_42	0.01%	Day 7
con_45	0.01%	Day 7
con_48	0.01%	Day 7
con_50	0.01%	Day 14
con_51	0.01%	Day 14
con_54	0.01%	Day 14
con_56	0.01%	Day 14
con_57	0.02%	Day 28
con_60	0.01%	Day 28
con_62	0.01%	Day 28
con_64	0.01%	Day 28
ws_1	0.01%	6 hours
ws_2	0.01%	6 hours
ws_4	0.03%	6 hours
ws_6	0.01%	6 hours
ws_11	0.01%	12 hours
ws_12	0.01%	12 hours
ws_14	0.01%	12 hours
ws_16	0.01%	12 hours
ws_17	0.01%	24 hours
ws_18	0.01%	24 hours
ws_20	0.01%	24 hours
ws_23	0.01%	24 hours
ws_26	0.01%	48 hours
ws_29	0.01%	48 hours
ws_30	0.01%	48 hours
ws_31	0.02%	48 Hours
ws_35	0.01%	Day 7
ws_36	4.87%	Day 7
ws_37	2.78%	Day 7
ws_40	0.01%	Day 7
ws_41	0.01%	Day 14
ws_42	0.01%	Day 14

ws_47	0.01%	Day 14
ws_48	8.08%	Day 14
ws_49	0.01%	Day 28
ws_50	4.85%	Day 28
ws_51	7.57%	Day 28
ws_52	0.33%	Day 28
ws_53	0.01%	Day 28

Table S 2 Quast statistics for WSSV transcriptomes based on ws_52

Quast	**ws_52**	**ws_52**
	GG Trinity	de novo Trinity
Contigs	299	307
Contigs (> 500bp)	121	122
Contigs (> 1000bp)	62	60
Total length	223473	222498
Total length (> 500bp)	167497	164623
Total length (> 1000bp)	128575	123852
Largest contig	5286	5286
GC %	40.82	40.85
N50	1752	1653
N75	1023	1016
L50	28	28
L75	60	60

Table S 3 Unexpressed WSSV ORFs

Unexpressed WSSV ORFs (AF332093.3)
wsv007, wsv011, wsv014, wsv015, wsv016, wsv018, wsv019, wsv026, wsv027, wsv030, wsv031, wsv032, wsv033, wsv035, wsv037, wsv038, wsv039, wsv041, wsv045, wsv047, wsv048, wsv068, wsv069, wsv070, wsv072, wsv078, wsv083, wsv084, wsv090, wsv097, wsv100, wsv101, wsv105, wsv113, wsv114, wsv116, wsv117, wsv119, wsv120, wsv121, wsv134, wsv137, wsv141, wsv142, wsv144, wsv146, wsv147, wsv148, wsv149, wsv149a, wsv150, wsv156, wsv157, wsv160, wsv162, wsv163, wsv166, wsv167, wsv168, wsv169, wsv178, wsv182, wsv187, wsv189, wsv190, wsv193, wsv195, wsv203, wsv204, wsv205, wsv206, wsv208, wsv209, wsv210, wsv214, wsv216, wsv217, wsv218, wsv219, wsv229, wsv230, wsv231, wsv242, wsv246, wsv249, wsv250, wsv251, wsv252, wsv258, wsv259, wsv263, wsv271, wsv275, wsv276, wsv277, wsv288, wsv293, wsv297, wsv299, wsv304, wsv307, wsv310, wsv312, wsv314, wsv315, wsv316, wsv317, wsv318, wsv328, wsv331, wsv332, wsv333, wsv335, wsv336, wsv337, wsv338, wsv339, wsv340, wsv341, wsv342, wsv343, wsv344, wsv345, wsv348, wsv349, wsv353, wsv356, wsv366, wsv378, wsv382, wsv395, wsv398, wsv399, wsv402, wsv404, wsv407, wsv409, wsv410, wsv415, wsv416, wsv420, wsv421, wsv423, wsv424, wsv445, wsv454, wsv455, wsv458, wsv468, wsv469, wsv470, wsv473, wsv474, wsv477, wsv478, wsv479, wsv480, wsv482, wsv483, wsv484, wsv492, wsv520, wsv522

6.8 Supplementary Files

Supplementary file S1: Illumina_Sequencing_results.xlsx

Supplementary file S2: Trinotate annotation report.xlsx

Supplementary file S3: Differentially_expressed_transcripts.xlsx

Supplementary file S4: Enriched_GO_terms.xlsx

Supplementary file S5: Cmaenas_miRNAs.fa

Supplementary file S6: Cmaenas_miRNA_expression.txt

Supplementary file S7: cma_pmiR-12_miranda_targets_GOenrichment.txt

Supplementary file S8: WSSV_miRNAs.fa

6.9 References

1. FAO: **Fishery Statistical Collections: Global Aquaculture Production** In.; 2015. http://www.fao.org/fishery/statistics/
2. Anderson JL: **Shrimp production review**. *Conference GOAL 2013, Paris* 2013.
3. Durand S, Lightner DV, Redman RM, Bonami JR: **Ultrastructure and morphogenesis of White Spot Syndrome Baculovirus (WSSV)**. *Diseases of Aquatic Organisms* 1997, **29**:205-211.
4. Pradeep B, Rai P, Mohan SA, Shekhar MS, Karunasagar I: **Biology, Host Range, Pathogenesis and Diagnosis of White spot syndrome virus**. *Indian Journal of Virology : an official organ of Indian Virological Society* 2012, **23**(2):161-174.
5. Stentiford GD, Oidtmann B, Scott A, Peeler EJ: **Crustacean diseases in European legislation: Implications for importing and exporting nations**. *Aquaculture* 2010, **306**(1-4):27-34.
6. Lightner DV: **Global transboundry disease politics: the OIE perspective.** *Journal of Invertebrate Pathology* 2012, **110**:184-187.
7. Flegel TW, Lightner DV, Owens L: **Shrimp disease control: past, present and future**. *Diseases in Asian Aquaculture* 2008, **VI**:355-378.
8. Stentiford GD, Neil DM, Peeler EJ, Shields JD, Small HJ, Flegel TW, Vlak JM, Jones B, Morado F, Moss S *et al*: **Disease will limit future food supply from the global crustacean fishery and aquaculture sectors**. *Journal of Invertebrate Pathology* 2012, **110**(2):141-157.

9. Bateman KS, Tew I, French C, Hicks RJ, Martin P, Munro J, Stentiford GD: **Susceptibility to infection and pathogenicity of White Spot Disease (WSD) in non-model crustacean host taxa from temperate regions**. *Journal of Invertebrate Pathology* 2012, **110**(3):340-351.

10. Chou H-Y, Huang C-Y, Wang C-H, Chiang H-C, Lo C-F: **Pathogenicity of a baculovirus infection causing white spot syndrome in cultured penaeid shrimp in Taiwan** *Diseases of Aquatic Organisms* 1995, **23**:165-173.

11. Lightner DV: **A Handbook of Shrimp Pathology and Diagnostic Procedures for Diseases of Cultured Penaeid Shrimp.** *World Aquaculture Society, Baton Roughe, Louisiana, USA, 304pp* 1996.

12. Stentiford GD, Bonami JR, Alday-Sanz V: **A critical review of susceptibility of crustaceans to Taura syndrome, Yellowhead disease and White Spot Disease and implications of inclusion of these diseases in European legislation**. *Aquaculture* 2009, **291**(1-2):1-17.

13. Verbruggen B, Bickley L, van Aerle R, Bateman K, Stentiford G, Santos E, Tyler C: **Molecular Mechanisms of White Spot Syndrome Virus Infection and Perspectives on Treatments**. *Viruses* 2016, **8**(1):23.

14. Li DF, Zhang MC, Yang HJ, Zhu YB, Xu X: **β-integrin mediates WSSV infection**. *Virology* 2007, **368**:122-132.

15. Huang J, Li F, Wu J, Yang F: **White spot syndrome virus enters crayfish hematopoietic tissue cells via clathrin-mediated endocytosis**. *Virology* 2015, **486**:35-43.

16. Liu WJ, Chang YS, Wang AH, Kou GH, Lo CF: **White spot syndrome virus annexes a shrimp STAT to enhance expression of the immediate-early gene ie1**. *Journal of Virology* 2007, **81**(3):1461-1471.

17. Li F, Li M, Ke W, Ji Y, Bian X, Yan X: **Identification of the immediate-early genes of white spot syndrome virus**. *Virology* 2009, **385**(1):267-274.

18. Liu WJ, Chang YS, Huang WT, Chen IT, Wang KC, Kou GH, Lo CF: ***Penaeus monodon* TATA box-binding protein interacts with the white spot syndrome virus transactivator IE1 and promotes its transcriptional activity**. *Journal of Virology* 2011, **85**(13):6535-6547.

19. Asgari S: **Role of MicroRNAs in Insect Host–Microorganism Interactions**. *Frontiers in Physiology* 2011, **2**:48.

20. Skalsky RL, Cullen BR: **Viruses, microRNAs, and Host Interactions**. *Annual Review of Microbiology* 2010, **64**:123-141.

21. Huang T, Cui Y, Zhang X: **Involvement of Viral MicroRNA in the Regulation of Antiviral Apoptosis in Shrimp**. *Journal of Virology* 2014, **88**(5):2544-2554.

