University of Rhode Island

DigitalCommons@URI

Open Access Master's Theses

2013

Transcriptome of the Eastern Oyster *Crassostrea Virginica* in Response to Bacterial Challenge

Ian McDowell

University of Rhode Island, ian.mcdowell@gmail.com

Follow this and additional works at: <https://digitalcommons.uri.edu/theses>

Recommended Citation

McDowell, Ian, "Transcriptome of the Eastern Oyster *Crassostrea Virginica* in Response to Bacterial Challenge" (2013). *Open Access Master's Theses*. Paper 19.
<https://digitalcommons.uri.edu/theses/19>

This Thesis is brought to you for free and open access by DigitalCommons@URI. It has been accepted for inclusion in Open Access Master's Theses by an authorized administrator of DigitalCommons@URI. For more information, please contact digitalcommons-group@uri.edu.

TRANSCRIPTOME OF THE EASTERN OYSTER CRASSOSTREA
VIRGINICA IN RESPONSE TO BACTERIAL CHALLENGE

BY

IAN MCDOWELL

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

BIOLOGICAL AND ENVIRONMENTAL SCIENCES

UNIVERSITY OF RHODE ISLAND

2013

MASTER OF SCIENCE THESIS
OF
IAN MCDOWELL

APPROVED:

Thesis Committee:

Major Professor Marta Gomez-Chiarri

Chris E. Lane

Matt P. Hare

Jason J. Kolbe

Nasser H. Zawia

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

2013

ABSTRACT

The Eastern oyster *Crassostrea virginica*, an ecologically and economically important estuarine organism, suffers mortalities as high as 90-100% in affected areas due to Roseovarius Oyster Disease (ROD), caused by the bacterial pathogen, *Roseovarius crassostreae*. Advanced genotypic breeding techniques necessitate information regarding markers and genes associated with disease resistance. As yet, the host-pathogen interaction between *C. virginica* and *R. crassostreae* is poorly understood at the molecular level. The identification of potential genes and pathways responsible for an effective host defense response in the Eastern oyster to *R. crassostreae* is important not only to provide a basis for enhanced breeding techniques, but also to enhance understanding of innate immunity in a broader, evolutionary sense. The present study proposed to uncover not only genes and general processes potentially involved in disease resistance to ROD in the Eastern oyster, but also diversified gene families. To that end the present study entailed a disease challenge exposing ROD-resistant and ROD-susceptible families of oysters to *R. crassostreae*, high-throughput cDNA sequencing of samples from several timepoints throughout the disease challenge, assembly of sequence data into a reference transcriptome, analysis of the transcriptome through differential gene expression and gene family similarity clustering, and single nucleotide polymorphism (SNP) detection to identify candidate gene markers. Oyster resistance to *R. crassostreae* was found to involve extracellular matrix

remodeling, cell adhesion, inflammation, metabolism, and other processes. Several gene families identified as putatively diversified and important in the oyster host defense response were enumerated and described, including serine proteases, serine protease inhibitors, c-type lectins, C1q domain-containing proteins, fibrinogen domain-containing proteins, scavenger receptors with class B SRCR domains, interferon-induced protein 44 (IFI44) family proteins, and GTPase of the immunity associated protein (GIMAP) family proteins. Further, similarity clustering of proteins and translated transcripts from diverse invertebrates suggested that GIMAP proteins are expanded in molluscs and IFI44 proteins are expanded in bivalves.

ACKNOWLEDGEMENTS

I thank my major professor, Marta Gomez-Chiarri, for her guidance, patience and expertise. I thank all of my labmates who have helped with this project and have helped to guide me in the lab including Chamilani Nikapitiya, Jess Piesz, Murni Karim, and Sae Bohm Sohn. I thank Ian Misner and Chris Lane who helped me greatly on the computational aspects of this project. I also thank the staff at Brown University's Center for Computing and Visualization, especially Lingsheng Dong and Mark Howison, for their timely and immensely helpful support. I thank my family, friends, and very supportive girlfriend, Stacey McMillin.

TABLE OF CONTENTS

ABSTRACT	ii
ACKNOWLEDGEMENTS	iv
TABLE OF CONTENTS.....	v
LIST OF TABLES	vi
LIST OF FIGURES	viii
INTRODUCTION	1
METHODS.....	7
RESULTS	25
DISCUSSION	40
TABLES.....	59
FIGURES.....	85
APPENDIX	97
BIBLIOGRAPHY	103

LIST OF TABLES

Table 1. Assembly metrics for transcriptome assembly.....	59
Table 2. Top 50 most highly differentially expressed genes (DEGs) unique to GX_early up-regulated relative to control pool (CGX_late).....	60
Table 3. Differentially expressed genes (DEGs) unique to GX_early down-regulated relative to control pool.....	63
Table 4. Top 50 most highly differentially expressed genes (DEGs) among GX_early shared with F3L_early and/or F3L_late up-regulated relative to control pool	66
Table 5. Top 50 most highly differentially expressed genes (DEGs) among GX_early DEGs shared with F3L_early and/or F3L_late down-regulated relative to control pool	69
Table 6. Pfam families enriched among the differentially expressed genes (DEGs) from resistant (GX) and susceptible (F3L) oysters at early (1 and 5 d) and late (15 and 30 d) timepoints compared to control (CGX) oysters, with rank order of significance of enrichment and significance of enrichment (p-value).....	72
Table 7. Pfam domains enriched among the differentially expressed genes	

(DEGs) from resistant (GX) and susceptible (F3L) oysters at early (1 and 5 d) and late (15 and 30 d) timepoints compared to control (CGX) oysters, with rank order of significance of enrichment and significance of enrichment (p-value).....74

Table 8. TribeMCL clusters enriched for differentially expressed genes (DEGs) from resistant (GX) and susceptible (F3L) oysters at early (1 and 5 d) and late (15 and 30 d) timepoints, with rank order of significance of enrichment and significance of enrichment (p-value).....76

Table 9. Comparison of select “diversified” groups of genes in the Eastern oyster to the Pacific oyster *C.gigas* genome and differential expression.....84

LIST OF FIGURES

Figure 1. Cumulative percent mortality in two families of oysters challenged with the bacterial pathogen <i>Roseovarius crassostreae</i> (F3L and GX) and in unchallenged controls (CF3L and CGX).....	85
Figure 2. Heatmap of all differentially expressed genes (DEGs) in two oyster families experimentally challenged with the bacterial pathogen <i>Roseovarius crassostreae</i> (GX and F3L) and unchallenged controls.....	86
Figure 3. Numbers of differentially expressed contigs (DEGs) shared and unique between GX_early (resistant oysters – days 1 and 5), F3L_early (susceptible oysters – days 1 and 5), and F3L_late (susceptible oysters – days 15 and 30).....	87
Figure 4. Functionally enriched Gene Ontology terms among DEGs from resistant (GX) and susceptible (F3L) oysters at early (1 and 5 d) and late (15 and 30 d) time points compared to control (CGX) oysters mapped by semantic similarity using SimRel method and REVIGO.....	88
Figure 5. Numbers of non-redundant transcripts/proteins and gene models in selected putative diversified gene families in multiple organisms from diverse taxa	90

Figure 6. <i>C. virginica</i> GIMAP translated transcripts clustered by similarity with color and size reflective of differential expression.....	91
Figure 7. <i>C. virginica</i> IFI44 translated transcripts clustered by similarity with color and size reflective of differential expression.....	92
Figure 8. Maximum likelihood phylogenetic tree of non-redundant GTPase of the immunity associated protein (GIMAP) transcripts/proteins from multiple organisms with leaves colored by organism	93
Figure 9. Maximum likelihood phylogenetic tree of non-redundant GTPase of the immunity associated protein (GIMAP) transcripts/proteins from multiple organisms with major groupings highlighted and summarized.....	94
Figure 10. Maximum likelihood phylogenetic tree of non-redundant interferon-induced protein 44 (IFI44) transcripts/proteins from multiple organisms with leaves colored by organism	95
Figure 11. Maximum likelihood phylogenetic tree of non-redundant interferon-induced protein 44 (IFI44) transcripts/proteins from multiple organisms with major groupings highlighted and summarized	96

INTRODUCTION

The Eastern oyster *Crassostrea virginica* is an ecologically and economically important estuarine organism cultured from Louisiana, USA to New Brunswick, Canada (Kennedy et al. 1996). Oyster production is an important sector of United States agriculture and the Eastern oyster was estimated in 2010 to have a farm gate value of \$90.4 million in the United States (NMFS 2010). Several oyster diseases, both protozoan and bacterial, have expanded in range and increased in severity during the latter half of the twentieth century, often causing staggering losses (Cook et al. 1998; Ford and Chintala 1996; Ford and Smolowitz 2007). Roseovarius Oyster Disease (ROD), caused by the bacterium *Roseovarius crassostreae*, went unreported before 1988 and presently affects oysters from the Long Island Sound north to Maine (Bricelj et al. 1992; Ford and Borrero 2001; Boettcher et al. 2005; Markey and Gomez-Chiarri 2009). ROD can cause mortalities as high as 90-100% in affected areas (Bricelj et al. 1992; Ford and Borrero 2001; Boettcher et al. 2005). ROD can cause mortality events that last a few weeks, often coinciding with or following closely upon peak summer temperatures. Gross clinical signs include uneven shell margins, soft tissue emaciation, and conchiolin depositions on the inner shell surfaces (Bricelj et al. 1992; Barber et al. 1998; Boettcher et al. 1999; Ford and Borrero 2001).

The host-pathogen interaction between *C. virginica* and *R. crassostreae* is poorly understood. It is known that *R. crassostreae* colonizes the oyster's

inner shell surface before lesions develop in the epithelial mantle (Boardman et al. 2008). Shell colonization may enable *R. crassostreae* to evade cell-mediated killing by hemocytes (Boardman et al. 2008), which are the active cells of the immune system present in the hemolymph and involved in triggering and sustaining both cell-mediated and humoral host defense, wound and shell repair, and other processes (Ford and Tripp 1996). Colonization likely stimulates the oyster to deposit conchiolin and it has been suggested that smaller oysters succumb to ROD because they lack adequate metabolic resources to fuel deposition, leading to emaciation (Bricelj et al. 1992; Ford and Borrero 2001; Boardman et al. 2008). *R. crassostreae* may produce a toxin with ciliostatic activity (Boettcher et al. 2000) and extracellular products from *R. crassostreae* have a cytotoxic effect on oyster hemocytes that cannot be solely attributed to lipopolysaccharide (LPS), a component of the membrane of gram-negative bacteria (Gomez-Leon et al. 2008).

At least two oyster lines with resistance to ROD are currently available, including the Frank M. Flowers (FMF) line and the University of Maine Flowers Select (UMFS) line, the latter of which contains germline materials from the former (Barber et al. 1998, Davis and Barber 1999; Lewis 2001). While traditional breeding practices have led to the production of ROD-resistant oysters, the genetic basis of resistance is presently unknown. Advanced genotypic breeding techniques, which necessitate information regarding markers and genes associated with disease resistance, have multiple

advantages over the current, traditional phenotypic breeding techniques including increased accuracy in estimating breeding values; amenability to disease resistance as a breeding trait, which only becomes apparent at death and may wane under traditional breeding regimes without consistent disease pressure; and enablement of simultaneous culture of mixed pedigrees whereby environmental effects can be minimized, which have been shown to be of great effect in oyster culture and may seriously confound main effects in traditional family-based selection (Langdon et al. 2003; Guo et al. 2008; Massault et al. 2008; Cancela et al. 2010).

The identification of potential genes and pathways responsible for an effective host defense response in the Eastern oyster to *R. crassostreae* is important not only to provide a basis for enhanced breeding techniques, but also to enhance understanding of innate immunity in a broader, evolutionary sense. Deep sequencing of the Eastern oyster transcriptome in response to bacterial challenge is a valuable and interesting contribution because the Eastern oyster is a member of Lophotrochozoa, a superphylum that has been poorly represented among genomic and transcriptomic datasets until the recent release and publication of the Pacific oyster, *Crassostrea gigas*, genome (Zhang et al. 2012). Research conducted on the innate immunity of model invertebrates with sequenced genomes has focused on deuterostomes like the purple sea urchin *Strongylocentrotus purpuratus*, the tunicate *Ciona intestinalis*, and the amphioxus *Branchiostoma floridae*, while research into the

innate immunity of protostomes has disproportionately focused on members of Pancrustacea and, more specifically, Arthropoda, including *Drosophila* spp. (Azumi et al. 2003; Kim and Kim 2005; Hibino et al. 2006; Huang et al. 2011). Invertebrate hosts lack the classical adaptive immune mechanism of receptor diversification through RAG-1- and RAG-2-mediated somatic recombination and gene conversion, yet they successfully combat widely varied types of microbes and parasites, which have comparatively high rates of mutation and rapid generation times (Flajnik and Du Pasquier 2004; Du Pasquier 2005). To mount effective and flexible defense responses to diverse pathogens, invertebrate hosts have developed diversified repertoires of receptors, regulators, and/or effectors (Messier-Solek et al. 2010). Though the precise role, relative importance, and presence/absence of specific genes and gene families relevant to host defense differs among the invertebrates in the taxa referred to above, several general strategies of diversification have been described including allelic diversity/SNPs/indels, alternative splicing, and gene family expansion (Ghosh et al. 2010; Messier-Solek et al. 2010). As an example of the latter, the genome of sea urchin *S. purpuratus* has undergone extensive expansions of Toll-like receptors (TLRs), NACHT- and leucine-rich repeat-containing (NLR) proteins and multi-domain scavenger receptors cysteine-rich (SRs), all of which directly or indirectly bind to pathogen associated molecular patterns (PAMPs) (Hibino et al. 2006). The *Dscam* gene, as studied in several arthropods, illustrates the strategy of diversification

through alternative splicing. The *Dscam* intron-exon architecture enables a possible production of tens of thousands of protein isoforms that function in homophilic binding and, likely, heterophilic, bacterial binding (Watson et al. 2005). Fibrinogen-related proteins (FREPs) identified in arthropods and molluscs have been shown to participate in agglutination and phagocytosis and have attained diversity through the multiple strategies of gene family expansion, alternative splicing, allelic diversity, and even somatic recombination (Leonard et al. 2001; Zhang et al. 2004; Ghosh et al. 2010).