22. He Y, Zhang X: **Comprehensive characterization of viral miRNAs involved in white spot syndrome virus (WSSV) infection**. *RNA Biology* 2012, **9**(7):1019-1029.

23. Huang T, Zhang X: **Functional analysis of a crustacean microRNA in host-virus interactions**. *Journal of Virology* 2012, **86**(23):12997-13004.

24. Wang PH, Huang T, Zhang XB, He JG: **Antiviral defense in shrimp: From innate immunity to viral infection**. *Antiviral Research* 2014.

25. Verbruggen B, Bickley LK, Santos EM, Tyler CR, Stentiford GD, Bateman KS, van Aerle R: ***De novo* assembly of the *Carcinus maenas* transcriptome and characterization of innate immune system pathways**. *BMC Genomics* 2015, **16**:458.

26. Holthuis LB: **FAO CATALOAGUE Vol.1 - Shrimps and Prawns of the World.** . *FAO Fisheries Synopsis* 1980, **1**(No. 125).

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

28. Gurevich A, Saveliev V, Vyahhi N, Tesler G: **QUAST: quality assessment tool for genome assemblies**. *Bioinformatics* 2013, **29**(8):1072-1075.

29. Parra G, Bradnam K, Korf I: **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes**. *Bioinformatics* 2007, **23**(9):1061-1067.

30. **TransDecoder** [http://transdecoder.sourceforge.net/]

31. Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching**. *Nucleic Acids Research* 2011, **39**(Web Server issue):W29-37.

32. Petersen TN, Brunak S, von Heijne G, Nielsen H: **SignalP 4.0: discriminating signal peptides from transmembrane regions**. *Nature Methods* 2011, **8**(10):785-786.

33. Lagesen K, Hallin P, Rødland EA, Stærfeldt H-H, Rognes T, Ussery DW: **RNAmmer: consistent and rapid annotation of ribosomal RNA genes**. *Nucleic Acids Research* 2007, **35**(9):3100-3108.

34. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: **KAAS: an automatic genome annotation and pathway reconstruction server**. *Nucleic Acids Research* 2007, **35**(Web Server issue):W182-185.

35. Huson DH, Mitra S, Ruscheweyh HJ, Weber N, Schuster SC: **Integrative analysis of environmental sequences using MEGAN4**. *Genome Research* 2011, **21**(9):1552-1560.

36. Yang F, He J, Lin X, Li Q, Pan D, Zhang X, Xu X: **Complete genome sequence of the shrimp white spot bacilliform virus**. *Journal of Virology* 2001, **75**(23):11811-11820.

37. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nature Methods* 2012, **9**(4):357-359.

38. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.

39. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer**. *Nature Biotechnology* 2011, **29**(1):24-26.

40. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics* 2010, **26**(6):841-842.

41. Ghosh S, Chan C-K: **Analysis of RNA-Seq Data Using TopHat and Cufflinks**. In: *Plant Bioinformatics.* Edited by Edwards D, vol. 1374: Springer New York; 2016: 339-361.

42. Kent WJ: **BLAT—The BLAST-Like Alignment Tool**. *Genome Research* 2002, **12**(4):656-664.

43. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome**. *BMC Bioinformatics* 2011, **12**:323.

44. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, Haag JD, Gould MN, Stewart RM, Kendziorski C: **EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments**. *Bioinformatics* 2013, **29**(8):1035-1043.

45. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN: **RNA-Seq gene expression estimation with read mapping uncertainty**. *Bioinformatics* 2010, **26**(4):493-500.

46. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research**. *Bioinformatics* 2005, **21**(18):3674-3676.

47. Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N: **miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades**. *Nucleic Acids research* 2012, **40**(1):37-52.

48. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ: **miRBase: microRNA sequences, targets and gene nomenclature**. *Nucleic Acids Research* 2006, **34**(suppl 1):D140-D144.

49. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ: **miRBase: tools for microRNA genomics**. *Nucleic Acids Research* 2008, **36**(suppl 1):D154-D158.

50. Kozomara A, Griffiths-Jones S: **miRBase: integrating microRNA annotation and deep-sequencing data**. *Nucleic Acids Research* 2011, **39**(suppl 1):D152-D157.

51. Kozomara A, Griffiths-Jones S: **miRBase: annotating high confidence microRNAs using deep sequencing data**. *Nucleic Acids Research* 2014, **42**(D1):D68-D73.

52. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics* 2010, **26**(1):139-140.

53. Thadani R, Tammi MT: **MicroTar: predicting microRNA targets from RNA duplexes**. *BMC Bioinformatics* 2006, **7 Suppl 5**:S20.

54. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS: **Human MicroRNA Targets**. *PLoS Biology* 2004, **2**(11):e363.

55. Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, Tahmasebpoor S, Danielsson A, Edlund K *et al*: **Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics**. *Molecular & Cellular Proteomics* 2014, **13**(2):397-406.

56. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW: **Reagent and laboratory contamination can critically impact sequence-based microbiome analyses**. *BMC Biology* 2014, **12**.

57. Marks H, Vorst O, van Houwelingen AM, van Hulten MC, Vlak JM: **Gene-expression profiling of White spot syndrome virus *in vivo***. *The Journal of General Virology* 2005, **86**(Pt 7):2081-2100.

58. Marks H, Mennens M, Vlak JM, van Hulten MC: **Transcriptional analysis of the white spot syndrome virus major virion protein genes**. *The Journal of General Virology* 2003, **84**(Pt 6):1517-1523.

59. Eddie Ip WK, Takahashi K, Alan Ezekowitz R, Stuart LM: **Mannose-binding lectin and innate immunity**. *Immunological Reviews* 2009, **230**(1):9-21.

60. Tellam RL, Wijffels G, Willadsen P: **Peritrophic matrix proteins**. *Insect Biochemistry and Molecular Biology* 1999, **29**(2):87-101.

61. Khayat M, Babin PJ, Funkenstein B, Sammar M, Nagasawa H, Tietz A, Lubzens E: **Molecular Characterization and High Expression During Oocyte Development of a Shrimp Ovarian Cortical Rod Protein Homologous to Insect Intestinal Peritrophins**. *Biology of Reproduction* 2001, **64**(4):1090-1099.

62. Du X-J, Wang J-X, Liu N, Zhao X-F, Li F-H, Xiang J-H: **Identification and molecular characterization of a peritrophin-like protein from fleshy prawn (*Fenneropenaeus chinensis*)**. *Molecular Immunology* 2006, **43**(10):1633-1644.

63. Hall TMT: **Structure and Function of Argonaute Proteins**. *Structure* 2005, **13**(10):1403-1408.

64. Yang L, Li X, Jiang S, Qiu L, Zhou F, Liu W, Jiang S: **Characterization of Argonaute2 gene from black tiger shrimp (*Penaeus monodon*) and its responses to immune challenges**. *Fish & Shellfish Immunology* 2014, **36**(1):261-269.

65. Shekhar MS, Ponniah AG: **Recent insights into host-pathogen interaction in white spot syndrome virus infected penaeid shrimp**. *Journal of Fish Diseases* 2014.

66. Wen R, Li F, Li S, Xiang J: **Function of shrimp STAT during WSSV infection**. *Fish and Shellfish Immunology* 2014, **38**(2):354-360.

67. Kim M, Jeon J-M, Oh C-W, Kim YM, Lee DS, Kang C-K, Kim H-W: **Molecular characterization of three crustin genes in the morotoge shrimp, *Pandalopsis japonica***. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* 2012, **163**(2):161-171.

68. Hauton C, Brockton V, Smith VJ: **Cloning of a crustin-like, single whey-acidic-domain, antibacterial peptide from the haemocytes of the European lobster, *Homarus gammarus*, and its response to infection with bacteria**. *Molecular Immunology* 2006, **43**(9):1490-1496.

69. Schnapp D, Kemp GD, Smith VJ: **Purification and characterization of a proline-rich antibacterial peptide, with sequence similarity to bactenecin-7, from the haemocytes of the shore crab, *Carcinus maenas***. *European Journal of Biochemistry* 1996, **240**(3):532-539.

70. Antony SP, Philip R, Joseph V, Singh ISB: **Anti-lipopolysaccharide factor and crustin-III, the anti-white spot virus peptides in *Penaeus monodon*: Control of viral infection by up-regulation**. *Aquaculture* 2011, **319** 11-17.

71. Mercer J, Schelhaas M, Helenius A: **Virus entry by endocytosis**. *Annual Review of Biochemistry* 2010, **79**:803-833.

72. Sritunyalucksana K, Wannapapho W, Lo CF, Flegel TW: **PmRab7 is a VP28-binding protein involved in white spot syndrome virus infection in shrimp**. *Journal of Virology* 2006, **80**(21):10734-10742.

73. Gimenez MC, Rodriguez Aguirre JF, Colombo MI, Delgui LR: **Infectious bursal disease virus uptake involves macropinocytosis and trafficking to early endosomes in a Rab5-dependent manner**. *Cellular Microbiology* 2015, **17**(7):988-1007.

74. Rizopoulos Z, Balistreri G, Kilcher S, Martin CK, Syedbasha M, Helenius A, Mercer J: **Vaccinia Virus Infection Requires Maturation of Macropinosomes**. *Traffic* 2015, **16**(8):814-831.

75. Sun E, He J, Zhuang X: **Dissecting the Role of COPI Complexes in Influenza Virus Infection**. *Journal of Virology* 2013, **87**(5):2673-2685.

76. Cureton DK, Burdeinick-Kerr R, Whelan SP: **Genetic inactivation of COPI coatomer separately inhibits vesicular stomatitis virus entry and gene expression**. *Journal of Virology* 2012, **86**(2):655-666.

77. Chang HC, Newmyer SL, Hull MJ, Ebersold M, Schmid SL, Mellman I: **Hsc70 is required for endocytosis and clathrin function in *Drosophila***. *The Journal of Cell Biology* 2002, **159**(3):477-487.

78. Eisenberg E, Greene LE: **Multiple Roles of Auxilin and Hsc70 in Clathrin-Mediated Endocytosis**. *Traffic* 2007, **8**(6):640-646.

79. Chuang CK, Yang TH, Chen TH, Yang CF, Chen WJ: **Heat shock cognate protein 70 isoform D is required for clathrin-dependent endocytosis of Japanese encephalitis virus in C6/36 cells**. *The Journal of General Virology* 2015, **96**(Pt 4):793-803.

80. Vale-Costa S, Amorim MJ: **Recycling Endosomes and Viral Infection**. *Viruses* 2016, **8**(3).

81. Yi G, Wang Z, Qi Y, Yao L, Qian J, Hu L: **Vp28 of Shrimp White Spot Syndrome Virus Is Involved in the Attachment and Penetration into Shrimp Cells**.*Journal of Biochemistry & Molecular Biology* 2004, **37**:726-734.

82. van Hulten MCW, Witteveldt J, Snippe M, Vlak JM: **White spot syndrome virus envelope protein VP28 is involved in the systemic infection of shrimp**. *Virology* 2001, **285**:228-233.

83. Waikhom G, John KR, George MR, Jeyaseelan MJP: **Differential host passaging alters pathogenicity and induces genomic variation in white spot syndrome virus**. *Aquaculture* 2006, **261**(1):54-63.

84. Huang XD, Zhao L, Zhang HQ, Xu XP, Jia XT, Chen YH, Wang PH, Weng SP, Yu XQ, Yin ZX *et al*: **Shrimp NF-kappaB binds to the immediate-early gene ie1 promoter of white spot syndrome virus and upregulates its activity**. *Virology* 2010, **406**(2):176-180.

85. Tang H: **Regulation and function of the melanization reaction in Drosophila**. *Fly* 2009, **3**(1):105-111.

86. Chen IT, Aoki T, Huang YT, Hirono I, Chen TC, Huang JY, Chang GD, Lo CF, Wang HC: **White spot syndrome virus induces metabolic changes resembling the warburg effect in shrimp hemocytes in the early stage of infection**. *Journal of Virology* 2011, **85**(24):12919-12928.

87. Yeh F-C, Wu S-H, Lai C-Y, Lee C-Y: **Demonstration of nitric oxide synthase activity in crustacean hemocytes and anti-microbial activity of hemocyte-derived nitric oxide**. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* 2006, **144**(1):11-17.

88. Griffiths-Jones S: **The microRNA Registry**. *Nucleic Acids Research* 2004, **32**(suppl 1):D109-D111.

89. Bach I: **The LIM domain: regulation by association**. *Mechanisms of Development* 2000, **91**(1–2):5-17.

90. Han-Ching W, Hao-Ching W, Guang-Hsiung K, Chu-Fang L, Wei-Pang H: **Identification of icp11, the most highly expressed gene of shrimp white spot syndrome virus (WSSV)**. *Diseases of Aquatic Organisms* 2007, **74**(3):179-189.

91. Ryabova LA, Pooggin MM, Hohn T: **Viral strategies of translation initiation: Ribosomal shunt and reinitiation**. In: *Progress in Nucleic Acid Research and Molecular Biology.* vol. Volume 72: Academic Press; 20021-39.

92. Kang ST, Leu JH, Wang HC, Chen LL, Kou GH, Lo CF: **Polycistronic mRNAs and internal ribosome entry site elements (IRES) are widely used by white spot syndrome virus (WSSV) structural protein genes**. *Virology* 2009, **387**(2):353-363.

93. Tsai JM, Wang HC, Leu JH, Hsiao HH, Wang AH, Kou GH, Lo CF: **Genomic and proteomic analysis of thirty-nine structural proteins of shrimp white spot syndrome virus**. *Journal of Virology* 2004, **78**(20):11360-11370.

Chapter 7

General Discussion

Chapter 7: General Discussion

The crustacean aquaculture sector is an important part of global food production. FAO statistics from 1985 to the current decade show that the sector has grown and is likely to continue to do so at a similar or greater rate in the future. The largest obstacle to further growth of the industry is the prevalence of disease epidemics, and over the last decade the most devastating disease has been caused by White Spot Syndrome Virus. The host range of WSSV is large and is thought to include most aquatic crustacean species, including the commercially important penaeid shrimp. Experiments carried out by Bateman *et al*. showed that the shore crab *C. maenas* is a crustacean with the greatest apparent resistance to this virus amongst a group of seven ecologically or economically important crustacean species from across Europe [1]. This observation was the basis of the hypothesis addressed in this thesis which aimed to identify the how this resistance operates. Understanding the resistance mechanism(s) of *C. maenas* might enable exploitation of this knowledge for the protection of disease sensitive and economically important penaeid shrimp species.

In the first phase of this thesis work knowledge was gathered on the process of WSSV infection, with a focus on the molecular interactions that occur between WSSV and proteins of the host cell during viral entry and replication. Because of the importance of WSSV to the aquaculture sector there has been a large body of research into this viral disease. The genomes of four WSSV isolates have been assembled and annotated. Furthermore, the function and interactions of many WSSV proteins have been elucidated. However there are still many significant knowledge gaps. It is known that WSSV enters host cells through Clathrin-mediated endocytosis but how the virions reach the host nucleus is unclear. The expression of WSSV genes has resulted in identification of immediate early, early and late genes along with transcription factors that can bind promoters. Changes that occur in the host cell in the presence of the virus have also been documented. The last stages of virion assembly and host cell exit have not been explored in much detail in the literature. The acquired knowledge of WSSV has been applied in the development of various treatments but thus far none have been applied outside the laboratory environment.

The experimental work in this thesis involved the generation of genomic resources for *C. maenas* and *H. gammarus*. For *H. gammarus* a transcriptome was assembled and for *C. maenas* a transcriptome and draft genome were assembled and annotated. Now follows a discussion of the main conclusions, strengths and weaknesses of each of the research findings in the different thesis chapters.

7.1 Carcinus maenas transcriptome assembly

The assembled transcriptome contained over 200,000 transcripts which is likely larger than the real number of transcripts *C. maenas* produces. The higher than expected number of transcripts may have resulted from a number of factors. Firstly, RNA sequencing does not distinguish reads based on their origin. This means that RNA from organisms, like bacteria, present in the sample will be sequenced along with RNA of the species of interest. Thus, transcripts of foreign origin are included in the transcriptome of the species of interest. Secondly, it can be the case that single transcripts are fragmented and subsequently represented by multiple smaller transcripts. This artefact of RNA-sequencing could be reduced by performing sequencing on platforms that offer longer reads. Furthermore, the *C. maenas* transcriptome was based on a pooled RNA sample, meaning that for every tissue RNA was pooled from four different individuals. Since the genotypes of these individuals are different, this can cause problems during the assembly process and a degree of redundancy can be introduced to the assembled transcriptome.

Benefits and drawbacks of pooled samples

Based on the issue of pooling samples, it can be debated whether it is better to generate sequencing data from a single individual or from a pooled sample of multiple individuals. Using a single individual, provided enough RNA can be isolated from the required tissues, would reduce redundancy due to genotype variation. However, using a single individual might result in a biased representation of transcripts since it is a snapshot of the status in that one individual at that point in time. If the individual experienced a stressor, such as an infection, then this will be reflected in the expression of its genes and thus the assembled transcriptome. Naturally, this variation can be reduced by using RNA pooled from different individuals. Once pooled, it is assumed that the resulting data is a better representation of a 'normal' transcriptome for that species. The pooling of individuals is easily done for microorganisms. However when considering larger eukaryotes it becomes more labour intensive and expensive to isolate RNA from many individuals, especially when RNA is separated based on tissue type. One of the aims for the transcriptome assembly was to investigate tissue specific expression. Unfortunately, the design of the RNA sequencing libraries in Chapter 3 was performed in a manner that did not allow separation of data based on the individual it was derived from. For example, the sequencing reads for the gills were generated from a pooled RNA sample of four individuals. This pooled sample was provided a certain barcode which allowed these data to be separated from the data of other tissues in the sequencing run. The individual resolution is lost in this manner. Should enough barcodes be available, it is possible to label every individual-tissue with a unique barcode which allows combining/separating data based on origin. However, it was also the case that the overall aim of the transcriptome assembly was not to investigate the effects of pooling on sequencing data/assembly but rather the analysis of tissue specific expression. To this end the experimental design (pooling prior to labelling) was sufficient for purpose. Also, improving the resolution to individual-tissue level represented a significant investment of labour and resources, which would not contribute to the aim of the experiment.