While high-throughput, digital transcriptomic studies of host defense have been performed in molluscs other than the Eastern oyster, including the Pacific oyster *C. gigas*, the mussel *Mytilus edulis*, the Manila clam *Ruditapes philipinarum*, and others (de Lorgeril et al. 2011; Brulle et al. 2012; Philipp et al. 2012), disease challenge transcriptomic studies specific to Eastern oyster host defense have used medium-throughput approaches including expressed sequences tag (EST) analysis and microarrays (e.g. Jenny et al. 2002; Wang et al. 2010). The present study proposed to uncover genes, gene families, and general processes potentially responsible for disease resistance to ROD in the Eastern oyster. Sequences of cDNA from ROD-resistant and susceptible families of oysters exposed to *R. crassostreae* were assembled into a reference transcriptome. Differential gene expression and gene family analyses were used to identify genes, gene families, and general processes involved in oyster immunity, and single nucleotide polymorphism (SNP) were

identified in candidate genes. The present study found that genes involved in extracellular matrix (ECM) restructuring, cell adhesion, inflammation, metabolism, catecholamine signaling, and several other processes distinguished resistance from susceptibility to Roseovarius Oyster Disease. The present study also identified several large gene families important in Eastern oyster host defense including two families poorly characterized in invertebrates, the GTPase of the immunity associated protein (GIMAP) family and the interferon-induced protein 44 (IFI44) family, which appear to be expanded in molluscs and bivalves, respectively.

METHODS

Bacterial Challenge of Eastern Oysters

Juvenile Eastern oysters from 2 families with known differential susceptibility to ROD as determined in a previous study were provided by the Rutgers University shellfish hatchery (Guo and Gomez-Chiarri, in preparation). F3L oysters were F₃ generation progeny derived from a single pair mating of a female oyster from Rutgers NEH (Northeastern High-survival) line and a male oyster from Louisiana (LA). GX09 (henceforth GX) oysters were progeny of two full-sib families of F₃ generation derived from three full-sib families whose parents contained germline material from the following stocks/lines: Rutgers NEH and DBH (Delaware Bay High-survival line), LA, University of Maine Flower's Select (UMFS), and Frank M. Flower's (FMF). Oysters from the F3L families had an admixture of four possible haplotypes, while GX oysters had an admixture of eight possible haplotypes. Two hundred seventy-three juvenile oysters (shell height, 10-15 mm) each from F3L and GX were divided into two replicate tanks with filtered sterile seawater (FSSW) for bacterial challenge (designated F3L and GX). Three groups of additional oysters (2 x 50 GX and 1 x 50 F3L) were kept in separate tanks as unchallenged controls (designated CGX and CF3L, respectively). Oysters were acclimated during a period of 2 weeks from conditions at origin to experimental conditions (salinity 28-30‰, temperature 15-19°C). Oysters in the challenge tanks were exposed

by bath to *R. crassostreae*, strain CV919-312^T (Boettcher et al. 2005) at a final tank concentration of 7.5×10^6 colony forming units (CFU) ml⁻¹ (day 0 of challenge), while oysters in control tanks were not exposed. Oysters were fed Instant Algae (Reed Mariculture) every other day and water was partially changed (50%) weekly. Oysters were monitored weekly for 93 days for mortalities and for the presence of clinical signs of ROD (uneven valves and conchiolin deposits in shells of dead oysters). Infection by *R. crassostreae* was confirmed by PCR (Maloy et al. 2005).

Sample collection, cDNA preparation, and sequencing

Oyster whole body tissue was collected from 5 randomly sampled oysters each from CGX, GX, and F3L at days 1, 5, 15, and 30 following challenge and stored in Qiagen RNeasy RNA Stabilization Reagent until time of RNA isolation. All RNA molecules >200 nucleotides were purified using Qiagen RNeasy Mini Kit. Samples were checked for RNA purity using the Nanodrop 8000 spectrophotometer and random samples were checked using the Agilent 2100 Bioanalyzer. Total RNA from whole body tissue from 5 oysters per time per treatment was pooled for a total of 12 experimental samples (3 treatments x 4 time points). Pools of RNA samples for each of the experimental groups were selectively enriched for poly-A containing mRNA using the Illumina mRNA-Seq-8 Sample Prep Kit which involved poly-A capture by poly-T oligo-attached magnetic beads, fragmentation, cDNA

synthesis, adapter ligation, and purification and enrichment by PCR. The resultant cDNA libraries were sequenced on the Illumina GAIIx platform (one sample/5 individuals per lane, 1 lane per sample for a total of 12 lanes) (Genome Quebec, Canada).

Read processing and *de novo* assembly

Raw sequencing reads of 108 bp were pooled then processed and filtered for contamination of mitochondrial and ribosomal sequences by mapping to all *Crassostrea* spp. rRNA and mtDNA in NCBI Genbank database, and were filtered for vector sequences by mapping to Univec using bowtie2 (Langmead and Salzberg 2012; <ftp://ftp.ncbi.nih.gov/pub/univec>). The 5'-end of reads was trimmed to reduce GC-bias (Hansen et al. 2010). Using the btrim software package, Illumina adapters were trimmed, low-complexity artifacts were removed, and reads were further trimmed using adaptive quality trimming (Kong 2011). Reads less than 20 bp in length at this stage were discarded. Processed transcriptome reads were assembled using Trinity (release 20111126) with default options (Grabherr et al. 2011). Only those assembled contigs ≥ 200 bp were retained.

Contamination Removal

Transcriptome contigs were compared to the RefSeq protein database (Sayers et al. 2012). A custom python script, created by L. Dong (Brown

University), was used to parse BLAST output and identify possible contaminants. Contigs that had all top blast hits (a maximum of 10) with associated e-value $\leq 1e-06$ to proteins from Archaeobacteria, Bacteria, or Protozoa were discarded. Additional mitochondrial and ribosomal contaminants were identified and discarded during prot4EST *in silico* translation (see below) and through text searching of BLAST results. Phage integrase sequences were identified and discarded by comparing the transcriptome to Pfam_A using Pfam scan (version 1.3) and HMMER (version 3.0) with hits retained where e-value $\leq 1e-05$ (Finn et al. 2011). DNA transposons were identified and discarded using RepeatMasker (Smit et al. 2010).

Differential Gene Expression

Reads from individual treatments were aligned to the reference transcriptome using bowtie with parameters, “-v 3 --a --best --strata,” such that 3 mismatches were allowed per read to account for the high rate of polymorphism in oysters, while only reporting the alignment with the least number of mismatches for each “stratum” (Langmead et al. 2009). Transcript abundances in reads per kilobase per million reads mapped (RPKM) were estimated using RSEM (RNA-Seq by Expectation Maximization) through the Trinity plug-in, run_RSEM.pl. (Mortazavi et al. 2008; Grabherr et al. 2011; Li and Dewey 2011). To reduce bias from differential sequencing depth across

lanes, the trimmed mean of M values (TMM) method was used to calculate normalization factors for each lane by which read abundances were multiplied and adjusted (Robinson and Oshlack 2010). Read abundances in the control (non-challenged CGX) samples from days 15 and 30 were considered jointly as a control pool to increase the accuracy of estimating the biological variability between samples and increase power in identifying DEGs. Read abundances from control samples (CGX 1 and 5 d) were not included in this pool because of a mortality event that occurred between days 1 and 7 for CGX (Fig. 1). Read abundances for contigs in each of the samples were compared to read abundances in the control pool (CGX 15 and 30 d). Only those contigs with at least 1 count-per-million in at least 2 samples were tested for differential expression. Differential gene expression (DGE) testing was performed using edgeR, which assumes a negative binomial distribution and uses an empirical Bayes estimation and exact test to identify differentially expressed genes (DEGs) (Robinson et al. 2010). Significance values yielded by hypergeometric test were adjusted using the False Discovery Rate (FDR) correction and a contig was considered differentially expressed (DE) if it had an FDR-adjusted p-value ≤ 0.05 (Benjamini and Hochberg 1995; Robinson et al. 2010). DGE testing was performed on all contigs (“gene isoform” testing) and on read abundances summed for all contigs within each Trinity component (“gene model” testing) (Grabherr et al. 2011). Only those DEGs that were significant at the level of “gene isoforms” and belonged to

components that were significant at the level of “gene models” were considered differentially expressed. While the relationship of Trinity contig to Trinity component works on the level of graph space and is not precisely biological (Grabherr et al. 2011), the relationship as it was applied only added a stringent filter to the pool of DEGs and did not contribute new DEGs.

Heatmap Analysis

For all genes differentially expressed in any one of the treatments GX-d1, GX-d5, F3L-d1, F3L-d5, F3L-d15, and F3L-d30, the $\log_2$ -transformed RPKM for each gene in each of eight treatment-days (including control, CGX-d15 and CGX-d30) were Z-score centered by subtracting then dividing by the mean $\log_2$ -transformed RPKM by gene. Genes were hierarchically clustered using Euclidean distance and complete linkage of the Z-score-transformed gene expression. Treatment-days were clustered using the complete linkage Euclidean distance of the Spearman correlation of the Z-score-transformed gene expression. Clustering and visualization were performed using the *fastcluster* and *gplots* packages, respectively, in the R programming environment (Bolker et al. 2010).

Annotation and Functional Enrichment

Transcriptome contigs were compared to the NCBI protein non-redundant database and Uniref100 using BLASTX (Altschul et al. 1997). Hits

were retained with $e\text{-value} \leq 1e\text{-}6$. For all subsequent methods herein employed, the annotation to NR was used because a greater percentage of the transcriptome could be annotated using this database. Gene Ontology (GO) terms were mapped to best BLASTX hits for each contig using the Blast2GO pipeline (version 2.3.5) (Conesa et al. 2005). Functional enrichment for each treatment comparison (GX_early, F3L_early, and F3L_late, where early includes days 1 and 5 after challenge, and late days 15 and 30) was performed by comparing the numbers of gene ontology (GO) terms associated with annotations of DEGs to the numbers of terms associated with all transcripts not DE using the R package topGO, which also accounts for the hierarchical topology of the GO graph (Alexa et al. 2006). The topGO “elim” algorithm eliminates significant child nodes from tests of significance for parent nodes, as significant child nodes may otherwise confer significance to their parent nodes. Fisher’s exact test was used to determine significance of enrichment of each GO term, with Bonferroni-adjusted p-values ≤ 0.05 taken as significant. Significantly functionally enriched GO terms were visualized in semantic space using SimRel functional similarity measure (Schlicker et al. 2006), REVIGO online visualization tool (Supek et al. 2011), and custom R scripts.

Protein Family Identification and Test for Enrichment

Transcriptome contigs were compared to Pfam_A using PfamScan (version 1.3) and HMMER (version 3.0) and only hits with an e-value $\leq 1e-5$ were retained (Eddy 1998; Finn et al. 2011; Pfam_A downloaded July 7, 2012). It is possible for a contig to have multiple hits to the same Pfam profile hidden Markov model (HMM). These hits were made non-redundant by retaining only the most significant contig-to-HMM hit. DEGs for each treatment comparison were tested for enrichment of both Pfam families and Pfam domains. Pearson's chi-squared test was performed to test for the enrichment of each unique HMM accession. If the expected count for any cell in the 2x2 contingency table was 5 or fewer, Fisher's exact test was performed instead. Significance p-values were adjusted for the number of independent enrichment tests by the False Discovery Rate correction method and p-values ≤ 0.05 were considered significant (Benjamini and Hochberg 1995).

Peptide Similarity Clustering

Prot4EST, a perl script that integrates several programs and approaches to find the best open reading frame, including, in order of priority, BLASTX similarity, codon usage bias, and longest open reading frame (Wasmuth and Blaxter 2004), was used to translate transcriptome contigs into a set of putative peptide sequences. For codon usage bias, all full-length Bivalvia Uniref100 proteins were used as the training set. Translated transcripts were then clustered into groups by similarity using TribeMCL. A

custom perl script provided by I. Misner (University of Rhode Island) was used to prepare the translated transcriptome for all-versus-all BLASTP and to execute this similarity search. Hit, or edges, with $e\text{-value} \leq 1e\text{-}5$ and hit identity $\geq 20\%$ were retained as significant. Negative $\log_{10}$ -transformed e -values above 99 were set to 99 so that subsequent edge weights for MCL clustering were not skewed (Chen et al. 2007). The Markov cluster (MCL) algorithm set on the default inflation index, 1.5, was then used to dissolve less reliable edges that may result from sequence similarity errors or from shared protein domains across families (Enright 2002). MCL returns connected components or clusters of contigs based on protein sequence similarity, which can be defined as a set of nodes such that any two nodes in the same cluster is connected by a path of edges and any two nodes in different clusters is not connected by any path of edges.

Two similarity graphs were constructed: a *C. virginica*-only graph, containing translated contigs from our *de novo* assembly only, and a multi-species sequence similarity graph, constructed using translated *C. virginica* transcripts along with sequences from evolutionarily diverse species including the basal eumetazoan *Nematostella vectensis*; the deuterostome *Strongylocentrotus purpuratus*; the basal chordates *Branchiostoma floridae*, *Saccoglossus kowalevskii*, and *Ciona intestinalis*; the basal vertebrate *Petromyzon marinus*; the pancrustaceans *Drosophila melanogaster* and *Daphnia pulex*; and several molluscs including the gastropods *Lottia gigantea*,

Aplysia californica, *Biomphalaria glabrata*, and *Lymnaea stagnalis*, and a bivalve species, the Pacific oyster *Crassostrea gigas*. Sequences were downloaded from a variety of sources in the form of ESTs and protein sequence (Appendix A, Table S1). ESTs were translated to peptide using the prot4EST pipeline. Full-length Gastropoda Uniref100 proteins were used as the training set for determination of the codon usage bias for the three gastropod species, while full-length Bivalvia Uniref100 proteins were used as the training set for *C. gigas* ESTs.

Protein sequences and translated ESTs from all organisms were concatenated, formatted, and filtered using the first two steps of the OrthoMCL pipeline, *orthomclAdjustFasta* and *orthomclFilterFasta* (Li et al. 2003; Fischer et al. 2011). Proteins were compared in a parallel run of all-versus-all BLASTP, retaining hits with $e\text{-value} \leq 1e\text{-}10$ (Altschul et al. 1997). The *orthomcl* utility *orthomclBlastParser* was used to calculate the percent identity of each hit. The list of significant hits, or edges, was then made non-redundant and with a maximum negative $\log_{10}$ -transformed $e\text{-value}$ of 99 using custom python scripts. The resultant network file was filtered using MCL set on the default inflation index of 1.5 (Enright et al. 2002).

Enrichment of DEGs among TribeMCL clusters

Each *C. virginica*-only TribeMCL cluster was independently tested for enrichment of the 3 DEG treatment comparison groups (GX_early, F3L_early,

and F3L_late). Pearson's chi-squared or Fisher's test was performed to test for the enrichment of each group of DEGs for each TribeMCL cluster as described above. Pragmatic rules were adopted for annotating clusters. Enriched TribeMCL clusters were annotated with a protein name if > 50% of the contigs had identical or nearly identical BLASTX best hits and the remainder of contigs had no significant hits, or if $\geq 80\%$ of the contigs had identical or nearly identical BLASTX best hits and the remainder of contigs had dissimilar best BLASTX hits (which may be the case in the sharing of domains, e.g., chymotrypsin and neurotrypsin). Enriched TribeMCL clusters were putatively annotated (noted with asterisks) if < 50% half of the contigs had identical or nearly identical BLASTX best hits and the remainder of contigs had no significant hits or had a small number of dissimilar but non-repeating hits (frequency equal to one). Multiple names were used for the annotation if a number of non-identical hits of comparable frequency composed a majority of the cluster (e.g., "hemicentin/rhamnospondin/thrombospondin*"). All other TribeMCL clusters were named "unknown."