Transcript annotation

After assembly, the annotation of the transcripts relies on sequence similarity to known transcripts/proteins. The success rate of annotation for a new transcriptome is highly dependent on the availability of known transcript/protein annotations of closely related species. For example, the annotation of a *de novo* transcriptome for a new *Drosophila* species is greatly facilitated by the broad knowledge available for other *Drosophila* species (e.g. *Drosophila melanogaster*). Unfortunately, aquatic crustaceans are not supported by such a large body of information. For *Carcinus maenas*, annotation had to be derived mainly from *Penaeus* shrimp species, *Daphnia*, planktonic crustaceans or insects like *T. castaneum* or *D. melanogaster.* Due to the very limited knowledge base for aquatic Crustacea only around 30 % of transcripts could be annotated, even when sequence similarity thresholds were lowered to BLASTX e-values below 1e^{-5}. At those e-values it is possible to get sequence similarities around 30%, but it becomes more difficult to justify whether the annotation is truly meaningful. Another pitfall can occur when an annotation yields similarity to a transcript in a database that was itself annotated by prediction through sequence similarity. For example: *C. maenas* transcript 1 shows BLASTX similarity to giant honeybee (*Apis dorsata*) protein A, which is predicted to be similar to *D. melanogaster* Dscam. In this case, how well *C. maenas* transcript 1 actually resembles *D. melanogaster* Dscam is open to major questioning. The size of *de novo* transcriptomes, often over 100k transcripts, makes it impractical to manually investigate every annotation. Furthermore, raising thresholds would result in an even lower rate of annotation. Missing annotation could also be due to an incorrect assembly of a transcript by the *de novo* assembler. For example, due to insufficient coverage a single mRNA can be fragmented and split into multiple shorter transcripts in the transcriptome assembly. Sequencing errors too can cause problems in assembly, often encountered in lowly expressed transcripts. Again, the transcriptome size often makes it intractable to check for these assembly artefacts.

Filtering

The low level of *C. maenas* transcript annotation complicated the subsequent analysis. As mentioned, it is likely that some of the transcripts in the transcriptome did not derive from *C. maenas* but instead from microorganisms present in the tissue at time of isolation. The annotation of transcripts was used to derive which transcripts were of metazoan origin and therefore likely to be genuine *C. maenas* transcripts. Applying such a filter to the transcriptome reduced its size significantly, but at the cost of removing potential novel *C. maenas* transcripts. Removal of novel *C. maenas* transcripts was not a major problem for the analysis however, because in the original work we were mainly interested in the presence of conserved pathways of the innate immune system. Where this could be detrimental is in the identification of novel antimicrobial peptides (AMP), especially ones that do not bear resemblance to known AMP families. So depending on the question asked of the data, one could consider removing the filter in order to identify novel transcripts. It is likely that over the coming years more genomic resources will become available for aquatic crustaceans and it could be beneficial to repeat the annotation step conducted in this thesis in order to benefit from this new information. Repeat analyses could include more transcripts into the metazoan-annotated group and thus improve the power of analysis performed on the generated dataset.

Gene expression analysis

Expression values for transcripts were used to derive differentially-expressed transcripts across tissues. Because the dataset did not contain different treatment conditions or whole body samples the analysis was performed by comparing a single tissue to all the remaining tissues. The latter were assumed to approximate for the whole body. However, for every tissue only a single sample is available. This lack of replicates is a weak point in the analysis since estimation of expression variation is not possible. As discussed in the introduction to this thesis, overdispersion is often observed in RNA sequencing datasets and its estimation is necessary for differential expression calculations. In the *C. maenas* analysis this issue was resolved by setting a constant dispersion for every tissue according to recommendations by the statistics package (edgeR), however, this is a suboptimal solution. In designing new NGS experiments it is recommended to always include replicates since it strengthens subsequent statistical analysis. A recent publication by Schurch *et al*. recommends using 6 to 12 biological replicates depending on the desired resolution in detection of differentially expressed transcripts [2].

Immune system annotation

The immune system of *C. maenas* was characterized through two methods. The first was through the use of the KEGG pathway database and secondly through a semi-automated software pipeline. The KEGG database was used because it provides good visual representation of curated pathways. However KEGG pathways are based mainly on vertebrate/mammalian species. As such its use is limited since KEGG orthology groups are assigned to *C. maenas* transcripts based on sequence similarity to vertebrate species. Furthermore, there are differences in the pathways for invertebrates compared to vertebrates. Therefore we also used a semi-automated software pipeline that allowed identification of components of the invertebrate system. This pipeline involved searching for known protein sequences in the NCBI protein database, potentially with added terms to filter undesired sequences. Searching for STAT sequences, for example, could be done through the following query: “STAT[Protein name] pancrustacea not hypothetical not putative not predicted not uncharacterized not partial”. The results of such queries were then used as input for a TBLASTN search against the *C. maenas* transcriptome in order to identify transcripts with significant similarity. However, the results required some manual post-processing, hence the use of a semi-automated pipeline, because the NCBI query could still yield undesirable proteins. When the “[Protein name]” filter was not used in the above query then the results would yield proteins like luciferase, which are not related to STAT. However, in some cases the [Protein name] addition was too strict for the NCBI search to yield results and thus had to be removed. Manual inspection was facilitated by a filter that produced the best hits for transcripts based on taxonomy. It would be interesting to see whether the capabilities of this pipeline can be developed further to remove the need for the manual aspect entirely, making it easier to use and less biased by user input. Overall, it was found that almost all the expected components of the invertebrate immune system were present in the *C. maenas* transcriptome, and those that are absent may be due to their absence in the sequencing dataset, for example because of low expression levels in the selected tissues analysed.

NGS technology can deliver large volumes of data that can be applied to effectively and comprehensively answer scientific questions and provide platforms onto which further studies can be based. We used *de novo* assemblers to generate a transcriptome that has great utility for future studies in *C. maenas* or closely related organisms. However, as the results of Chapter 3 have shown, the application of RNA-sequencing to organisms in species with little genomic resources is challenging due to problems with annotation. It is thus likely that the results in Chapter 3 contain errors in the assembly and annotation, but the size of data makes it extremely difficult to manually deduce these errors. Improvements in assemblers and sequence annotations in *C. maenas* and its close relatives would strengthen of the current dataset.

7.2 *Carcinus maenas* genome assembly

DNA isolated from *C. maenas* muscle was used as a basis for a sequencing experiment with the aim to produce a genome scaffold. Analysis of the generated sequencing data indicated issues with quality, in particular in the matepair libraries. In order to resolve the sequence quality problem, a repeat sequencing run was performed and stringent quality filters were applied to the data. Combining sequencing data into a single genome scaffold can be done through many different assembly pipelines. Currently there are no agreed standards by which the quality of assemblies can be assessed. It would be valuable to have an international standard for such assessments, which would allow comparisons between laboratories. Often assessment of quality and choice for final assembly are based on statistics describing the total number of contigs and distribution of contig sizes (e.g. N50), mainly because these are the only metrics that are available after assembly. The scaffolds generated in Chapter 4 were relatively similar in respect to these statistics, assemblies encompassing several millions of contigs. Metrics that take content of the assembly into account rather than pure size distribution statistics might be advisable for assembly evaluation. For example, one could consider including annotation into assembly choice, preferring assemblies that cover certain genes. The choice of genes can either be well-conserved genes, enabling comparison across laboratories/species, or specific for a project, favouring presence of genes likely to be related to a certain research topic. Again here a form of international guideline or evaluation method is advisable. However, annotation of an assembly is a computationally and time consuming process and repeating it for a set of assemblies would result in a significant delay of analysis with no guaranteed output, therefore incorporation into assembly choice is intractable at this time. Should faster annotation methods become available it is worth considering. Unlike the majority of DNA-sequencing projects, an RNA-sequencing data set was available for *C. maenas* which could be used as a proxy for annotation, as its mapping to a genome scaffold informs on total exon content. In the end, a choice was made for an assembly of limited size (338,980 scaffolds) that still incorporated a large portion of RNA-sequencing reads.

Library construction

The comparison between the draft genome assemblies for *C. maenas* (Chapter 4), *E. sinensis* [3] and *N. denticulata* [4] illustrates that genome sequencing experiments in relatively closely related species can have different outcomes depending on sequencing strategy. These three studies were all performed on the Illumina HiSeq platform, Illumina HiSeq-2500 for *C. maenas* and Illumina HiSeq-2000 for *E. sinensis* and *N. denticulata*. However the library construction schemes and sequencing depth were different in the different studies. Table 1 lists the study designed applied for each species. The *C. maenas* and *E. sinensis* studies both employed libraries with a large range of insert sizes whereas the sequencing of the *N. denticulata* genome was performed on a single library of a small size. The total amount of generated sequence data also showed a wide range, from the 12 x coverage of the *N. denticulata* to the 155 x coverage of the *E. sinensis* data. Out of the three species, the draft genome for *E. sinensis* showed the best results, achieving coverage of 67.5 % of the estimated genome size with 17,553 scaffolds (Table 2). The draft genomes for the other two species were highly fragmented, consisting of 338,980 to 3,346,358 scaffolds covering less than half of the estimated genome sizes.