Selection of Putative Diversified Families

Gene diversification implies a gene family expansion comparative to other lineages (e.g. by gene duplication) and/or gene diversity significantly greater than that which is present in the genomic sequence alone (e.g. by alternative splicing). To identify "diversified" families with host defense

relevance, working definitions were adopted for the terms: gene family, expansion, and host defense relevance. A gene family was defined as a group of transcripts with similarity between one another (membership in the same TribeMCL cluster) and/or similarity to the same gene family based on BLAST matches (e.g. two transcripts with BLAST similarity to C1qDC proteins). Gene families were considered to be expanded if: 1) families consisted of at least 50 non-redundant transcripts (in order to find the largest families and to significantly reduce the number of gene families under consideration); and 2) the number of non-redundant transcripts in that gene family in *C. virginica* exceeded the number of non-redundant transcripts/proteins in that gene family in at least half of the other 14 organisms considered. The putative diversified gene families were regarded as relevant to the oyster host defense to bacterial challenge if they contained a significant portion of DEGs as determined by Pfam and/or TribeMCL-DEG enrichment.

Certain diversified gene families were selected for further analyses based on the following criteria: 1) high rank of significance in the enrichment of DEGs belonging to the gene family in Pfam and/or TribeMCL-DEG enrichment sets (preferably both); 2) large sizes of TribeMCL cluster(s) composing the gene family; and 3) previously known importance in immunity, based on literature searches. Those clusters containing proteins with repetitive domains (i.e. leucine-rich-repeats, LRR, or epidermal-growth factor, EGF, domains)

were not included in further analyses, due to the fact that the presence of multiple repeats frustrated similarity clustering. When a diversified family was found that met several of the conditions stated above, a term search was conducted to find relevant hits among the BLAST results (e.g. terms like “serine protease,” “trypsin,” etc. were used to find serine proteases). The contig list in each TribeMCL cluster was expanded to include all contigs in a TribeMCL cluster if and only if more than or equal to half of the contigs in that cluster had BLASTX best hits to the group of interest and the remainder of contigs had no significant or conflicting hits, whereas the contig list was contracted if less than half of the contigs in that cluster had BLASTX best hits to the group of interest and the remainder of contigs had hits to dissimilar proteins, in which case, all contigs in that cluster were excluded from that group of interest. The contig list was neither contracted nor expanded but remained unchanged in the cases of a TribeMCL cluster in which more than or equal to half of the contigs in that cluster had BLASTX best hits to the group of interest but the remainder of contigs had significant hits to dissimilar proteins or if less than half of the contigs in that cluster had BLASTX best hits to the group of interest and the remainder of contigs had no significant hits. Those contigs (with hits to the gene family of interest) that were not contained in the TribeMCL graph were also retained.

Translated transcripts for each of the select diversified families of interest were reduced to non-redundant sets using CD-HIT, on settings “-G 0 –

aS 0.50 –c 0.95,” which reduced the sets of translated transcripts by 95% redundancy for an alignment length at least 50% of the smaller translated transcript (Weizhong and Godzik 2006). The nucleotide sequence of the non-redundant transcripts for each group of interest was mapped to the *C. gigas* genome using BLAT (Kent 2002; oyster.v9.fa, http://gigadb.org/pacific_oyster/, Zhang et al. 2012). BLAT hits were scored as #Matches + #Repmatches - #Mismatches - #Query Gap Count - #Target Gap Count and scores ≥ 30 were retained. The number of unique *C. gigas* loci was determined by summing the number of non-overlapping regions of the *C. gigas* genome regions to which queries aligned, with each BLAT target region padded by an additional 200 bp upstream and downstream to account to some extent for the possibility of truncated mappings. The estimation of the number of genome loci for each diversified family was separate from the estimation of the number of genes for that family in the *C. gigas* genome. For the remainder of the diversified families of interest, nucleotide sequences for the non-redundant transcripts and protein sequences from *C. gigas* GLEAN gene models were reciprocally compared by BLASTX and TBLASTN, respectively (oyster.v9.glean.rename.gff.pep.gz, http://gigadb.org/pacific_oyster/, Zhang et al. 2012). The number of genes in each family present in the *C. gigas* genome was estimated as the number of peptides that had reciprocal hits to the non-redundant *C. virginica* transcript sequences with e-value $\leq 1e-05$ and that had

best BLASTP hits to NCBI's NR database correspondent to the family of interest.

SNP detection and polymorphism in candidate oyster host defense genes

For single nucleotide polymorphism (SNP) detection, processed reads were aligned to the transcriptome using Bowtie2 on "--very-sensitive" setting. PCR duplicates were removed with the Picard Tools utility MarkDuplicates (<http://picard.sourceforge.net/>). SNPs were called using samtools mpileup with "-g" flag to generate genotype likelihoods (Li et al. 2009). The output was filtered using vcfutils varFilter with minimum mapping quality of 25, minimum read depth of 25, and maximum read depth of 200 (Q 25 -d 25 -D 200). In the R programming environment, SNPs were further filtered for minor allele frequency $\geq 10\%$ and minor allele count ≥ 5 . All transcriptome contigs were aligned to their protein translations using estwise (version 2.2.0) to find the reading frame and start and stop sites (<http://www.ebi.ac.uk/Wise2/>). Custom R scripts utilizing the R packages, ShortRead and SeqinR, were used to determine whether SNPs were synonymous or non-synonymous based on the reading frame, start and stop sites (Charif et al. 2007; Morgan et al. 2009). For each contig containing SNPs within the determined coding region, a multi-sequence file was generated consisting of the original nucleotide contig sequence, and one nucleotide sequence per SNP consisting of the original

nucleotide sequence modified to reflect the base composition of the SNP. Each multi-sequence file in categories of interest like DEG groups (GX_early, F3L_early, F3L_late) and diversified gene families, was run through the codeml program in the PAML package (version 4.1b) to obtain the rate of non-synonymous mutations (dN), the rate of synonymous mutations (dS) and the ratio of these two rates (dN/dS) (Yang 2007). The M0 model was assumed of a single ω for all lineages, initial K and ω values of 1.0 and 0.5, respectively, and the F3x4 codon frequency model (Goldman and Yang 1994). The final K and ω values were estimated by codeml (Yang 2007).

Comparison of putative diversified families between diverse taxa

The diversified gene families analyzed in depth using the *C. virginica*-only TribeMCL graph were tracked and enumerated in the multi-species TribeMCL graph. For each diversified family, the set of nodes first consisting of the known *C. virginica* nodes was reiteratively expanded with members from other taxa, adding neighbor nodes until the set failed to grow. The final subgraph was reduced by examination of the *C. virginica* annotations and the *C. gigas* annotations (downloaded from http://public-contigbrowser.sigenae.org:9090/Crassostrea_gigas/index.html). Multi-species families were retained if at least half of the contigs had BLASTX hits to the gene family of interest and the remainder of contigs had no significant hits, otherwise, the cluster was excluded. The final set of proteins for each

diversified family of interest were subsequently reduced to non-redundant sets, independently for each family, using CD-HIT, on settings “-G 0 -aS 0.50 -c 0.95,” which reduced the sets of proteins by 95% redundancy for a alignment length at least 50% of the smaller protein (Li and Godzik 2006). The number of non-redundant proteins was enumerated for each species for each diversified gene family. In a few select cases, proteins from different species were so similar as to cluster together, in which case, proteins were counted for both species. For example, two serine proteases from *D. melanogaster* that reduced to one sequence by CD-HIT, would have been counted as one serine protease, while one serine protease each from *D. melanogaster* and *D. pulex* that reduced to one sequence by CD-HIT would have been counted as one serine protease for each species. In addition to an estimation of the number of non-redundant proteins for each species for each diversified family, the number of gene models was estimated for each species for each diversified family, according to the formula:

$$NPX \times (NGM \div NP) = NGMX$$

where NPX = Number of non-redundant proteins in diversified family X,

species Y

NGM = Number of total gene models in species Y

NP = Number of total non-redundant proteins used for similarity
clustering in species Y

$NGMX$ = Number of gene models in diversified family X, species Y

Estimations of the total number of gene models for each species (*NP*, above) were taken from various sources including the published literature and websites of genome sequencing centers (Appendix A, Table S2). Estimates for total number of non-redundant proteins and number of gene models for *L. stagnalis* are not reproduced here because the estimates were unreasonably low.

Phylogeny of diversified families

For a select few diversified groups of interest, the non-redundant protein sequences from the multi-species TribeMCL graph were aligned using the E-INS-I option of MAFFT (version 6, Katoh and Toh, 2008). Multiple sequence alignments were viewed in JalView (version 2.6.1) and manually curated to extract blocks of well-aligned sequence (Waterhouse et al. 2009). The models of sequence divergence that best fit the multiple sequence alignments were found using ProtTest (version 3, Abascal et al. 2005). FastTree2 was used to generate a phylogenetic tree (Price et al. 2010). For both IFI44 and GIMAP families, the WAG model was used to model sequence divergence (Whelan and Goldman 2001), and a hybrid of CAT (Stamatakis 2006) and gamma approximations were used to account for rate variation across sites (Price et al. 2010). Trees were viewed as circular cladograms in Dendroscope (version 3.2.2, Huson et al. 2007).

RESULTS

Oyster survival in response to bacterial challenge

Oysters from the F3L family experienced a consistent and high rate of mortality after challenge with the bacterial pathogen *R. crassostreae*, reaching over 90% cumulative mortality by the end of the 93-day period (Fig. 1). The survival curve of the challenged F3L oysters was significantly different from all other groups (log-rank survival, $p < 0.01$). Pearson's chi-squared test with Bonferroni corrections was used to compare oyster survival in the 3 groups used for sequencing, CGX, GX, and F3L, at day 28 (the closest time point at which mortality was tallied before collection of samples for RNA isolation at day 30), and at day 93 (the final time point of the bacterial challenge). F3L had a significantly higher cumulative mortality than GX at day 28 ($p < 0.01$), and a significantly higher cumulative mortality than GX and CGX at day 93 ($p < 0.01$). No significant differences in mortality were observed between unchallenged control oysters (CF3L and CGX) and oysters from the resistant family challenged with *R. crassostreae* at days 28 and 93 after challenge. Oysters from the control ROD-resistant family (CGX) suffered a mortality event between days 1 and 7 (20% cumulative percent mortality by day 7), with an additional 10% mortality for the 86 days following that event (Fig. 1).

Oyster transcriptome assembly

From a total of 4.1×10^8 Illumina GAIIx-sequenced cDNA reads of 108 bp, after filtering and trimming, the final set consisted of 3.8×10^8 reads of 94 ± 5 bp (Appendix B. Fig. S1). After the Trinity assembly of 374,029 contigs was further filtered for contaminants, 356,237 contigs remained with a mean length of 440 bp and an N50 of 487 bp (Table 1). A BLASTX search to the NCBI NR protein database led to annotation of 19.8% of the transcriptome. Of the total transcriptome, 22,934 contigs (16.3%) were at least 1 Kb in length.

Differential gene expression in oysters in response to bacterial challenge

Differential gene expression analyses of samples GX-1d, GX-5d, F3L-1d, F3L-5d, F3L-15d, and/or F3L-30d yielded a total 6,296 differentially expressed genes (DEGs), or 1.8% of the total *de novo*-assembled transcriptome. When samples were clustered by gene expression patterns, two major clusters separating F3L and GX/CGX treatments were evident. Furthermore, within each of these clusters, the following subclusters were detected: F3L 1 and 5d (F3L_early); F3L 15 and 30d (F3L_late); GX 1 and 5d (GX_early); and CGX 15 and 30d (control) (Fig. 2). This pattern of treatment-day clustering (GX_early, F3L_early, and F3L_late) was used in further analyses. The focus of the analysis was placed on GX_early DEGs according to the rationale that genes involved in disease resistance would likely be expressed in resistant GX oysters at early time points after exposure to the pathogen.

The more dramatic response to bacterial challenge of F3L compared to GX in terms of cumulative mortality was reflected in a similar differential response in terms of gene expression (Fig. 3). Of the 356,237 total transcripts tested for differential expression, 6,097 transcripts were differentially expressed in F3L_early and/or F3L_late, compared to only 552 transcripts differentially expressed in GX_early. F3L DEGs were described at the gene level where overlap was found with GX_early DEGs and were further described at the scale of Gene Ontology functional enrichment, Pfam enrichment, and TribeMCL-DEG enrichment (Tables 2 - 5). A greater share of GX_early DEGs was shared with F3L_early and/or F3L_late DEGs (64%) than was unique to GX_early (36%) (Fig. 3).

DEGs shared by GX_early and F3L treatments should include (among others) genes associated with host defense and supporting functions (Tables 4 & 5). DEGs unique to GX_early should include genes contributing to disease resistance in the GX family (Tables 2 & 3). The most highly differentially expressed, annotated, up-regulated DEGs unique to GX_early (potentially involved in disease resistance) included, among others, transcripts that annotated to 2 scavenger receptor cysteine-rich proteins, 2 fibropellin ia proteins, 2 inhibitor of apoptosis (IAP) proteins, cytochrome p450, interleukin 17d, a fibrinolytic enzyme, and a disintegrin and metalloprotease with thrombospondin motifs 8 (ADAMTS8) (Table 2). The most highly differentially expressed, annotated, down-regulated DEGs unique to GX_early included

transcripts that annotated to the development-related protein rapunzel, 2 collagen proteins, tenascin xb, 2 cubilin proteins, inhibitor of apoptosis (IAP), a c-type lectin, a melatonin receptor (Table 3). The most highly differentially expressed, annotated, up-regulated DEGs shared between GX_early and F3L_early and/or F3L_late (potentially involved in responses to bacterial infection) included transcripts that annotated to 2 serine protease inhibitors, 2 dopamine beta-hydroxylases, 2 fatty acid synthases, 2 sulfatases, cytochrome p450, a C1q domain-containing (C1qDC) protein, and heat shock protein 60 (HSP60)(Table 4). The most highly differentially expressed, annotated, down-regulated DEGs shared between GX_early and F3L_early and/or F3L_late included transcripts that annotated to a C1qDC protein, two monocarboxylate transporters, multiple epidermal growth factor 11 (MEGF 11), deleted in malignant brain tumors 1 (DMBT1), and sushi-repeat-containing x-linked 2 (Table 4). Because different contigs sometimes shared the similar annotations (suggesting that they could potentially be members of a gene family, alternatively spliced forms, or parts of the same transcript that were not assembled together), annotations were manually compared to find truly unique transcripts among GX_early unique DEGs. Sixteen non-redundant transcript annotations were found in the GX_early sample, including arginase I, arginase II, rho gtpase, and cubilin (down-regulated at day 5); and unc-5, furin, and interleukin 17d (up-regulated at day 1) (Appendix A. Table S3).

The relative ratio of up-regulated and down-regulated DEGs differed between GX and F3L, with a significantly greater number of up-regulated transcripts in GX_early DEGs (Pearson's chi-squared analysis, $p < 0.01$) and a significantly greater number of down-regulated transcripts in F3L DEGs ($p < 0.01$).

Gene Ontology categories enriched among oyster DEGs upon bacterial challenge

As there were fewer DEGs in GX_early compared to the F3L groups, so there were comparatively fewer enriched GO terms (Fig. 4). The most significantly enriched biological process GO term among GX_early up-regulated DEGs was “protein folding”, corresponding to 3 DEGs annotated to HSP60 that were up-regulated only at day 1. “Protein folding” was also enriched among F3L_early DEGs, also corresponding to HSP60 transcripts, some of which were down-regulated and some up-regulated. The enrichment in “defense response” among GX_early DEGs corresponded to transcripts that annotated to interleukin 17 (IL17), which were strongly up-regulated at day 1. Terms closely allied to defense response were found enriched among F3L_early DEGs including “defense response to bacterium” and “response to molecule of bacterial origin”, corresponding to transcripts that annotated to angiotensin-converting enzyme, defensin, and immune-responsive gene 1, the latter of which was also up-regulated in GX at day 1. The other biological

process terms enriched among up-regulated GX_early DEGs were the related terms “programmed cell death” and “apoptotic process”, corresponding to transcripts that annotated to IAP transcripts. Though related cell death terms were not found to be enriched in F3L treatments, 3 IAP transcripts were differentially expressed in F3L, 2 of which were down-regulated at all time points. The most significantly enriched F3L early biological process term, found among up-regulated DEGs, was “cholesterol transport”, corresponding to epididymal secretory protein E1. Other transport terms enriched among F3L_early DEGs include “hexose transport” (up-regulated) and “amino acid transport”-related and “carboxylic acid transport” (down-regulated). Various development-related terms, including the closely related terms “blood vessel morphogenesis,” “angiogenesis,” and “vascular development” as well as “fin development” were enriched among F3L_early down-regulated DEGs while “fin development” and neuron-related development terms were enriched among F3L_late down-regulated DEGs. Several terms related to carboxylic acid metabolism were uniquely enriched among F3L_late up-regulated DEGs, corresponding to various decarboxylases. Also uniquely enriched among F3L_late up-regulated DEGs was “oxidation-reduction process”, corresponding to several cytochrome p450 transcripts.

With respect to enriched molecular function GO terms, commonalities across treatments included “enzyme inhibitor activity” among GX_early and F3L_late up-regulated DEGs (Fig. 4). The closely related terms

“endopeptidase inhibitor/regulator activity” were enriched among F3L_early up-regulated DEGs, while “peptidase inhibitor activity” was enriched among F3L_late up-regulated DEGs and was the most significantly enriched molecular function term among F3L_late DEGs. Also enriched among F3L_early up-regulated DEGs was the highly significant term “monooxygenase activity”, corresponding to several cytochrome p450 and dopamine beta hydroxylase transcripts, while among F3L_late up-regulated DEGs the term “oxidoreductase activity” was enriched, also corresponding to several cytochrome p450 and other transcripts. There were several “hydrolase”-related terms enriched among F3L_early DEGs including “hydrolase activity, hydrolyzing O-glycosyl compounds”, corresponding to CHIT3 protein among others.

The GO Cellular Component hierarchical superstructure is not as ramified as are Biological Process and Molecular Function, and accordingly, the enriched terms found here were fewer. “Extracellular region” was the only cellular component GO term enriched among GX_early up-regulated DEGs, and was also enriched among F3L_early up-regulated and F3L_late down-regulated DEGs (Fig. 4). The transcripts that annotated to this term were diverse and included tissue inhibitors of metalloproteinase (TIMPs), HSP60, fibropellin, and others. Interestingly, all enriched cellular component terms were related to the membrane or extracellular matrix.

Pfam protein families and domains enriched among oyster DEGs upon bacterial challenge

Of 356,237 contigs, 21,446 contigs (6%) had significant hits to 2,367 different Pfam families. The most significantly enriched Pfam family among GX_early DEGs, which was also enriched among F3L_early DEGs, was NAD_binding_5, a family of myo-inositol-1-phosphate synthases involved in signal transduction and the mobilization of calcium (Table 5; Berridge 1984). Several proteins in the thrombospondin_1 (TSP_1) family, including 7 hemicentin-like transcripts, were differentially expressed in all treatments, with one of these up-regulated at days 1 and 5 in GX. Transcripts matching the AIG1 family included many annotating as GIMAP proteins; nearly all GIMAP DEGs were down-regulated in F3L (data not shown). The remainder of the Pfam families enriched among GX_early DEGs were enriched solely in that class of DEGs including IL17, HSP20, bZIP_2 (a family of transcription factors including creb-binding protein; Schumacher et al. 2000), patched (involved in developmental signaling; Ingham et al. 1998), and sulfatase. Families enriched among DEGs from both F3L_early and F3L_late and not GX_early included An_peroxidase (response to oxidative stress and bactericidal defense; Zamocky et al. 2008), mannose-6-phosphate receptors (Man-6-P_recep, involved in biogenesis of lysosomes; Griffiths et al. 1998), Dynamin_N (necessary for endocytosis; McClure and Robinson 1996), and ApoL (role in lipid transport, linked to innate immunity in humans; Perez-Morga

et al. 2005; Vanhollebeke et al. 2006). Slightly greater than half (18) of the 39 Pfam families enriched among F3L_early DEGs were uniquely enriched among that class of DEGs. Among Pfam families uniquely enriched among F3L_early DEGs, some have obvious relevance to host defense including tissue inhibitors of metalloproteinase (TIMP), Pacifastin_1 (serine protease inhibitor; Simonet et al. 2003), and Von Willebrand Factor (VWF) type D domain-containing proteins (involved in blood clotting and cell adhesion; Jorieux et al. 1998). Other Pfam families uniquely enriched among F3L_early DEGs included several types of transporters including SSF, Sugar_tr, SNF, and AA_permease_2. Fewer than half (8) of the 21 Pfam families enriched among F3L_late DEGs were uniquely enriched among that class of DEGs, including ovomucin-binding proteins (VOMI), proteins that contain the C-terminal domain of Chitobiase/beta-hexosaminidase (CHB_HEX_C_1, which degrade chitin), and Mucin2_WxxW, (help form an insoluble extracellular matrix that protects epithelial linings; <http://pfam.sanger.ac.uk/>; Johansson et al. 2011).

Of 356,237 contigs, 23,220 contigs (6.5%) had significant hits to 1,591 different Pfam domains. The most significantly enriched Pfam domain among F3L_early and second most significantly enriched among F3L_late DEGs, was C1q, which was also enriched among GX_early DEGs (Table 6). Only three other domains were enriched among all classes of DEGs, namely, SRCR, Kunitz_BPTI, and HYR. Only two domains were enriched among F3L_early

DEGs and GX_early DEGs, including Trypsin (protease involved in digestion and innate immunity; Rawlings and Barrett 1994; Ross et al. 2003). Most of the domains that were enriched among GX_early DEGs were unique to that class of DEGs and included baculovirus inhibitor of apoptosis repeat (BIR) protein (involved in the inhibition of apoptosis; Silke and Vaux 2001); Complement C1r-like EGF (cEGF)-like (involved in blood coagulation; Wouters et al. 2005); and CUB (involved in a variety of processes including inflammation, angiogenesis, and endocytosis; Blanc et al. 2007). Ten Pfam domains were enriched only among F3L_early and F3L_late DEGs including Fibrinogen_C; Stichodactyla toxin (ShK, present in proteins that block voltage-gated potassium channels and found in antimicrobial proteins in *C. elegans*; Tudor et al. 1996; Troemel et al. 2006); and Copper type II ascorbate-dependent monooxygenase (Cu2_monooxygen; present in dopamine beta-hydroxylases and associated with immunity through the regulation of catecholamine biosynthesis; Flieri et al. 2009). Twenty-eight of the 44 Pfam domains enriched among F3L_early DEGs were uniquely enriched among that class of DEGs and included Von Willebrand Factor A (VWA, present in mostly extracellular proteins involved in cell migration, cell adhesion, and other processes; Colombatti et al. 1991; Colombatti et al. 1993) and Kazal_1 (present in serine protease inhibitors; Rawlings et al. 1994). Only 4 of 18 Pfam domains enriched among F3L_late DEGs were uniquely enriched among that class of DEGs and included Cupin_8, found here in transcripts that annotated

as jumonji domain-containing proteins (involved in histone demethylation; Klose et al. 2006) and phosphatidylserine receptors (involved in apoptotic cell clearance; Li et al. 2003).

Sequence similarity clusters enriched for oyster DEGs upon bacterial challenge

Of 356,237 contigs, 82,498 contigs (23%) were clustered by TribeMCL similarity clustering into 18,873 clusters, 187 of which were enriched for F3L_early, F3L_late, and/or GX_early DEGs (Fisher's exact test or Pearson's chi squared test, $p < 0.01$). The top six clusters most significantly enriched for GX_early DEGs were also enriched for F3L_early DEGs and sometimes for F3L_late DEGs. Generally, the clusters most significantly enriched for F3L_early DEGs were also enriched for F3L_late DEGs and *vice versa*.

Several putative diversified gene families were selected for further analyses based on the criteria listed in the methods. These included serine proteases (SPs), serine protease inhibitors (SPIs), fibrinogen domain-containing proteins/ fibrinogen-related proteins (FREDs/FREPs), C1qDC proteins, c-type lectin domain-containing (CTLDC) proteins, deleted in malignant brain tumors 1 (DMBT1) and scavenger receptor cysteine-rich type 12, interferon-induced protein 44 (IFI44), and GTPase of the immunity associated protein family (GIMAP) proteins (Table 9).

***Characteristics of selected putative diversified gene families in oysters
in terms size, polymorphism, and response to bacterial challenge***

C1qDC was among the largest gene families differentially expressed in *C. virginica* in response to bacterial challenge, with 391 non-redundant annotated transcripts, 323 of which mapped to 149 genomic loci (non-overlapping mappings ± 200 bp) in the *C. gigas* genome (Table 9). The estimated number of C1qDC transcripts in the *C. virginica* transcriptome is consistent with the recent estimate made by Zhang et al. (2012) of 321 C1qDC genes in the *C. gigas* genome. The widest disparity between the number of family members identified in the *de novo* transcriptome for *C. virginica* and the numbers estimated from the *C. gigas* genome was observed for the GIMAP family (Table 9). Of a total of 173 non-redundant GIMAP *C. virginica* transcripts, 158 transcripts mapped to 33 *C. gigas* loci. Only 19 genes were found in the *C. gigas* set of gene models. Certain *C. gigas* loci had a great many mapped GIMAP transcripts, including one locus of 11 transcript hits, two loci of 12 transcript hits, one locus of 17 transcript hits, and one locus of 25 transcript hits (data not shown).

The average rate of polymorphism for the entire transcriptome, based on our stringent thresholds, was 272 ± 299 bp/SNP (mean $\pm$ SD). Forty-nine non-synonymous SNPs and 109 non-synonymous SNPs were identified in transcripts uniquely differentially expressed in GX_early and in both GX_early and F3L, respectively (Appendix A, Table S4). There was a wide range of

variation (from tens to hundreds) in the number of non-synonymous SNPs that were identified in each of the selected diversified gene families (Appendix A, Table S4). Mean dN/dS must be interpreted with caution given the limited numbers of individuals in each sample (limited SNP density) and the necessity of adopting certain assumptions (initial ω/K , codon freq. model, etc.). While many individual transcripts had dN/dS values > 1 , limited SNP density led to a consideration of dN/dS by gene family rather than by sequence or site. Serine proteases had the lowest mean dN/dS, by at least an order of 2, suggesting that this group is under stronger functional constraint than other groups. The highest mean dN/dS belonged to C1qDC proteins and IFI44.

In terms of differential expression within the diversified families of interest, nearly all differentially expressed SPs were up-regulated at GX day 5 or F3L day 5; nearly all SPIs were up-regulated at GX day 1 and 5 or F3L day 1 and 5; all FREDs/FREPs were down-regulated in F3L at one or more time points and no FREDs/FREPs were differentially expressed in GX; nearly all GIMAP DEGs were down-regulated in F3L at one or more time points and a few were down-regulated in GX day 1, while one transcript was up-regulated in GX day 5 (Fig. 6); nearly all IFI44 DEGs were down-regulated in F3L at one or more time points, while one transcript was up-regulated at F3L day 5, and one was up-regulated in GX at day 5 (Fig. 7). For the remainder of the diversified families, C1qDC, CTLDC, and DMBT1/SRCR type 12, a consistent pattern of differential expression could not be observed (data not shown). For

the relatively few transcripts in the select diversified families that were differentially expressed in both GX_early and F3L_early (19 transcripts total), the direction of differential expression, that is, up- or down-regulation relative to control, tended to match between GX_early and F3L_early, except for 2 CTLDC transcripts and 1 DMBT1/SR type 12 transcript (data not shown).

Comparison of size and phylogeny of selected diversified gene families in oysters and other organisms

Certain patterns could be seen in the size of selected diversified gene families across taxa and across the diversified groups (Fig. 5). First, the number of proteins and the number of gene models for each diversified family varied widely, especially for species for which ESTs or poorly-curated gene models were used. FREPs/FREDs and CTLDC proteins, known to be highly expanded in molluscs (Ghosh et al. 2010; Gerdol et al. 2011), appeared to have the greatest number of members. Compared to the number of C1qDC proteins present in molluscan species, few are present in non-molluscan species. Scavenger receptors with class B SRCR domains, well studied in *S. purpuratus*, showed tens of proteins/gene models in the bivalve species considered here. The warm colors signifying a high column z-score for IFI44 for bivalve species and for GIMAP proteins for molluscan species indicates that IFI44 is greatly expanded in bivalves and GIMAP is greatly expanded in molluscs. No IFI44 transcripts were found in the gastropod species considered

here and few were found in other non-bivalve invertebrate species.

Phylogenetic analysis of GIMAP sequences showed 3 major clusters (Fig. 8 and 9). One major cluster of sequences contained, in addition to many *C. virginica* sequences, comparable numbers of sequences from the other bivalves species, including *P. fucata* and *C. gigas*, and a comparable number of sequences from one gastropod, *L. gigantea*, along with one sequence each from the hemichordate, *S. kowalevskii*, the amphioxus, *B. floridae*, and the basal deuterostome, *N. vectensis*. The second major cluster contained many sequences from all mollusc species, *S. kowalevskii*, and *B. floridae*. The third major cluster contained the majority of *C. virginica* sequences (>100), about 10 *C. gigas* sequences, 1 *P. fucata* sequence, and 1 *B. floridae* sequence. Because the 150+ non-redundant GIMAP *C. virginica* transcripts mapped to only 33 *C. gigas* loci, I conclude that some of the diversification (expansion in protein number relative to other organisms) derives from mechanisms different from gene duplication like polymorphism and/or alternative splicing.