The reasons for these different results could be related to three aspects: library design and coverage, genome assembly pipeline and/or biological basis. Certain genomes can be more difficult to assemble than others due to large quantities of repeats or strong GC bias. Indeed, the presence of repeats is hypothesized as the reason for assembly fragmentation in *C. maenas* and *N. denticulata*. However, repeat analysis of the *E. sinensis* draft genome indicated that 50 % of the assembled genome consisted of repeat structures thus repeat structures were not causing significant fragmentation in this assembly. As illustrated by results in Chapter 4, the choice of assembler influence the statistics applied to the resulting assemblies. The *C. maenas* and *N. denticulata* assemblies were attempted using several assemblers whereas the *E. sinensis* only listed the platanus assembler. It is always worth exploring different assemblers since simply applying the platanus assembler, as used in the *E. sinensis* study, to the *C. maenas* dataset did not produce the optimal results nor relieved fragmentation. Another lesson that can be learned relates to sequencing study design. The *E. sinensis* study had the largest range of insert sizes, seven in total, as compared to the five for *C. maenas* and the single library for *N. denticulata*. Additionally, the total amount of sequence generated was the largest for *E. sinensis*, although this choice is dependent on available budget. A larger quantity of information combined with a larger spread of insert sizes could thus provide the best opportunity for reducing fragmentation. Overall, it can be concluded that it is beneficial to design an Illumina technology based genome sequencing study with as many insert sizes as possible, to apply different assembly strategies and generate as much coverage as the budget allows. The developments of alternative sequencing technologies provide more opportunities to optimize study design. A combination of the longer reads offered by the PacBio platform with high coverage Illumina data and hybrid assembly methods could make improved assemblies at reasonable budgets more likely.

Table 1 Aquatic crustacean sequencing library designs

Species	Platform	Library insert sizes (bp)	Total sequence yield (Gb)	Estimated genome size (sequence coverage, given estimated size)
C. maenas	Illumina HiSeq 2500	300-500, 500-700, 3500-5624, 7361-11741	~ 135 Gb	1 Gb (135 x)
E. sinensis [3]	Illumina HiSeq 2000	170, 250, 500, 800, 2000, 5000, 10000	~ 258 Gb	1.66 Gb (155 x)
N. denticulata [4]	Illumina HiSeq 2000	167	~ 36 Gb	3 Gb (12 x)

Table 2 Aquatic crustacean draft genome assembly results

Species	Assembler (alternatives explored)	Total scaffolds	Total length (%)	N50 (bp)
C. maenas	SOAP-denovo2 + BESST (*SPAdes, platanus, ALLPATHS*)	338,980	0.36 Gb (36.0 %)	1,107
E. sinensis [3]	Platanus	17,553	1.12 Gb (67.5 %)	224,000
N. denticulata [4]	ABySS (*Velvet, SOAP-denovo*)	3,346,358	1.28 Gb (42.7 %)	400

Gene model

One of the major outputs of the *C. maenas* genome assembly is the gene model that contains the location of predicted genes and their respective exons/introns. The gene model was generated through a combination of *ab initio* predictions and predictions based on known sequence information. As mentioned earlier, sequence information for aquatic crustaceans is scarce which complicates this avenue for prediction. Also the *ab initio* predictors can be trained with prior information which propagates into better gene models. The *C. maenas* genome *ab initio* prediction benefitted from the RNA-sequencing data described in chapter 3 [5]. Should such a dataset not be available, as for the *E. sinensis* and *N. denticulata* draft genomes, training of predictors like AUGUSTUS was based on more distant species like *Homo sapiens*, *Crassostrea gigas*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Daphnia pulex*. Despite the availability of RNA-sequencing data and the combination of multiple gene prediction methods the resulting gene model is often still improved through manual examination. The characterization of the predicted Dscam gene on *C. maenas* scaffold 192 showed that the final gene model incorrectly split up the *C. maenas* Dscam gene into two parts. It is likely that such artefacts occur in other predicted genes as well. However, due to the volume of data, manual examination is often not tractable with limited financial and labour resources. Therefore it can be recommended that prior to using draft genomes, like those of *C. maenas*, *E. sinensis* and *N. denticulata*, a manual inspection of the scaffold assembly and gene model support is performed on loci of interest.

miRNAs

The absence of introns and the clear structure of their precursors facilitate the identification of miRNAs in the draft assemblies. Investigation of miRNA precursor structures was not performed for *E. sinensis* and *N. denticulata*, possibly due to not having a small RNA sequencing dataset available. Combination of small RNA sequencing data and the draft genome resulted in the prediction of 184 miRNA precursors, with 31 showing similarity to known miRNAs. Thus a significant number of novel miRNAs were identified in *C. maenas*. Considering that in miRbase (release 21) there are a total of 223 species represented (only 27 arthropods) the *C. maenas* miRNAs are of real value to the scientific community – particularly in the invertebrate clade. Since miRNAs fulfil an important regulatory role across many cellular processes their identification can aid in understanding these processes. However, a complete picture of the role if miRNAs would require knowledge on the mRNA targets that each miRNA can bind to in order to influence translation/degradation.

Contrary to the less complex precursor structures, the prediction of the gene targets of miRNA was difficult. The applied miRNA target prediction algorithms, MicroTar and miRanda, did not tend to agree on miRNA-mRNA target combinations. Because animal miRNAs tend to bind to the 3' or 5' UTR regions of mRNA, some miRNA target prediction methods like MicroTar require only 3' UTR sequences as input. The transcriptome for *C. maenas* is available, but precise determination of 3' UTR regions in the transcripts was not performed. A more detailed investigation of these, for example by identification of transcript regions after the stop-codon of the coding areas in the assembled transcripts, could lead to improvements of the miRNA target predictions. However even with improvements it is likely that miRNA target prediction remains capricious, as precision of target prediction algorithms can be around 50 % with a sensitivity ranging from 6 to 12 % [6]. It is likely that many of the miRNA targets will have to be identified through experimental means rather than rely on bioinformatic predictions.

Viral insertions

Rozenberg *et al.* 2015 showed the presence of WSSV ORFs in the Jamaican bromeliad crab (*Metopaulias depressus*) genome, as determined through Roche 454 sequencing [7]. Within their sequencing data, 186,890 reads with an average length of 265.5 bp, they found 10 % of reads coming from viral origins. In the *C. maenas* such evidence was not found. While there were reads with significant similarity to WSSV sequences, their quantity was low (0.00158 % of reads in the paired-end library with insert size 300-500). The integration events in *M. depressus* appear to have been facilitated through the action of transposable elements which provide the means for multiplication across the genome. Should a similar integration even have taken place in *C. maenas* the percentage of WSSV derived sequences would be higher. The small amount of sequence identified in the *C. maenas* data is more likely to derive from contamination of previous sequencing experiments with WSSV performed on the same machine. Nevertheless genome-walking experiments with primers based on *wsv514* (RefSeq: NP_478036.1), *wsv209* (RefSeq: NP_477731.1) or *wsv037* (RefSeq: NP_477559.1) could verify a possible integration event since these are WSSV ORFs with reads in *C. maenas* sequencing data.

The transcriptome assembly described in Chapter 3 is by itself a good fundamental resource for *C. maenas*. But placing this transcriptome in the context of the genome scaffold assembly enables a wider perspective that is greater than the sum of its parts. For example: knowledge on the introns/exons of a gene enables tracking of splice variation (e.g. Dscam transcript variations), regulatory elements in gene promoter regions sheds light on signalling circuitry and miRNA identification allows progress in studying their role. The availability of these datasets thus allows scientists to ask more detailed questions, design more powerful experiments with greater ease and facilitates interpretation of their results. This could potentially result in a snowball effect which propels *C. maenas* to become an important model organism for aquatic invertebrates/arthropods.

However there is room for significant improvement in the assembly and annotation. The assembly remains fragmented and annotation could benefit from manual interventions. The coverage of the *C. maenas* genome by Illumina sequencing reads ought to be sufficient for assembly; therefore additional sequencing with this technology should have little impact on improving the assembly. Should funds be available for additional sequencing, it is advisable that this is performed on a different platform. Sequencing technologies like PacBio and Oxford Nanopore can offer read lengths that extend far beyond those delivered by Illumina. It has been suggested that a 5-10x coverage of the *C. maenas* genome with PacBio data, which would amount to a cost of around £ 40,000 can significantly improve the genome scaffold via hybrid assembly strategies. The annotation is more difficult to improve. While a higher quality, less fragmented, assembly would achieve better *in silico* annotation, significant improvements would have to be derived either through manual inspections, novel algorithms or incorporating additional data sources on *C. maenas* genes.

7.3 Homarus gammarus transcriptome assembly

The European lobster (*H. gammarus*) is an important aquatic crustacean species in Europe and the United Kingdom in particular. Research efforts in these species focus on topics like conservation, stock enhancement, improvement of husbandry techniques, impacts of environmental threats and diseases. Many of these studies can benefit from the molecular platform offered by the assembled *H. gammarus* transcriptome. For example, identification of genetic components linked to growth and disease resistance of the lobster could be meaningful in optimizing stocking programmes and potentially enhancing their aquaculture. In this thesis we aimed at the immune system, characterization of which can aid in the study of disease like WSD and allows for comparisons to results from *C. maenas*.