Phylogenetic analysis of IFI44 sequences showed two main clusters, one of which was composed solely of bivalve sequences (159 sequences) and the other of which was composed of 18 sequences from 8 diverse organisms including *C. virginica* (Fig. 10 and 11). The main bivalve-only group could be further subdivided into subgroups that had variable numbers of sequences from both *C. virginica* and *C. gigas*.

DISCUSSION

The present study provides a rich view of the processes and genes that constitute the oyster host defense responses to bacterial challenge.

Extracellular matrix (ECM) restructuring, cell adhesion, inflammation, metabolism, catecholamine signaling, and several other processes are key in distinguishing resistance from susceptibility to Roseovarius Oyster Disease.

The present study also identified 8 putative diversified gene families important in the host defenses of Eastern oysters, including two families, GTPase of the immunity associated protein (GIMAP) and interferon-induced protein 44 (IFI44) families, which appear to be expanded in molluscs and bivalves, respectively.

Oysters from the resistant family did not show clinical signs of infection and suffered mortalities comparable to non-challenged oysters, suggesting that these oysters were able to eliminate the pathogen rapidly. Prior to a discussion of differential gene expression and gene family analysis results, it should be noted that only 20% of the transcriptome could be annotated by BLAST and only 30% of the transcriptome could be described by the combination of BLAST similarity, Pfam, and TribeMCL clustering. Conclusions should be regarded with caution and viewed as foundation for future study. With that in mind, comparison of the patterns of gene expression between resistant and susceptible oysters suggest that resistance to *R. crassostreae* may involve a targeted hemocytic response followed by tight control of inflammatory processes and detoxification.

ROD-resistant juvenile oysters responded to the bacterial pathogen *R. crassostreae* mainly by up/down-regulating the expression of transcripts coding for proteins that modify the extracellular matrix, proteins that bind self or non-self ligands, stress proteins, and receptors and/or proteins involved in signaling. The interaction between *R. crassostreae* and its oyster host occurs primarily in the extracellular matrix based on DEG annotations and enriched GO terms (Fig. 4). The unique up-regulation in resistant oysters of several subtilisin-like pro-protein convertases (PPC) (Table 2) suggests the involvement of neuroendocrine signaling and/or host defense-relevant protein processing. PPCs are involved in the processing of von Willebrand Factor, matrix metalloproteinases, and antimicrobial peptides in invertebrates, having multiple downstream effects on cell migration, differentiation, inflammation control, and the restructuring of the ECM (Sato et al. 1996; Parks et al. 2004). The importance of ECM proteolysis and restructuring in the response of resistant oysters to bacterial challenge is corroborated by the up-regulation of ADAMTS8 (Table 2), a matrix metalloproteinase that negatively regulates proliferation and participates in ECM proteolysis (Apte 2004, Feinberg and Weiss 2009), the up-regulation of a fibrinolytic enzyme (Table 2), and the down-regulation of tenascin-xb (Table 3), a glycoprotein involved in wound healing and matrix maturation with anti-adhesive properties (Egging et al. 2007). The multiple transcripts that annotate as tenascin in oysters further dispel the notion that tenascins are unique to chordates (Tucker et al. 2006),

as tenascin-like transcripts have also been found in the transcriptome of the Antarctic bivalve *Laternula elliptica* (Clark et al. 2010). Other transcripts involved in ECM restructuring and cell adhesion identified here as distinguishing resistance and susceptibility include those coding for hemicentin and fibropellin-ia. Hemicentin increases cell adhesion and re-shapes areas of cell contact in *C. elegans* (Vogel and Hedgecock 2001). Fibropellin-ia increases cell adhesion in sea urchin embryos (Burke et al. 1998). By sequence similarity, fibropellin transcripts separated into a cluster enriched for susceptible DEGs and another cluster enriched for resistant DEGs (Table 7). Cell adhesion has long been known to be important in invertebrate innate immunity in general (Johansson 1999) and in oyster immunity (Gueguen et al. 2003). I hypothesize that the differential response of resistant oysters in respect to ECM restructuring and cell adhesion molecules enabled a more effective hemocytic response in these oysters, facilitating cell migration to the extrapallial cavity (the space between the oyster mantle tissue and the shell, where ROD is known to have its primary effects, Paillard et al. 1996; Boardman et al. 2008), likely followed by aggregation, phagocytosis, and apoptosis (Terahara et al. 2005; Anderson et al. 2011; de Lorgeril et al. 2011). Phagocytosis of the bacterial pathogen *Vibrio tapetis* by carpet shell clam (*Ruditapes decussatus*) hemocytes has been shown to have an important role in the resistance of this clam species to Brown Ring Disease (BRD), a disease with many similarities in pathology to ROD (Allam et al. 2001; Allam and Ford

2006). I hypothesize that the cell adhesion transcripts seen in the present study may play a role in the defense capabilities of hemocytes in resistant oysters.

The early resistant response also involved the pro-inflammatory mediator, interleukin 17 (IL17), and the nitric oxide modulator, arginase. Previous research shows that injection of heat-killed gram-positive and gram-negative bacteria into *C. gigas* oysters produces a rapid and transient up-regulation in IL17 transcript abundance in hemocytes, suggesting that IL17 is an important mediator of the pro-inflammatory response in oysters (Roberts et al. 2008). My results support an important role for IL17 in the immune response of oysters against bacterial infection and a potential role in disease resistance to ROD. Here, while IL17 was uniquely up-regulated in resistant oysters, arginase was uniquely down-regulated (Table 3). Arginases have been shown in macrophages to modulate the production of nitric oxide (Chang et al. 1998), which is an immune effector in the Eastern oyster (Villamil et al. 2007). Using microarray technology, a transcript annotating as arginase was shown to increase rapidly after 6 h of heat stress in *C. gigas* (Lang et al. 2009). The down-regulation of arginase in resistant oysters on day 5 may signalize a down-regulation of the inflammation and stress response following a successful defense response.

Genes and processes activated in susceptible oysters in response to bacterial challenge and absent or present to a much lesser degree in resistant

oysters provide potential information on the molecular basis for disease susceptibility or are signs of an unsuccessful defense response. Many transcripts involved in metabolic processes (e.g. carbohydrate metabolism) were differentially expressed in susceptible oysters at both early and late time points following bacterial challenge, but not in resistant oysters as illustrated by enriched molecular function GO terms (Fig. 4). A decrease in energy metabolic enzyme activity and a down-regulation of genes related to energy metabolism have been shown to coincide with mortality events in the Eastern oyster (Genard et al. 2011; 2012).

The up-regulation of multiple dopamine beta-hydroxylase (DBH) transcripts early in both resistant and susceptible oysters (Table 4) suggests the role of catecholamine signaling in host defenses against bacteria. However, both DBH and Dopa decarboxylase (DDC) families/clusters, as well as several transcripts containing the tyrosinase domain were uniquely enriched for susceptible DEGs (Table 7 & 8) suggesting that catecholamine signaling and/or melanization were greater in susceptible oysters. DBH and DDC produce/modify catecholamines, which have been shown to modulate both the immune and stress response in the scallop *Chlamys farreri* (Chen et al. 2008; Zhou et al. 2011a; Zhou et al. 2011b). *Crassostrea gigas* hemocytes have been shown to respond to neuroendocrine signaling with changes in gene expression (Bricelj et al. 1992; Lacosta et al. 2001). It can be imagined that hemocytes at the site of ROD infection and injury are both signaling and

responding to neuroendocrine signals to coordinate a response. DDC, along with enzymes with phenoloxidase activity like tyrosinases, also participates in melanization, a process whose products can kill bacteria (Kan et al. 2008, Sideri et al. 2008) and which may be responsible for the pigmentation characteristic of the conchiolin depositions characteristic of ROD (ref). Both resistant and susceptible oysters appear to use catecholamine signaling to coordinate a response but they may have a differential response in terms of melanization, as supported by the up-regulation of tyrosinase-like transcripts in susceptible oysters. It is not known, however, whether the differentially expressed tyrosinases in the present study had phenoloxidase activity. While tyrosinase has been shown to have phenoloxidase and antibacterial activities in several bivalves including the scallop *Chlamys farreri* (Zhou et al. 2012), Manila clams, and Sydney rock oysters, laccase and catecholase but not tyrosinase were shown to be responsible for phenoloxidase activity in *C. gigas* or *C. virginica* (Luna-Acosta et al. 2011 and references therein). A BLAST search for reciprocal hits between the tyrosinase transcripts in the present transcriptome and the *C. gigas* translated GLEAN gene models revealed 24 tyrosinase-like peptides therein, providing evidence that *C. gigas*, too, has tyrosinase proteins.

The host defense response to bacterial challenge in resistant and susceptible oysters shared some commonalities, including the involvement of catecholamine signaling (discussed above), detoxification, and apoptosis.

Transcripts that were highly up-regulated in both resistant and susceptible oysters included transcripts annotating as glutathione s-transferase, cytochrome p450, and heat shock protein 60 (HSP60), which are involved in preventing oxidative damage (Table 4). Glutathione s-transferase is an anti-oxidant and is up-regulated in hemocytes of oysters challenged with a pathogenic *Vibrio* sp. (de Lorgeril et al. 2011). Although cytochrome p450s have been best studied in detoxification of xenobiotics in bivalves (Snyder 2000), they have also been implicated in the host defense response of the flat oyster *Ostrea edulis* to the parasite *Bonamia ostreae* (Morga et al. 2011) and the clam *Ruditapes phillipinarum* to *Vibrio tapetis* (Brulle et al. 2012). HSP60 is involved in xenobiotic detoxification and the stress response in oysters (Ivanina et al. 2008). Generally, detoxification was intensified in susceptible oysters as related detoxification terms like “monooxygenase activity” and “oxidoreductase activity” were functionally enriched among the susceptible but not the resistant DEGs (Fig. 4). Detoxification-related transcripts are highly up-regulated in *C. virginica* prior to mass mortality events (Genard et al. 2012). While an early response of detoxification/stress transcripts may contribute to resistance, a persistent and more generalized response (that is, a greater number of up-regulated stress transcripts) may signalize imminent mortality. Another process involved in both the resistant and susceptible response was apoptosis. Inhibitor of apoptosis (IAP) transcripts were found to be both up- and down-regulated among resistant oyster DEGs and largely down-regulated

among susceptible oyster DEGs. IAP proteins are associated with molluscan immunity (Sokolova 2009), participating in the defense response of clams to BRD (Donaghy et al. 2009; Brulle et al. 2012), so they likely play a role in the oyster response to ROD.

Transcripts uniquely expressed in susceptible oysters (F3L) at both early and late time points may reveal genes linked to stress or mortality, since this family suffered consistent levels of mortality throughout the challenge. The most significantly enriched Pfam family among late susceptible oyster DEGs was ankyrin Ank_2 (Table 6). The ankyrin domain serves as a mediator of protein-protein interactions in proteins with a wide variety of functions (Bennett and Baines 2001). Little is known about the function of ankyrin transcripts, which includes ankyrin unc-44, nacht and ankyrin domain containing, and other ankyrin repeat-containing transcripts, in oysters. Ankyrin repeat-containing proteins have been shown to be differentially expressed in *C. gigas* as a result of heat shock, exposure to *Vibrio*, and between resistant and susceptible *C. gigas* to summer mortality (Lang et al. 2009; de Lorgeril et al. 2011; Fleury et al. 2012). Further research is necessary to elucidate the role of these proteins in oyster immunity.

Studies in multiple marine invertebrates suggest that diversified groups of receptors, regulators, and/or effectors enable these organisms to meet the challenge of counteracting pathogens and parasites with relatively short generation times and high mutation rates, without the adaptability of the

adaptive immune system (Messier-Solek et al. 2010). Two gene family analysis approaches, Pfam annotation and TribeMCL clustering, followed by enrichment testing for DEGs enabled the identification of putative “diversified” gene families. Together the two approaches covered some 96K contigs (27% of transcriptome) at the protein/gene family level. Pfam annotated 13K contigs not clustered by TribeMCL and TribeMCL clustered 50K contigs not annotated by Pfam and both approaches covered 38K contigs not annotated using BLAST. When used to describe differential expression at the gene family and domain level (in the case of Pfam domains) both the Pfam and TribeMCL approaches largely complimented one another and helped to distinguish the resistant from susceptible responses. The Pfam approach was more specific and automatically included annotation, yet it described a smaller portion of the transcriptome and was of no use for previously unannotated or shallowly annotated gene families. While TribeMCL clusters sometimes contained false positives (likely because of domain sharing), these were filtered when examining specific families. Because TribeMCL clustering was unbiased and more sensitive, it was used as the primary technique for the identification of enumeration of the members of 8 potentially diversified gene families of immune relevance. Some candidate diversified gene families, like C1q domain-containing proteins spanned multiple TribeMCL clusters.

The abundance and diversity of serine protease (SP) and serine protease inhibitor (SPI) transcripts, combined with the observed patterns of

differential expression, suggest a role of these gene families as diversified effectors of oyster host defense to bacterial challenge. The role of SPIs in oyster host defense has long been recognized, although previous studies mainly focused on the role of these molecules in the interplay between the Eastern oyster and the protozoan parasite *Perkinsus marinus*. The hemolymph of *Crassostrea* spp. contains effective protease inhibitors and *C. gigas*, naturally resistant to *P. marinus*, has significantly greater inhibitory activity than *C. virginica*, which has been interpreted as suggestive of the role of protease inhibitors in host defense and resistance (Faisal et al. 1998; Jenny et al. 2002). The list of known oyster SPs and SPIs has grown with each EST analysis (Gueguen et al. 2003; Roberts et al. 2008) and several SPIs have been biochemically characterized (Xue et al. 2006; Xue et al. 2009; La Peyre et al. 2010). Polymorphism in the promoter of an Eastern oyster SPI has been associated with disease resistance to *P. marinus* (Yu et al. 2011). Proteases from *Perkinsus* sp. inhibit phagocytosis of *Vibrio tapetis* in clams (Ordas et al. 1999) and the virulent effects of *V. tapetis* on clam hemocytes are consistent with the effects of bacterial proteases (Borrego et al. 1996; Allam and Ford 2006). I hypothesize that SPIs can neutralize *R. crassostreae* proteases. Less work has been conducted on the role of SPs in oyster host defense, yet SPs may take part in both digestion and host defense, as in *Drosophila* spp. (Ross et al. 2003). Here, none of the *C. virginica* SPs annotated as clip domain SPs, which are important in insect immunity. However, because the clip domain is

N-terminal to the chymotrypsin domain, true clip domain SP transcripts in the transcriptome may have had their clip domains truncated. A clip domain SP has been found in the scallop *C. farreri* (Zhu et al. 2008) and the pearl oyster *P. fucata* (Zhang et al. 2009). I speculate that at least some of the SPs identified here are truly clip domain SPs. More work will need to be done to definitively describe the domain architecture of Eastern oyster SPs and their role in the host defense response.

Another putative diversified gene family herein identified as abundant and differentially expressed in response to bacterial challenge was the fibrinogen domain-containing gene family. The patterns of expression of members of the FBG domain-containing family in oysters in response to bacterial challenge differed between resistant and susceptible oysters, suggesting a potential role in disease resistance/susceptibility. Invertebrate fibrinogen domain-containing proteins –all of which have a C-terminal fibrinogen (FBG) domain– include fibrinogen-related proteins (FREPs), which contain one or two N-terminal IgSF domains; fibrinogen-related molecules (FREMs), which contain epidermal growth factor-like repeats; and fibrinogen-related domain-containing (FREDs) which includes all FBG domain-containing proteins that do not fit into the other aforementioned groups. These proteins function in pathogen recognition, agglutination, and parasite resistance (Hanington and Zhang 2011). FBG domain-containing proteins, particularly FREPs, have been studied in depth in the gastropod *B. glabrata*, perhaps

partially because of the surprising finding that FREP genes can diversify somatically (Zhang et al. 2004). FREPs have been shown in *B. glabrata* to contribute to resistance against the parasite *Schistosoma mansoni* (Hertel et al. 2005). A FREP in the bay scallop *Argopecten irradians* has been shown to have agglutinating activity against chicken and human erythrocytes and bacteria and to increase in expression following challenge by gram-negative bacteria (Zhang et al. 2009). No IgSF domains were found in the identified FBG domain-containing transcripts in the Eastern oyster, while one potential FREM-like transcript with five EGF-like repeats was identified. This does not mean that FREPs are absent from the oyster transcriptome –as 190 FREPs were recently found in the *C. gigas* genome (Zhang et al. 2012)– but underscores the 3'-bias of cDNA sequencing following poly-A capture and the limitation of annotation with transcripts that are rarely full-length.

My results also suggest a role of the gene families DMBT1/SR type 12, C1qDC, and CTLDC in the oyster host defense response. Scavenger receptors with class B SRCR domains have undergone expansions in *S. purpuratus* and show differential expression following challenge by fungi and bacteria (Pancer 2000; Hibino et al. 2006). Scavenger receptors have not undergone comparable expansion/diversification in *D. melanogaster*, *C. intestinalis*, nor in *C. elegans* (Hibino et al. 2006). Scavenger receptor diversity in lophotrochozoans has yet to be adequately addressed. These proteins are differentially expressed in oyster species in response to summer

mortality (Huvel et al. 2004; Fleury et al. 2010) and hypoxia (David et al. 2005). Recently, a SR protein has been characterized in the scallop *C. farreri* that is up-regulated significantly by exposure to PAMPs like LPS, peptidoglycan and β -glucan and can bind LPS and peptidoglycan (Liu et al. 2011). The discovery here of tens to hundreds of non-redundant sequences in each of several oyster species and the differential expression of many SR sequences in response to bacterial challenge in *C. virginica* suggests that SR proteins may play a role in oysters comparable to that in *C. farreri* and with diversity, as a gene family, that may approach that of *S. purpuratus*. Several oyster transcripts also annotated to class A and F SRs, the latter of which consist mostly of annotations to cell death abnormality-1, which has been linked in *C. elegans* to the unfolded protein response, apoptotic cell debris engulfment, and resistance to at least one species of bacteria (Lamitina and Cherry, 2008; Haskins et al. 2008). While apparently very abundant in the transcriptome, class F SRs contain epidermal growth factor-like (egf-like) domains present in serial repetitions that frustrate similarity clustering by TribeMCL. Enrichment for DEGs was a prerequisite for consideration of a gene family as relevant to immunity. Adulteration of class F SR clusters with other multiple egf-like domain-containing proteins thwarted enrichment testing and might have possibly precluded definitive enumeration of class F SRs, which illustrates the limitations of similarity clustering, especially when using transcripts assembled from short reads in the absence of a reference genome.

The most abundant putative diversified family studied was the C1qDC gene family. C1qDC proteins can participate in self and non-self binding and function in a variety of processes such as agglutination, cell adhesion, inflammation, and clearance of apoptotic bodies (Kishore et al. 2004). The high sequence variability of C1qDC transcripts in *C. virginica* has also been shown in transcripts in another bivalve, the mussel *M. galloprovincialis* (Gerdol et al. 2011). C1qDC transcripts are up-regulated in *M. galloprovincialis* following gram-positive and gram-negative bacterial challenge and show highest tissue-specific expression in hemocytes (Gerdol et al. 2011). The role of C1qDC proteins as pattern recognition receptors (PRRs) has been solidified by a demonstration of the ability of a recombinant C1qDC protein from the scallop *Argopecten irradians* to bind PAMPs from diverse pathogens including gram-negative and gram-positive bacteria and fungi (Kong et al. 2010). Results of the present study suggest that the Eastern oyster expresses a great number of C1qDC proteins and that they play a role in host defense against gram-negative bacteria. While > 100 non-redundant translated transcripts/proteins were found by network similarity clustering in all three oyster species examined, only a handful of transcripts were found in gastropod species (Fig. 5). These observations are consistent with the hypothesis that C1qDC genes likely expanded in bivalves, independent from the expansion in the chordate lineage (Gerdol et al. 2011).

Another abundant putative diversified gene family consisted of CTL domain-containing (CTLDC) transcripts. CTLDC proteins are extracellular proteins that contain conserved carbohydrate recognition CTL or CRD domains and function in processes as diverse as cell adhesion, endocytosis, activation of antimicrobials, and pathogen recognition and agglutination (Weise et al. 2006). In *C. elegans*, a total of 278 diversified CTLDC genes have been identified, some of which show differential expression upon pathogen challenge (Schulenburg et al. 2008). Recently a CTLDC from *C. farreri* was shown to act as a PRR, binding LPS and β -glucan, and as an opsonin, enhancing the phagocytic capabilities of *C. farreri* hemocytes (Yang et al. 2011). CTLDC proteins may play a role as diversified PRRs and/or activators of the host defense response of the Eastern oyster to gram-negative bacteria.

The above three groups of proteins, DMBT1/SR type 12, C1qDC, and CTLDC were demonstrably diversified relative to other taxa, yet showed differential expression patterns in oysters in response to bacterial challenge that were not easily interpretable. For each group, while some transcripts were up-regulated following challenge, other transcripts were down-regulated. Several explanations for this apparent lack of consistency include dynamic expression regimes, as demonstrated in sea urchin SRs (Pancer 2000), complex/compensatory regulation as demonstrated in mussel C1qDC transcripts (Gerdol et al. 2011), variability of function within gene family, high

polymorphic rate, basal expression variability, which is high in many *C. gigas* genes (Rosa et al. 2011), and the intrinsic difficulty of mapping reads to a large gene family in a *de novo*-assembled transcriptome.

Interestingly, I have identified two families that have not been previously identified as diversified in other studies. To my knowledge, the present study is the first to identify GIMAP genes in invertebrates as diversified mediators of the invertebrate immune response. Previously, similarity searches for GIMAP genes within the genomes of fission yeast *Saccharomyces pombe* and brewer's yeast *S. cerevisiae*, *C. elegans*, and *D. melanogaster* turned up no hits, and consequently, it was concluded that GIMAP genes are present only in vertebrates and angiosperms (Filen and Lahesmaa 2010). While no GIMAP proteins were found in the proteomes of two protostomes herein studied, *D. melanogaster* and *D. pulex*, GIMAP sequences were plentiful in molluscs (Fig. 5). My study advances the importance of including lophotrochozoans in genomic surveys for genes of interest, now enabled by the recent release of the Pacific oyster genome (Zhang et al. 2012). Phylogenetic analysis of GIMAP sequences supports the possibility of several gene expansion events (with an expanded set of GIMAP sequences from basal chordates and molluscs grouping together and expanded sets of GIMAP sequences unique to bivalves/molluscs) likely combined with diversification through alternative mechanisms (e.g. alternative splicing/INDELs/allelic variation/somatic diversification) (Fig. 9). In vertebrates, GIMAP proteins have been best

characterized in their role as regulators of apoptosis (Nitta et al. 2007), though it has been shown that GIMAP family members show differential expression patterns across tissue types and may serve varying functions at different times and in different tissues (Wang and Li 2009). Exposure of human monocytes to LPS induces the down-regulation of 28 genes by >4 fold, four of which are GIMAP proteins (Dower et al. 2008). It has been suggested that the down-regulation of GIMAP proteins in humans may serve to promote the survival of monocytes by negatively regulating apoptosis (Dower et al. 2008). It may be that GIMAP proteins fulfill a parallel role in oyster hemocytes though further work will be needed to define this role.

Another gene family proposed as a novel diversified mediator of the oyster immune response is the IFI44 gene family. IFI44, inducible by interferon- α , is implicated in antiviral host defense (Kitamura et al. 1994) and shows antiproliferative activity, possibly by contributing to cell cycle arrest (Hallen et al. 2007). IFI44 transcripts are up-regulated in *C. gigas* in response to challenge with the virus OsHV-1 (Renault et al. 2011) and a pathogenic *Vibrio* (de Lorgeril et al. 2011). Phylogenetic analysis of IFI44 sequences from diverse taxa supports the hypothesis of a bivalve-only gene expansion(s) likely combined with diversification through alternative mechanisms (e.g. alternative splicing/INDELs/allelic variation/somatic diversification) (Fig. 11). Further expression profiling on a finer time scale and in a variety of conditions may

help to determine why the IFI44 gene family has become diversified in bivalves and what specific challenges induce IFI44 expression.

Many unannotated gene families were certainly involved in the defense and/or stress response. Eighty-six of the 187 TribeMCL clusters that were found to be enriched for DEGs could not be annotated. The inability to annotate a great portion of the transcriptome (in this case 80% without BLAST similarity) remains a challenge in describing the oyster host defense response. I show here that similarity clustering does offer the means of transferring annotations. The majority of clusters that were herein annotated included one or more transcripts that could not be annotated by BLAST alone, yet whose identity could be inferred from its neighbors. Moreover, the very process of reducing a set of transcripts to a smaller set of connected components has the promise to focus efforts and resources in the effort to characterize genes and gene families that presently cannot be annotated by similarity search to the public databases. While these clusters could not be used to describe the oyster host defense response here, TribeMCL or an analogous clustering technique could be used to facilitate annotation in future studies. By considering transcripts at the level of TribeMCL clusters, sequence similarity motifs may be extracted to aid eventual characterization.

While much work remains to be done in characterizing the present Eastern oyster transcriptome and describing the oyster host defense response to bacterial challenge, the present study has made great advances to these

ends. When ROD-resistant and ROD-susceptible oysters were exposed *R. crassostreae* and gene expression was compared throughout the challenge by high-throughput cDNA sequencing, several processes emerged as key to resistance including ECM remodeling, cell adhesion, and inflammation. The present study has generated a pool of candidate disease resistance genes for advanced genotypic selection regimes. Additionally, several gene families were identified as putatively diversified and of immune relevance in the Eastern oyster, two of which, IFI44 and GIMAP families, are of especial interest as expansions were found to be specific to bivalves and molluscs, respectively. Transcript translation and similarity clustering followed by gene family analysis should prove useful in describing the transcriptomes of other invertebrates in response to immune and/or stress challenge as an unbiased means of identifying putatively diversified groups of host receptors, regulators, and effectors.

TABLES

Table 1. Assembly metrics for transcriptome assembly

Number of contigs	356237
Total span (bp)	156920694
Number of contigs (> 1Kb)	22934
Max Contig Length (bp)	16256
Mean Contig Length (bp)	440
N50 (bp)	487
Number of contigs with BLAST hits*	70621
% of contigs with BLAST hits*	19.8

*Contigs compared to NCBI's non-redundant protein database using BLASTX, significant hits retained with e-value $\leq 1e-06$.

Table 2. Top 50 most highly differentially expressed genes (DEGs) unique to GX_early up-regulated relative to control pool (CGX_late)

Contig	Reg. GX-1d	Reg. GX-5d	logFC	log CPM	p-value	Best blastx hit to nr db	Hit accession	Hit e-value
comp17501_c0_seq1	up	-	6.1	4.55	2.01E-14	-	-	-
comp5950_c0_seq1	up	up	5.21	3.76	3.51E-12	-	-	-
comp4755_c0_seq2	up	-	5.13	2.95	1.03E-10	-	-	-
comp2875_c0_seq4	up	up	4.41	2.89	2.50E-09	-	-	-
comp1023_c0_seq2	up	up	4.28	3.85	5.46E-09	scavenger receptor cysteine-rich	ACT53266	1.30E-15
comp24124_c0_seq1	-	up	4.49	2.59	9.31E-09	ched related family member (ptr-19)	XP_002734100	1.39E-138
comp12059_c0_seq1	up	up	4.18	4.67	1.46E-08	-	-	-
comp1165_c0_seq2	-	up	3.99	4.75	1.48E-08	-	-	-
comp2015_c0_seq24	up	-	3.95	2.67	8.73E-08	inhibitor of apoptosis	AEB54800	4.91E-09
comp1023_c0_seq5	up	up	3.95	3.23	9.03E-08	-	-	-
comp2870_c0_seq2	up	up	3.79	4.29	1.07E-07	-	-	-
comp31762_c0_seq1	up	up	3.73	2.96	1.20E-07	-	-	-
comp657_c0_seq3	-	up	3.57	3.74	2.18E-07	-	-	-
comp3858_c0_seq5	up	-	3.63	2.82	6.54E-07	isoleucyl-trna synthetase	NP_001090690	2.18E-59
comp9303_c0_seq3	-	up	3.67	1.49	8.37E-07	-	-	-
comp3628_c0_seq2	up	-	3.62	3.14	8.53E-07	-	-	-
comp7475_c2_seq3	up	-	3.66	1.75	8.71E-07	-	-	-
comp18756_c0_seq2	up	-	3.83	2.06	9.27E-07	sushi repeat-containing	XP_002664481	2.19E-22
comp20853_c1_seq3	-	up	3.57	2.16	1.19E-06	af397902_1egf-like	XP_002601693	1.28E-35
comp6834_c0_seq1	-	up	3.75	2	1.42E-06	Protein	XP_002592396	8.35E-12
comp6161_c0_seq5	up	-	3.49	3.24	1.57E-06	type 2 proinsulin processing endopeptidase	2206277A	2.33E-42
comp1023_c0_seq1	up	up	3.53	3.08	1.67E-06	-	-	-