Assembly and annotation

Many of the previous comments made regarding the *C. maenas* transcriptome are also applicable to the *H. gammarus* transcriptome. While data generation and assembly are feasible with current technologies, the annotation of the transcriptome remains difficult. Transcriptome annotation relying on sequence similarities has a low rate of success, even when thresholds are lowered. However, availability of the transcriptome does facilitate wet lab experiments that can aid in elucidating the function of transcripts, thus widening the knowledgebase for annotation. The identification and comparison of relevant biological pathways in *H. gammarus* and *C. maenas* depended partially on the KEGG database. As mentioned in the text, KEGG is focused on mammalian/vertebrate species which makes it less viable in invertebrates like crabs and lobsters. It would be beneficial for all studies conducted in invertebrate species to have an equivalent resource to KEGG available to facilitate pathway visualization and analysis. Given that currently there are large invertebrate genome sequencing projects underway, e.g. the i5k projects that aims to sequence 5000 invertebrate genomes (including aquatic crustacea) [8] and that analysis of all this data would benefit from such a resource, it could be possible for an invertebrate KEGG to be developed.

Through comparing transcript expression in *H. gammarus* samples, profiles of differentially expressed transcripts for nine lobster tissues were generated. Amongst these profiles the male gonad tissue stood out in its quantity of differentially expressed transcripts. Further analysis showed that for those transcripts with functional annotation, there was enrichment of terms not typically associated with testes. Cell differentiation of various types like osteoBLASTs, mesenchymal cells, epidermal cells, keratinocytes, astrocytes and more were identified but not spermatozoa. In addition, terms related to higher rates of cell division that accompany production of spermatozoa were also not identified. Underlying reasons for the large count of differentially expressed transcripts observed and their functional annotation could involve: previously discussed difficulties of transcript annotation, presence of complex parasites, other invading agents in the sampled testes or even mislabelling of samples during tissue isolation and subsequent RNA extraction (e.g. whole body juveniles labelled as male gonad tissue).

Comparison of the immune systems of *H. gammarus* and *C. maenas* showed many overlapping components. Most components could be identified in both species. Notable exceptions were: peptidoglycan recognition protein (present in *H. gammarus* but not in *C. maenas*), Dredd (present in *C. maenas* but not in *H. gammarus*) and Dicer2 (present in *C. maenas* but not in *H. gammarus*). The first two could have implications on how both organisms deal with pathogens of bacterial origin whereas the latter can be related to antiviral defence. In invertebrates like *Drosophila* Dicer2 is important in the siRNA component of the RNAi system. siRNAs can be created from virus derived dsRNA which in combination with the RISC complex promote degradation of complementary mRNAs. WSSV is a dsDNA virus, but because transcription can take place in both sense and antisense directions from its genome dsRNA molecules can be formed [9]. Since Dicer2 is lacking in *H. gammarus*, the efficacy of its siRNA could be compromised as Dicer1 is less efficient in the siRNA-RISC complex [10]. Experiments involving introduction of *C. maenas* Dicer2 in *H. gammarus* or silencing of Dicer2 in *C. maenas*, both in combination with WSSV exposure, could reveal the impact of these differences in the immune system WSD development.

For many viruses the success of infection is largely dependent on the process of viral entry [11]. Prior to viral entry via endocytosis WSSV bind to proteins present on the host cell surface, the efficiency of which is determinant of whether infection will be successful. Furthermore, the target host proteins are important determining factors for the range of species a virus can infect. Often viruses have interactions with conserved proteins which enables a greater host range and potential host jumping [12]. In the 1990s WSSV was identified as a virus that mainly impacts penaeid shrimp. Therefore the more closely a cellular environment matches that in penaeid shrimp, the more likely it is that WSSV can successfully infect that host [12]. According to this theory the sequences of three WSSV receptors in the transcriptomes of *H. gammarus* and *C. maenas* were compared to those of penaeid shrimp species. One of the major interactions between WSSV and host proteins is the interaction between WSSV VP28 and host Rab7 [13]. For all the WSSV receptors (integrin, lectin and Rab7) the *H. gammarus* sequences were closer to those of penaeid shrimp than the *C. maenas* sequences. This may offer a partial explanation as to why *C. maenas* are less susceptible to the virus compared to *H. gammarus*.

7.4 *Carcinus maenas* response to WSSV infection

For the study on the response of *C. maenas* to WSSV, the virus was introduced to *C. maenas* through injection and animals were subsequently kept for a period of up to 28 days. Over this period *C. maenas* individuals were sampled at regular timepoints. The response was measured in the gill tissue, which is relatively easy to dissect from an individual crab and tends to be one of the tissues where the virus first appears and replicates [14].

After sequencing it was found that only five (out of 28) samples contained RNA derived from WSSV. The presence of WSSV RNA indicates that the gill tissues of those crabs contained cells wherein WSSV was expressing its genes and replicating. It would be interesting to compare this result to those of certified WSSV PCR detection kits used on the same samples. There should be differences in results since the PCR kits will pick up presence of virus, even if it is not replicating. This would give an indication on how long the virus remains present in the *C. maenas* after injection and the fraction of individuals that develop infection over time. Additionally, histology data would provide extra depth of insight in the fate of WSSV in *C. maenas*. To observe what cellular compartments the virions are present, e.g. the nucleus, endosomes and lysosomes, would be highly informative. Thus together histology, PCR and RNA-sequencing can describe the whole infection process from confirming virus presence to entry, localization and ultimately replication. Collectively this would help to provide additional depth and understanding of the infection process.

As mentioned, replicating WSSV was detected in 5 out of 28 *C. maenas* individuals injected with WSSV. Infected individuals were spread over three timepoints: two at 7 days post injection, one at 14 days post injection and another two at 28 days post injection. Since there were four individuals sampled at every timepoint it was clear there was individual variation in relation to infection success of WSSV. This intraspecies variation could provide greater detail to understanding *C. maenas* resistance. The exposure experiment in Chapter 6 sampled different individuals at every timepoint. Elucidating the individual variation in response to the virus requires repeated sampling if the same individual over the set time period. Sampling gill tissue multiple times is intractable and instead only sampling of haemocytes can be considered. Studies of WSSV in crustacean haemocytes have been performed [15, 16]; however even repeated blood sampling might incur significant stress on individuals that might act as a confounding factor in the study.

A further point of note is that the infected individuals were all sampled at the later timepoints; 7-28 days post infection. Thus there is a slower speed of infection compared with in penaeid shrimp, where infection takes places within 12-48 hours [17]. This slower rate confirmed the previously established recalcitrance *C. maenas* has towards WSSV infection [1]. One of the main objectives of this thesis was to identify the (molecular) basis of this apparent resistance. One explanation for this has already been provided through the comparison of sequences of the identified WSSV receptors. Sequence variation could cause the interaction between host receptor and viral proteins to lose stability, slowing down the viral entry process and thus only infections at later timepoints are observed. However, gene and miRNA expression measurements showed there are other interesting interactions which are considered later in this discussion.

While genomes of four WSSV isolates have been sequenced, a large fraction of their annotation is based on ORF predictions. The availability of RNA-sequencing data for WSSV provided an opportunity to improve annotation by determining which transcripts were expressed. To this end a WSSV transcriptome was assembled. Application of both genome-guided and *de novo* assemblers resulted in assemblies containing long transcripts with multiple ORFs. It appeared that the assemblies were capricious in regard to the used settings in the assemblies (see Chapter 6). Assembly was complicated by two factors. Firstly, the genome of WSSV is densely packed with ORFs which was confirmed by the fact that *de facto* the whole genome was covered by RNA-sequencing reads. This results in some assemblers like cufflinks to generate a single transcript from the RNA sequencing data. The second complication is the presence of clusters of WSSV genes that are represented by multiple polycistronic mRNAs. Having many individual regions of the genome being represented by multiple distinct transcripts is a situation that many assemblers apparently struggle with. As was shown for the VP60/wsv419/wsv420/VP28 cluster, the monocistronic mRNA for VP28 was not represented and RNA likely originating from this transcript was ended up as part of a polycistronic mRNA. The difficulty in WSSV transcriptome assembly points to a need for a novel assembler that is able to deal with data of this nature. As was shown, adjustment of parameters for assemblies did not yield consistent results across the whole transcriptome. Perhaps assemblers have to estimate the location of potential clusters and adjust parameters within said clusters to derive better results.

The transcriptomic response in *C. maenas* to WSSV was determined by comparing injected with control samples at each timepoint of the infection study. The first observation was that overall there was only a limited change in the expression of immune system components of the *C. maenas* immune system. Increased expression was observed for some pattern recognizing and response proteins. This stands in contrast to studies in penaeid shrimp species wherein significant upregulation of the immune system is often observed. Paradoxically a limited response could be in favour of *C. maenas* because WSSV hijacks transcription factors that are activated by the immune system. The absence of a transcriptomic response of the immune system could thus be part of the molecular basis of *C. maenas* resistance to WSSV infection.