comp21651_c0_seq1	up	-	3.58	2.33	2.23E-06	-	-	-
comp4755_c0_seq1	up	-	3.52	2.56	2.40E-06	-	-	-
comp1316_c0_seq1	-	up	3.26	4.6	2.79E-06	-	-	-
comp1157_c0_seq3	up	up	3.34	2.72	3.28E-06	-	-	-
comp4626_c0_seq4	up	-	3.34	2.55	3.32E-06	alpha-ketoglutarate-dependent hypophosphite dioxygenase-like	XP_002944900	1.40E-10
comp18756_c0_seq3	up	-	3.48	2.61	4.03E-06	fibropellin ia	XP_002601363	2.21E-13
comp24671_c0_seq1	-	up	3.57	1.99	4.38E-06	-	-	-
comp5314_c0_seq1	up	up	3.36	3.57	4.85E-06	-	-	-
comp1788_c0_seq4	-	up	3.26	2.21	5.26E-06	fibrinolytic enzyme	CAA64472	5.10E-12
comp22438_c0_seq1	-	up	3.26	2.81	5.49E-06	cytochrome p450 family 4	ACM16804	4.94E-106
comp2015_c0_seq17	up	-	3.23	3.14	6.25E-06	-	-	-
comp7137_c0_seq2	-	up	3.26	3.02	7.06E-06	organic solute transporter subunit alpha	XP_002732822	4.93E-20
comp6713_c0_seq2	up	up	3.02	2.8	8.59E-06	-	-	-
comp1506_c0_seq4	up	up	3.08	5.47	1.01E-05	a disintegrin and metalloproteinase with thrombospondin motifs 8	XP_002940685	2.89E-12
comp6837_c0_seq1	up	-	3.15	4.45	1.16E-05	interleukin 17d	A9XE49	1.05E-56
comp18756_c0_seq5	up	-	3.27	3.01	1.18E-05	fibropellin ia	XP_002599260	2.50E-27
comp3628_c0_seq3	up	-	3.34	2.03	1.18E-05	-	-	-
comp14520_c0_seq2	up	-	3.21	2.31	1.26E-05	-	-	-
comp274_c0_seq1	-	up	3	6.82	1.29E-05	-	-	-
comp4755_c0_seq3	up	-	3.1	4.57	1.34E-05	hypothetical protein	ACU33972	3.51E-35
comp15440_c0_seq1	up	-	3.18	1.8	1.34E-05	inhibitor of apoptosis	XP_002426441	1.01E-13
comp11365_c0_seq2	-	up	2.93	2.83	1.36E-05	-	-	-
comp664_c0_seq5	up	up	3.26	2.38	1.43E-05	Protein	XP_001642030	4.44E-10

comp28783_c0_seq1	-	up	3.1	2.08	1.52E-05	BRAFLDRAFT_129258	XP_002612219	1.03E-31
comp6161_c0_seq11	up	-	3.1	3.16	1.59E-05	furin	AAA49718	1.42E-22
comp9269_c0_seq3	up	-	3.2	2.3	1.76E-05	-	-	-
comp1023_c0_seq3	up	up	3.07	3.85	1.98E-05	scavenger receptor cysteine-rich	XP_001622238	4.33E-27
comp22224_c0_seq1	-	up	3.07	2.07	2.01E-05	-	-	-

The top fifty most differentially expressed and up-regulated genes unique to GX days 1 and 5 (genes not differentially expressed in any F3L treatment) are shown, ranked by false discovery rate-adjusted p-value. Magnitude of differential expression is expressed as log-10 fold change over control pool (logFC) and abundance is expressed as log-10 counts per million (logCPM). Regulation, or “Reg.”, (up- or down-regulated) is shown for each contig. Hyphen (-) indicates that the contig is not differentially expressed at that timepoint (Reg. columns) or that the contig does not have BLASTX hit to the NCBI non-redundant protein database with associated e-value $\leq 1e-6$ (annotation columns). Where contigs are differentially expressed in both timepoints, logFC, logCPM, and p-value correspond to the timepoint in which the contig was most highly differentially expressed and that timepoint is indicated in the regulation columns in bold.

Table 3. Differentially expressed genes (DEGs) unique to GX_early down-regulated relative to control pool

Contig	Reg. GX-1d	Reg. GX-5d	logFC	log CPM	p-value	Best blastx hit to nr db	Hit accession	Hit e-value
comp24428_c0_seq1	down	down	-8.64	1.68	6.67E-07	rapunzel 5	NP_0011 03594	2.10E-10
comp1572_c0_seq4	-	down	-8.67	1.53	1.31E-06	-	-	-
comp5722_c1_seq2	-	down	-4.84	2.94	1.91E-06	collagen alpha	XP_0015 12734	2.87E-32
comp14853_c0_seq2	down	down	-8.29	1.28	3.49E-06	-	-	-
comp1572_c0_seq3	-	down	-4.34	4.24	5.84E-06	-	-	-
comp7972_c0_seq1	-	down	-8.09	1.08	7.32E-06	cubilin	XP_0027 34392	0.00E+0 0
comp869_c0_seq4	down	-	-8.02	0.95	1.25E-05	-	-	-
comp1572_c0_seq9	-	down	-4.64	1.93	2.13E-05	-	-	-
comp11408_c0_seq2	-	down	-7.89	0.77	2.30E-05	-	-	-
comp1285_c1_seq8	-	down	-3.9	3.84	2.38E-05	arginase type i-like	AEB7096 5	5.87E-28
comp3240_c1_seq2	-	down	-4.27	1.99	4.15E-05	-	-	-
comp25746_c0_seq4	-	down	-7.6	0.55	5.20E-05	tenascin xb	XP_0027 41293	4.52E-38
comp19167_c0_seq1	-	down	-4.79	1.39	5.95E-05	-	-	-
comp34093_c0_seq1	-	down	-7.54	0.48	6.63E-05	-	-	-
comp22172_c0_seq3	down	-	-7.57	0.44	7.22E-05	-	-	-
comp810_c1_seq1	-	down	-3.4	6.65	8.89E-05	heat shock protein 22	ACU832 31	2.79E-28
comp1285_c1_seq3	-	down	-3.54	3.50	9.56E-05	arginase ii	XP_0021 30834	6.53E-12
comp39520_c0_seq1	-	down	-4.31	1.47	1.30E-04	polyprote in	XP_0027 40782	0.00E+0 0
comp10161_c0_seq1	down	-	-3.47	3.97	1.41E-04	-	-	-
comp9135_c0_seq1	-	down	-3.54	3.17	1.59E-04	-	-	-
comp1190_c0_seq7	down	-	-3.44	3.60	2.04E-04	-	-	-
comp38620_c0_seq1	-	down	-7.26	0.16	2.14E-04	-	-	-

comp2015_c0_seq13	down	-	-7.19	0.12	2.29E-04	inhibitor of apoptosis	AEB54799	1.28E-41
comp43138_c0_seq1	down	-	-7.22	0.12	2.29E-04	-	-	-
comp5608_c0_seq1	-	down	-3.35	3.50	2.31E-04	c-type lectin	ABB71672	7.01E-16
comp16886_c0_seq1	-	down	-7.17	0.11	2.36E-04	-	-	-
comp18757_c0_seq1	-	down	-7.14	0.09	2.36E-04	hla-b associated transcript 1	XP_003217350	6.91E-21
comp1190_c0_seq5	down	down	-3.25	4.38	2.48E-04	-	-	-
comp7972_c0_seq4	-	down	-3.11	4.03	3.13E-04	cubilin	XP_002612977	0.00E+00
comp7814_c1_seq2	-	down	-7	0.04	3.58E-04	-	-	-
comp16058_c0_seq1	-	down	-3.9	1.18	3.76E-04	-	-	-
comp24625_c0_seq1	down	-	-7.09	0.03	3.81E-04	-	-	-
comp5396_c0_seq1	-	down	-3.01	3.96	4.75E-04	melatonin receptor 1a	ADM73175	1.28E-66
comp27900_c0_seq1	-	down	-4.2	0.79	5.18E-04	-	-	-
comp1190_c0_seq8	down	-	-3.1	3.94	5.22E-04	-	-	-
comp16567_c0_seq1	-	down	-4.58	0.39	5.85E-04	-	-	-
comp5722_c1_seq1	-	down	-2.86	4.28	6.15E-04	collagen alpha	XP_001512734	3.63E-32
comp13269_c0_seq1	-	down	-3.06	3.19	6.31E-04	DAPPUDRAFT_309315	EFX70737	1.87E-17
comp18902_c0_seq1	-	down	-3.43	1.62	6.49E-04	rho gtpase	XP_002739105	1.37E-58
comp17170_c0_seq1	-	down	-3.42	1.63	6.75E-04	-	-	-
comp19167_c0_seq2	-	down	-3.48	1.42	7.44E-04	-	-	-
comp1285_c1_seq1	-	down	-4.44	0.25	9.04E-04	-	-	-
comp7712_c0_seq2	-	down	-3.36	1.33	1.07E-03	-	-	-
comp14853_c0_seq6	-	down	-3.34	1.12	1.09E-03	-	-	-

comp216_c0_seq3	-	down	-2.84	4.41	1.13E-03	-	-	-
comp8760_c0_seq2	-	down	-3.14	1.47	1.19E-03	-	-	-

Differentially expressed contigs unique to GX days 1 and 5 (genes not differentially expressed in any F3L treatment) that were down-regulated are shown, ranked by false discovery rate-adjusted p-value. Magnitude of differential expression is expressed as log-10 fold change over control pool (logFC) and abundance is expressed as log-10 counts per million (logCPM). Regulation, or “Reg.”, (up- or down-regulated) is shown for each contig. (-) indicates that the contig is not differentially expressed at that timepoint (Reg. columns) or that the contig does not have BLASTX hit to the NCBI non-redundant protein database with associated e-value $\leq 1e-6$. Where contigs are differentially expressed in both timepoints, logFC, logCPM, and p-value correspond to the timepoint in which the contig was most highly differentially expressed and that timepoint is indicated in the regulation columns in bold.

Table 4. Top 50 most highly differentially expressed genes (DEGs) among GX_early shared with F3L_early and/or F3L_late up-regulated relative to control pool

Contig	Reg. GX- 1d	Reg. GX- 5d	logFC	log CPM	p-value	Best blastx hit to nr db	Hit accession	Hit e-value
comp866_c0_seq1	-	up	5.43	4.01	1.11E-12	BRAFLDRAFT_227853	XP_002612894	6.68E-31
comp619_c0_seq1	up	up	4.84	5.8	2.34E-11	serine protease inhibitor cvsi-2	B9A8D7	1.55E-11
comp1523_c0_seq2	-	up	4.67	4.37	5.40E-11	-	-	-
comp2870_c0_seq3	up	up	4.82	3.88	1.19E-10	predicted protein	XP_001632962	3.82E-17
comp2870_c0_seq1	up	up	4.34	5.13	1.97E-09	hypothetical protein	XP_002416173	2.97E-59
comp11276_c0_seq3	-	up	4.67	2.68	2.88E-09	Polyketide synthase pks2	XP_002734101	3.26E-44
comp985_c0_seq1	-	up	4.34	4.38	3.18E-09	-	-	-
comp3562_c1_seq1	up	-	4.27	3.64	3.21E-09	-	-	-
comp631_c0_seq2	-	up	4.09	5.3	3.99E-09	dopamine beta hydroxylase- like	XP_002117559	1.97E-08
comp9303_c0_seq5	-	up	4.17	2.48	9.67E-09	omega class glutathione s- transferase	CAD89618	3.39E-13
comp1165_c0_seq1	-	up	4.07	5.51	9.80E-09	-	-	-
comp928_c0_seq1	-	up	3.92	6.29	1.93E-08	serine protease inhibitor cvsi-2	B9A8D7	3.21E-12
comp300_c0_seq1	-	up	3.87	8.08	2.17E-08	-	-	-
comp1165_c0_seq3	-	up	3.89	6.63	3.08E-08	-	-	-
comp1893_c0_seq1	up	up	3.95	4.01	4.61E-08	dopamine beta hydroxylase- like	AAS92605	3.53E-27
comp2875_c0_seq2	-	up	3.9	3.61	5.23E-08	serine protease	XP_002593726	2.71E-08
comp8625_c0_seq1	up	-	4.04	3.44	6.69E-08	fatty acid synthase-like	ACZ55138	0.00E+0 0
comp298_c0_seq1	up	up	3.74	7.91	8.70E-08	-	-	-

comp2550_c0_seq1	up	up	3.83	4.73	9.59E-08	-	-	-
comp3607_c0_seq3	-	up	3.88	3.27	1.09E-07	arylsulfatase a-like	XP_002607295	1.00E-12
comp11276_c0_seq6	-	up	4.12	2.19	1.48E-07	-	-	-
comp631_c0_seq1	-	up	3.58	6.81	1.60E-07	dopamine beta hydroxylase- like	XP_002117561	2.29E-31
comp1880_c0_seq2	up	up	3.85	3.59	1.62E-07	-	-	-
comp281_c1_seq1	-	up	3.54	8.32	1.78E-07	-	-	-
comp7186_c1_seq4	up	-	3.92	2.39	2.12E-07	dna damage- regulated autophagy modulator protein 2	NP_001230625	1.82E-14
comp1199_c0_seq1	up	-	3.74	4.66	2.23E-07	-	-	-
comp335_c0_seq1	-	up	3.6	5.15	3.26E-07	-	-	-
comp12125_c0_seq1	-	up	3.49	3.4	4.15E-07	-	-	-
comp1479_c0_seq1	up	up	3.67	3.44	5.43E-07	-	-	-
comp1199_c1_seq2	up	-	3.57	5.05	6.03E-07	-	-	-
comp1853_c0_seq4	-	up	3.57	3.25	7.21E-07	-	-	-
comp985_c0_seq2	-	up	3.51	4.9	7.43E-07	-	-	-
comp7828_c0_seq1	up	-	3.63	2.6	7.76E-07	-	-	-
comp1199_c1_seq3	up	-	3.52	5.31	9.49E-07	-	-	-
comp1037_c0_seq1	-	up	3.27	6.27	1.28E-06	-	-	-
comp985_c0_seq3	-	up	3.39	3.06	1.38E-06	-	-	-
comp88_c0_seq1	up	up	3.4	2.85	1.79E-06	-	-	-
comp88_c0_seq2	up	up	3.4	2.85	1.79E-06	-	-	-
comp88_c0_seq3	up	up	3.4	2.85	1.79E-06	-	-	-
comp88_c0_seq4	up	up	3.4	2.85	1.79E-06	-	-	-
comp1197_c1_seq1	-	up	3.4	5.68	2.07E-06	-	-	-
comp887_c0_seq1	up	up	3.32	5.73	2.13E-06	c1q domain containing protein 1q13	CBX41662	1.54E-07
comp437_c0_seq1	up	up	3.41	5.69	2.20E-06	-	-	-
comp437_c0_seq2	up	up	3.35	5.33	2.30E-06	-	-	-

comp1229_c0_seq1	-	up	3.16	4.59	2.59E-06	-	-	-
comp11715_c0_seq1	-	up	3.4	2.6	2.63E-06	-	-	-
comp4943_c0_seq1	-	up	3.35	3.49	3.04E-06	Cytochrome p450	XP_002594971	1.92E-58
comp2451_c0_seq15	-	up	3.59	1.94	3.19E-06	galactosamine (n-acetyl)-6-sulfate sulfatase-like	XP_002605064	5.12E-16
comp3498_c0_seq2	up	-	3.18	4.32	3.45E-06	heat shock protein 60	ABN11936	7.59E-83
comp186_c0_seq1	up	up	3.25	5.44	3.57E-06	-	-	-

The top fifty most differentially expressed and up-regulated genes shared between GX days 1 and 5 and F3L days 1, 5, 15, and/or 30 are shown, ranked by false discovery rate-adjusted p-value. Magnitude of differential expression is expressed as log-10 fold change over control pool (logFC) and abundance is expressed as log-10 counts per million (logCPM). Regulation, or “Reg.”, (up- or down-regulated) is shown for each contig. Hyphen (-) indicates that the contig is not differentially expressed at that timepoint (Reg. columns) or that the contig does not have BLASTX hit to the NCBI non-redundant protein database with associated e-value $\leq 1e-6$ (annotation columns). Where contigs are differentially expressed in both timepoints, logFC, logCPM, and p-value correspond to the timepoint in which the contig was most highly differentially expressed and that timepoint is indicated in the regulation columns in bold.