Expression analyses at the 12h timepoint indicated significant changes to expression of genes involved in the endocytosis pathway. It appeared that the flow of the endocytosis system is changed due to lower expression of several of its key regulators, the Rab GTPases. Firstly, there would be reduced uptake of cargo via Clathrin-mediated endocytosis which reduces the ability of WSSV to enter the cell. This is the first barrier of defence for *C. maenas* against WSSV infection. Should WSSV still manage to enter endocytic vesicles it faces the next barrier: the early endosome. Like many viruses, WSSV probably requires a pH trigger in order to escape the endosomes. The lower pH in maturing/late endosomes is thought to be such trigger and indeed WSSV proteins interact with one of the key regulators of endosome maturation: Rab7. It appears that in *C. maenas* the process of endosome maturation is slowed down due to lower expression of Rab7 which cause WSSV particles to remain in early endosomes. The slow recycling transport mechanism, which flows through the endosomal recycling complex, is also reduced while the fast recycling pathway remains at equal levels. All these transport mechanisms combined would result in accumulation of WSSV in early endosomes and increased recycling to the cellular surface, thereby hindering successful viral infection. Combining the fact that this change was only observed at the 12 hour timepoint with the low infection rate of WSSV lead us to hypothesise that passing through *C. maenas* early endosomes may influence subsequent pathogenicity of the virions. While similar defence mechanisms have been observed in other virus/host interactions, it has to be noted that the described hypothesis is purely based on observed changes in gene expression [18]. Endocytosis is a complex process, the complexities of which might not be fully captured in expression data alone. Therefore, it is recommended that an attempt to observe accumulation of WSSV in early endosomes and their recycling to the cell surface through microscopy would be useful to clarify if this hypothesis explains, at least to some extent, the resistance to WSSV in the crab. Studying the effects of endocytosis mediating chemical agents and knockdown/overexpression of important regulators like Rab5, Rab7, Rab11 or Phosphoinositides kinases on WSSV infection in *C. maenas* would also be good steps in verifying this hypothesis. In addition it would also be of significant interest to identify the signal through which these changes are initiated and the regulators that bring them to fruition.

MicroRNAs could represent one class of such regulators. Over the course of the exposure experiment, components of the RNAi system (e.g. ago2) were upregulated, indicating an increased activity of miRNA mediated expression regulation. Small RNA sequencing allowed detection of differentially expressed miRNAs. Across most timepoints there was not a lot of coherence, apart from cma_pmiR-12 which showed significant downregulation in WSSV exposed samples. There is thus a contrast between cma_pmiR-12 downregulation and the overall increase of RNAi activity. Nevertheless, this miRNA is thus of major interest in regards to WSSV infection in *C. maenas*. To determine the function of cma_pmiR-12 its targets were predicted using two software packages: miRanda and MicroTar. Analysis of results showed a low level of agreement between the two packages, indicating that *in silico* miRNA-target prediction is not consistent. A factor that has been indicated previously [19]. Targets predicted by miRanda showed enrichment for genes involved in the endomembrane system and this could thus be one of the regulators of the *C. maenas* endocytosis-mediated resistance to WSSV infection. Confirmation through proteomics experiments following silencing of cma_pmiR-12, e.g. by using antagomirs [20], might reveal which mRNAs' translation is inhibited by this miRNA. Should cma_pmiR-12 indeed be directly involved in *C. maenas* resistance to WSSV, the next step would be to investigate whether this knowledge can be applied to commercially important species like penaeid shrimp. If shrimp such as *L. vannamei* also produce a miRNA that is similar to cma_pmiR-12 then silencing experiments are a good start. If not, the focus should shift on the targets of cma_pmiR-12.

After examining the data, several hypotheses on why *C. maenas* is relatively resistant to WSSV infection can be formulated. First, *C. maenas* viral receptors deviate from those in susceptible shrimp species. Secondly, the immune system of *C. maenas* does not undergo large changes in gene expression which could be beneficial for the host. Thirdly, at 12 hours post injection there are significant changes in the *C. maenas* endocytosis and RNAi systems that could disrupt viral entry. Finally, the novel miRNA cma_pmiR-12 may be involved in mediating *C. maenas* response to WSSV exposure, but thus requires further investigation to provide more strength to this hypothesis. Applications of these hypotheses to other aquatic crustacean species can only be done for the third and fourth hypotheses. Sequences of viral receptors and the immune system are difficult to change without genetic modification. Influencing host endocytosis and cma_pmiR-12 can be applied more readily in a form of treatment. For example, it has already been shown that injection of *P. monodon* Rab7 or antibody for *P. monodon* Rab7 could increase the survival rate upon WSSV challenge [13]. As mentioned, treatments based on cma_pmiR-12 would depend on whether a similar miRNA is present in the species of interest or alternatively design treatments that act on the targets of cma_pmiR-12.

7.5 Results in the wider context

The work presented in this thesis had several practical outputs. The second chapter summarized current knowledge on the WSSV infection process which is beneficial to anyone working in, or being introduced to, this field. We generated transcriptomes for *C. maenas* and *H. gammarus* and a draft genome assembly for *C. maenas.* Taken together, this body of work is a valuable resource for future studies involving these species as has been elaborated earlier in this discussion.

In this work, one particular component of the immune system, *C. maenas* Dscam, received special attention as there are suggestions of its link to pancrustacean immune memory. A gene model for *Dscam* was identified on a scaffold in the *C. maenas* genome. The gene model recovered the latter domains of the gene whereas the earlier domains were not present on the scaffold. Sequencing based on gene walking is thus necessary to extend the scaffold in the 5' direction of the *C. maenas* scaffold. An understanding of immune memory in *C. maenas* and penaeid shrimp could be valuable to the crustacean aquaculture sector. Work by Ng *et al.* has already shown that WSSV-induced Dscam from the Australian freshwater crayfish (*Cherax quadricarinatus*) could neutralize the virus and sustain the host during persistent infection (> 1 month) [21]. Identifying whether a similar form of Dscam can be induced in shrimp would be of high importance to penaeid shrimp aquaculture. Given that farmed shrimp only have to survive 3 – 6 months in order to reach marketable size, any protective effect from Dscam can aid in achieving that biomass without succumbing to WSSV infection. It would be an asset if production of WSSV-neutralizing Dscam could be initiated prior to placement in ponds and through addition of e.g. neutralized virus as is done in vertebrate vaccinations. As is the case for cattle and other farmed animals, vaccinations that prevent diseases are preferable to drugs that cure a disease. Therefore, exploring whether Dscam-associated 'vaccination' is a plausible route for crustacean aquaculture would be preferable to deploying resources to develop drugs that cure diseases like WSD.

The *C. maenas* – WSSV exposure study yielded insights on how a relatively resistant crustacean responds to the pathogen. As described, RNA- and miRNA sequencing data analysis pointed to changes in endocytosis and miRNA expression. Both of these provide angles of research into WSD. Firstly, it can be explored whether temporarily changing endocytosis via small molecules or gene knockdown has a significant effect on WSSV infection and replication efficiency. If this be the case, endocytosis-related targets can be explored for the development of pharmaceuticals. However, endocytosis is an important cellular process and interfering with its regulation can result in significant side effects. In Drugbank, a database for human drugs and their targets, there are no pharmaceuticals that specifically target the Rab family of endocytosis regulators (accessed in June 2016) [22]. Given that many human diseases can be linked to endocytosis (e.g. Griscelli syndrome), mental retardation, neuropathy (Charcot–Marie–Tooth), kidney disease (tuberous sclerosis), and blindness (choroideremia), yet the fact that no endocytosis targeting drugs exist might be circumstantial evidence that they are impractical targets [23].

The RNAi pathway is a well-studied antiviral pathway in invertebrates and tackling WSSV via this route might be easier than interfering with endocytosis. The identification of a *C. maenas* miRNA that is downregulated in every WSSV exposed sample, cma_pmiR-12, provides a great opportunity for potential treatment development and should be given preference. Firstly, it might be insightful to identify whether or not penaeid shrimp species expressed a similar miRNA to cma_pmiR-12. If this is the case, testing the effects of reduced expression on WSSV pathogenicity can be measured. If shrimp do not possess a cma_pmiR-12 focus should shift to its targets. As discussed previously, target identification for miRNAs *in silico* is problematic, thus laboratory experiments are a definite requirement. Identified cma-pmiR-12 targets can be investigated for potential therapeutic development.

While the insights gained on *C. maenas*-WSSV interactions are interesting and yielded angles for drug target discovery, it should be questioned whether new drug targets should be the top priority for WSSV-related research. The current body of knowledge on WSSV infection provides many potential pathways/genes/miRNAs for treatment development. And indeed several treatments have been shown to increase shrimp survival after WSSV exposure e.g. injection of plasmids that enable expression of long hairpin RNAs that target viral genes like *vp19* and *vp28*. To date, however, the treatments that work in laboratory have not found application in the field. The main hurdle is the lack of an effective delivery method. Delivery of small molecules to shrimps in ponds is complex, since they would have to be administered either through the food or water. Laboratory treatments for WSSV are often delivered via injection, which is impractical in an aquaculture context. In order to combat future epidemics as well as the current threat of WSSV it would be important to develop a cost effective system that can deliver drugs (be it small molecule, miRNA or protein based) to shrimps in ponds. An example of a realistic method of delivery and treatment is through food. Valdez *et al.* showed that oral administration of *B. subtilis* spores expressing VP26 on their surface was shown to convey a 100 % protection against WSSV in *L. vannamei* [24]. It is claimed that this method of treatment is easy to produce, practical to handle, environmentally stable, human-safe and economically feasible. Furthermore, the system could be adapted by changing the presented protein to one that is relevant to the disease to combat. Alternatively, one could consider to genetically modify shrimp to make them resistant to WSSV infection, e.g. through expression of small RNA molecules targeted to viral genes like VP28. However, applicability of this method depends on global attitudes toward GMO food sources. Thus development of delivery methods should be given a higher priority than additional target discovery and optimization. Furthermore, once a delivery method is developed there is potential for it to be applied to other diseases in addition to WSSV specifically.

While WSSV has been the most devastating viral epidemic in crustacean aquaculture thus far, it is likely that other diseases will emerge. The recent outbreak of Early Mortality Syndrome in South East Asia is such an example. Should these new diseases develop into global epidemics, resulting in significant losses of food and resources, again methods for vaccination and/or treatment have to be developed. Such studies would probably follow a similar pattern as those for WSSV research: identify the pathogen, the major host pathogen interactions and identification of potential targets for treatment. To this date, most of the work in aquatic crustacea has been performed on whole organisms. For human disease studies, most of the experimental work has moved toward immortalized cell lines platforms since experimenting on whole organisms is time consuming, costly and ethically undesirable. Cell lines are available for insects, and indeed some studies referenced in this thesis were conducted on Sf9 insect cells. However to date there are no lines available for aquatic crustacea like penaeid shrimp. Jayesh *et al.* 2012 [25] described problems with cell line development, and why it has been unsuccessful to date. Despite of 25 years of ongoing attempts to develop shrimp cell lines, a continued effort is highly desirable given the multibillion dollar size of the shrimp aquaculture industry and its importance in supplying the world with a source of protein. An additional benefit is that not only pathogen-related studies can be facilitated through cell lines and research in aquatic crustaceans can be sped up resulting in increased understanding of these invertebrates. Thus datasets like the *C. maenas* transcriptome and exposure can be interpreted with greater detail.

It is likely that development of antiviral vaccinations, delivery methods and treatments for shrimp diseases will take more time. Once developed there will be the considerable challenge of actually implementing said treatment in an efficient manner. This can be due to production and distribution limitations, potential costs of treatments and education how and when to administer it. With these problems, one should accept that WSSV and other (viral) diseases will be endemic in the aquaculture industry in the future. In such an environment it is vital to have an understanding on how diseases emerge and develop into epidemics and to identify whether a disease is affected by conditions in the ponds – preferably those that can be controlled. Establishing good guidelines based on such real world data might be equally important as the treatment development in itself.

To conclude, the work presented in this thesis delivers a significant contribution the genomic resources available for aquatic crustacea. In addition we identified mechanisms behind the resistance of *C. maenas* to WSSV infection and identified potential targets for treatment development. Endocytosis regulators, the cma_pmiR-12 miRNA and its targets are angles that can be explored in future experiments. Current understanding of WSSV is at a good level and most importantly a series of targets is available to probe for intervention strategies. What is lacking, however, is field/commercial scale application of this knowledge. Therefore although further molecular studies will likely help to better identify targets for possible optimised treatment interventions, it might be prudent also to investigate the potential of (Dscam related) vaccination, development of effective treatment delivery methods and development of penaeid shrimp cell lines in the research route(s) to help combat WSD.

7.6 References

1. Bateman KS, Tew I, French C, Hicks RJ, Martin P, Munro J, Stentiford GD: **Susceptibility to infection and pathogenicity of White Spot Disease (WSD) in non-model crustacean host taxa from temperate regions**. *Journal of Invertebrate Pathology* 2012, **110**(3):340-351.
2. Schurch NJ, Schofield P, Gierlinski M, Cole C, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson GG, Owen-Hughes T *et al*: **How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?** *RNA* 2016, **22**(6):839-851.
3. Song L, Bian C, Luo Y, Wang L, You X, Li J, Qiu Y, Ma X, Zhu Z, Ma L *et al*: **Draft genome of the Chinese mitten crab, *Eriocheir sinensis***. *Gigascience* 2016, **5**:5.
4. Kenny NJ, Sin YW, Shen X, Zhe Q, Wang W, Chan TF, Tobe SS, Shimeld SM, Chu KH, Hui JH: **Genomic sequence and experimental tractability of a new decapod shrimp model, *Neocaridina denticulata***. *Marine Drugs* 2014, **12**(3):1419-1437.
5. Verbruggen B, Bickley LK, Santos EM, Tyler CR, Stentiford GD, Bateman KS, van Aerle R: ***De novo* assembly of the *Carcinus maenas* transcriptome and characterization of innate immune system pathways**. *BMC Genomics* 2015, **16**:458.
6. Min H, Yoon S: **Got target? Computational methods for microRNA target prediction and their extension**. *Experimental & Molecular Medicine* 2010, **42**(4):233-244.
7. Rozenberg A, Brand P, Rivera N, Leese F, Schubart CD: **Characterization of fossilized relatives of the White Spot Syndrome Virus in genomes of decapod crustaceans**. *BMC Evolutionary Biology* 2015, **15**:142.
8. i KC: **The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment**. *Journal of Heredity* 2013, **104**(5):595-600.
9. Umbach JL, Cullen BR: **The role of RNAi and microRNAs in animal virus replication and antiviral immunity**. *Genes & Development* 2009, **23**(10):1151-1164.
10. Lee YS, Nakahara K, Pham JW, Kim K, He Z, Sontheimer EJ, Carthew RW: **Distinct Roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA Silencing Pathways**. *Cell* 2004, **117**(1):69-81.

11. Mercer J, Schelhaas M, Helenius A: **Virus entry by endocytosis**. *Annual Review of Biochemistry* 2010, **79**:803-833.

12. Longdon B, Brockhurst MA, Russell CA, Welch JJ, Jiggins FM: **The Evolution and Genetics of Virus Host Shifts**. *PLoS Pathogens* 2014, **10**(11):e1004395.

13. Sritunyalucksana K, Wannapapho W, Lo CF, Flegel TW: **PmRab7 is a VP28-binding protein involved in white spot syndrome virus infection in shrimp**. *Journal of Virology* 2006, **80**(21):10734-10742.

14. Wu W, Wang L, Zhang X: **Identification of white spot syndrome virus (WSSV) envelope proteins involved in shrimp infection.** *Virology* 2005, **332**:578-583.

15. Huang Z-J, Kang S-T, Leu J-H, Chen L-L: **Endocytic pathway is indicated for white spot syndrome virus (WSSV) entry in shrimp**. *Fish & Shellfish Immunology* 2013, **35**(3):707-715.

16. Jiravanichpaisal P, Sricharoen S, Söderhäll I, Söderhäll K: **White spot syndrome virus (WSSV) interaction with crayfish haemocytes**. *Fish & Shellfish Immunology* 2006, **20**(5):718-727.

17. Sanchez-Paz A: **White spot syndrome virus: an overview on an emergent concern**. *Veterinary Research* 2010, **41**(6):43.

18. Rizopoulos Z, Balistreri G, Kilcher S, Martin CK, Syedbasha M, Helenius A, Mercer J: **Vaccinia Virus Infection Requires Maturation of Macropinosomes**. *Traffic* 2015, **16**(8):814-831.

19. Ekimler S, Sahin K: **Computational Methods for MicroRNA Target Prediction**. *Genes* 2014, **5**(3):671-683.

20. Krutzfeldt J, Rajewsky N, Braich R, Rajeev KG, Tuschl T, Manoharan M, Stoffel M: **Silencing of microRNAs in vivo with 'antagomirs'**. *Nature* 2005, **438**(7068):685-689.

21. Ng TH, Hung H-Y, Chiang Y-A, Lin J-H, Chen Y-N, Chuang Y-C, Wang H-C: **WSSV-induced crayfish Dscam shows durable immune behavior**. *Fish & Shellfish Immunology* 2014, **40**(1):78-90.

22. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J: **DrugBank: a comprehensive resource for in silico drug discovery and exploration**. *Nucleic Acids Research* 2006, **34**(Database issue):D668-D672.

23. Stein M-P, Dong J, Wandinger-Ness A: **Rab proteins and endocytic trafficking: potential targets for therapeutic intervention**. *Advanced Drug Delivery Reviews* 2003, **55**(11):1421-1437.

24. Valdez A, Yepiz-Plascencia G, Ricca E, Olmos J: **First *Litopenaeus vannamei* WSSV 100% oral vaccination protection using CotC::Vp26 fusion protein displayed on Bacillus subtilis spores surface**. *Journal of Applied Microbiology* 2014, **117**(2):347-357.

25. Jayesh P, Seena J, Singh IS: **Establishment of shrimp cell lines: perception and orientation**. *Indian Journal of Virology : an official organ of Indian Virological Society* 2012, **23**(2):244-251.