Table 5. Top 50 most highly differentially expressed genes (DEGs) among GX_early DEGs shared with F3L_early and/or F3L_late down-regulated relative to control pool

Contig	Reg. GX-1d	Reg. GX-5d	logFC	log CPM	p-value	Best blastx hit to nr db	Hit accession	Hit e-value
comp12483_c0_seq1	down	down	-11.19	4.06	3.29E-11	c1q domain containing protein 1q13	CBX41662	4.49E-09
comp1102_c0_seq1	-	down	-7.59	4.92	7.33E-11	-	-	-
comp9636_c0_seq2	-	down	-10.66	3.55	2.99E-10	-	-	-
comp1102_c0_seq3	-	down	-6.92	5.29	3.29E-10	collagen alpha-5 chain	XP_002595170	1.40E-08
comp6091_c0_seq2	-	down	-10.58	3.45	4.66E-10	camp responsive element binding 2	AAU93879	3.04E-18
comp14386_c0_seq1	-	down	-7.3	4.42	6.49E-10	-	-	-
comp3971_c0_seq2	-	down	-6.02	5.04	7.78E-09	monocarboxylate transporter 14	EGI68511	5.46E-27
comp9636_c0_seq6	-	down	-9.83	2.70	1.11E-08	-	-	-
comp8174_c0_seq1	down	-	-6.45	3.28	3.28E-08	-	-	-
comp10675_c0_seq1	-	down	-6.02	3.86	4.08E-08	-	-	-
comp30091_c0_seq2	-	down	-9.22	2.10	1.33E-07	Nudt9	EGD73755	1.12E-27
comp35662_c0_seq1	-	down	-8.93	1.97	2.23E-07	Pol	XP_786277	0.00E+00
comp13170_c1_seq3	-	down	-9.02	1.96	2.30E-07	deleted in malignant brain tumors 1	XP_002833280	4.81E-30
comp513_c0_seq3	-	down	-8.98	1.93	2.64E-07	-	-	-
comp3971_c0_seq4	-	down	-4.74	5.33	5.66E-07	monocarboxylate transporter	XP_002573719	1.50E-17
comp7347_c0_seq1	-	down	-6.41	2.21	5.79E-07	-	-	-
comp23074_c0_seq1	-	down	-5.9	2.53	6.43E-07	-	-	-
comp35580_c0_seq2	-	down	-8.66	1.57	1.15E-06	-	-	-
comp3971_c0_seq1	-	down	-8.56	1.52	1.31E-06	monocarboxylate transporter	XP_001606814	2.19E-26
comp5775_c0_seq3	-	down	-5.3	2.59	1.47E-06	-	-	-
comp11368_c0_seq1	-	down	-6.16	1.88	2.26E-06	-	-	-

comp9636_c0_seq8	-	down	-5.57	2.14	3.08E-06			-
comp14613_c0_seq2	down	down	-5.06	2.12	3.59E-06			-
comp50794_c0_seq1	-	down	-8.23	1.15	5.51E-06	novel protein human megf11	EGW0405 8	1.99E-25
comp3971_c0_seq5	-	down	-4.6	3.13	5.80E-06			-
comp33670_c0_seq1	-	down	-8.23	1.10	6.91E-06			-
comp30670_c0_seq1	down	dow n	-8.15	1.05	8.24E-06			-
comp10350_c0_seq1	-	down	-8.17	1.04	8.74E-06	c-type lectin 2	XP_00260 3342	5.30E-10
comp11802_c0_seq2	-	down	-8.12	1.03	8.74E-06			-
comp869_c0_seq2	-	down	-4	6.49	9.60E-06	x-box binding	XP_00273 2738	1.19E-07
comp5787_c0_seq1	-	down	-7.99	0.98	1.05E-05			-
comp6966_c1_seq2	-	down	-8.04	0.98	1.05E-05			-
comp30235_c0_seq3	down	-	-8.14	0.97	1.17E-05			-
comp47716_c0_seq1	-	down	-8.01	0.95	1.19E-05			-
comp5396_c0_seq2	-	down	-3.91	4.71	1.25E-05	melatonin receptor 1a	ADM73175	2.20E-58
comp10976_c1_seq1	-	down	-4.41	2.80	1.26E-05			-
comp55655_c0_seq1	down	-	-8.1	0.94	1.32E-05			-
comp908_c1_seq3	-	down	-4.15	3.53	1.43E-05			-
comp25817_c0_seq1	-	down	-7.89	0.83	1.87E-05	protein tyrosine phosphatase	ACH42087	2.41E-30
comp6700_c2_seq2	-	down	-3.88	5.01	2.15E-05			-
comp27010_c0_seq1	down	-	-5.51	1.27	2.45E-05			-
comp7978_c0_seq1	down	-	-4.37	2.63	2.63E-05			-
comp28180_c0_seq1	-	down	-4.91	1.59	2.69E-05	sushi-repeat- containing x-linked 2	XP_00293 2840	3.55E-16
comp908_c1_seq2	-	down	-3.85	4.36	2.82E-05			-
comp14067_c0_seq3	-	down	-5.28	1.23	2.82E-05			-
comp908_c1_seq1	-	down	-3.75	4.90	2.98E-05			-
comp18320_c0_seq1	down	down	-3.93	4.06	3.30E-05			-

comp22724_c0_seq1	down	-	-7.8	4.92	3.36E-05	-	-	-
comp8272_c0_seq2	-	down	-7.67	3.55	3.54E-05	-	-	-

The top fifty most differentially expressed and down-regulated genes shared between GX days 1 and 5 and F3L days 1, 5, 15, and/or 30 are shown, ranked by false discovery rate-adjusted p-value. Magnitude of differential expression is expressed as log-10 fold change over control pool (logFC) and abundance is expressed as log-10 counts per million (logCPM). Regulation, or “Reg.”, (up- or down-regulated) is shown for each contig. Hyphen (-) indicates that the contig is not differentially expressed at that timepoint (Reg. columns) or that the contig does not have BLASTX hit to the NCBI non-redundant protein database with associated e-value $\leq 1e-6$ (annotation columns). Where contigs are differentially expressed in both timepoints, logFC, logCPM, and p-value correspond to the timepoint in which the contig was most highly differentially expressed and that timepoint is indicated in the regulation columns in bold.

Table 6. Pfam families enriched among the differentially expressed genes (DEGs) from resistant (GX) and susceptible (F3L) oysters at early (1 and 5 d) and late (15 and 30 d) timepoints compared to control (CGX) oysters, with rank order of significance of enrichment and significance of enrichment (p-value)

Pfam families	Rank order of significance of enrichment in F3L_early	p-value of enrichment, adjusted by FDR correction, F3L_early	Rank order of significance of enrichment in F3L_late	p-value of enrichment, adjusted by FDR correction, F3L_late	Rank order of significance of enrichment in GX_early	p-value of enrichment, adjusted by FDR correction, GX_early
TSP_1	1	4.16E-12	13	0.00598	10	0.01299
Acyl_transf_3	2	1.87E-06	9	0.00079	-	-
AIG1	3	1.90E-05	2	1.44E-09	11	0.02949
MFS_1	4	0.00011	-	-	3	0.00156
RVT_1	5	0.00011	6	0.00053	-	-
TIMP	6	0.00013	-	-	-	-
Perilipin	7	0.00018	-	-	-	-
SSF	8	0.00019	-	-	-	-
Sugar_tr	9	0.00023	-	-	-	-
An_peroxidase	10	0.00033	3	1.87E-05	-	-
Pacifastin_I	11	0.00104	-	-	-	-
GlcNAc_2-epim	12	0.00194	-	-	-	-
Transposase_21	13	0.00194	4	5.35E-05	-	-
VWD	14	0.00194	-	-	-	-
SNF	15	0.00211	-	-	-	-
Ank_2	16	0.00222	1	8.50E-11	-	-
7tm_1	17	0.00243	-	-	-	-
Nucleoplasmin	18	0.00271	-	-	-	-
T2SE	19	0.00271	7	0.00076	-	-
Dam	20	0.00480	-	-	-	-
Man-6-P_recep	21	0.00612	10	0.00079	-	-
Pro_dh	22	0.00612	-	-	-	-
TauD	23	0.00790	-	-	-	-
Dynamin_N	24	0.00847	11	0.00092	-	-
AA_permease_2	25	0.01120	-	-	-	-
Methyltransf_FA	26	0.01123	-	-	-	-
NAD_binding_5	27	0.01123	-	-	1	0.00069
Pyridoxal_deC	28	0.01440	16	0.01161	-	-
Sulfate_transp	29	0.01582	-	-	-	-
Cpn60_TCP1	30	0.01692	-	-	2	0.00156
Peptidase_M13_N	31	0.02113	-	-	-	-
HTH_Tnp_Tc3_2	32	0.02465	-	-	-	-
ApoL	33	0.02787	5	0.00053	-	-
Gal_Lectin	34	0.02841	-	-	-	-
zf-TAZ	35	0.02841	15	0.00819	-	-
A2M	36	0.03380	-	-	-	-
GCC2_GCC3	37	0.04229	-	-	-	-
HSP70	38	0.04229	-	-	-	-
KR	39	0.04229	-	-	-	-
VOMI	-	-	8	0.00076	-	-
CHB_HEX_C_1	-	-	12	0.00299	-	-
RPE65	-	-	14	0.00819	-	-
DDE_1	-	-	17	0.02056	-	-
Mucin2_WxxW	-	-	18	0.02422	-	-
MMR_HSR1	-	-	19	0.03240	-	-
Peptidase_M84	-	-	20	0.04388	-	-
zf-MYND	-	-	21	0.04388	-	-

Solute_trans_a	-	-	-	-	4	0.00156
IL17	-	-	-	-	5	0.00206
HSP20	-	-	-	-	6	0.00355
bZIP_2	-	-	-	-	7	0.01033
Patched	-	-	-	-	8	0.01299
Sulfatase	-	-	-	-	9	0.01299
HSP70	-	-	-	-	12	0.02965
GCC2_GCC3	-	-	-	-	13	0.04324

Pearson's chi-squared test was performed to test for the enrichment of each unique HMM accession. If the expected count for any cell in the 2x2 contingency table was 5 or fewer, Fisher's exact test was performed instead. Significance p-values listed above were adjusted for the number of independent enrichment tests performed for each group of DEGs by the False Discovery Rate correction method. Pfam families are ordered above by (1) rank order of significance among F3L_early DEGs, (2) rank order of significance among F3L_late DEGs, then (3) rank order of significance among GX_early DEGs.

Table 7. Pfam domains enriched among the differentially expressed genes (DEGs) from resistant (GX) and susceptible (F3L) oysters at early (1 and 5 d) and late (15 and 30 d) timepoints compared to control (CGX) oysters, with rank order of significance of enrichment and significance of enrichment (p-value)

Pfam domains	Rank order of significance of enrichment in F3L_early (days 1, 5)	p-value of enrichment, adjusted by FDR correction, F3L_early	Rank order of significance of enrichment in F3L_late (days 15, 30)	p-value of enrichment, adjusted by FDR correction, F3L_late	Rank order of significance of enrichment in GX_early (days 1, 5)	p-value of enrichment, adjusted by FDR correction, GX_early
C1q	1	5.51E-53	2	1.57E-70	4	0.00761
Fibrinogen_C	2	2.86E-24	8	1.22E-15	-	-
Lectin_C	3	1.67E-21	13	3.81E-07	-	-
VWA	4	6.91E-14	-	-	-	-
ShK	5	8.62E-09	17	6.59E-06	-	-
Kazal_1	6	8.29E-07	-	-	-	-
SRCR	7	2.35E-06	18	0.00555	15	0.01437
Kunitz_BPTI	8	8.56E-06	12	0.00770	12	0.03973
Chitin_bind_3	9	6.40E-05	-	-	6	0.01029
Cu2_monooxygen	10	0.00014	4	6.59E-06	-	-
E1_DerP2_DerF2	11	0.00101	-	-	-	-
BBE	12	0.00251	1	0.00019	-	-
Cu-oxidase	13	0.00251	-	-	-	-
Defensin_2	14	0.00251	6	0.00011	-	-
Sulfotransfer_1	15	0.00313	-	-	-	-
CBM_14	16	0.00362	-	-	-	-
p450	17	0.00362	-	-	-	-
GOLD_2	18	0.00438	-	-	-	-
Kazal_2	19	0.00438	11	0.03974	-	-
Sulfate_tra_GLY	20	0.00438	-	-	-	-
Cystatin	21	0.00688	-	-	-	-
HYR	22	0.00751	10	0.01121	10	0.00419
Thioredoxin_4	23	0.00846	-	-	-	-
TILa	24	0.00846	-	-	-	-
Cu2_monoox_C	25	0.01419	3	0.00301	-	-
Glyco_hydro_9	26	0.01419	-	-	-	-
PAX	27	0.01419	-	-	-	-
DOMON	28	0.01976	7	0.00770	-	-
CBM_4_9	29	0.02046	-	-	-	-
Lipase	30	0.02509	-	-	-	-
adh_short_C2	31	0.02823	-	-	-	-
F5_F8_type_C	32	0.02823	-	-	-	-
GBP	33	0.03023	9	0.01014	-	-
Cu-oxidase_2	34	0.03302	-	-	-	-
PBP	35	0.03302	-	-	-	-
T-box	36	0.03553	-	-	-	-
Tyrosinase	37	0.03784	-	-	-	-
CRAL_TRIO	38	0.03850	-	-	-	-
Trypsin	39	0.03850	-	-	16	0.00060
Transglut_N	40	0.04209	-	-	-	-
FTCD	41	0.04469	-	-	-	-
Ras	42	0.04469	-	-	-	-
DUF4218	43	0.04731	-	-	-	-
EGF_3	44	0.04731	-	-	-	-
Cupin_8	-	-	5	0.02629	-	-

Myb_DNA-bind_4	-	-	14	0.03888	-	-
rve	-	-	15	0.00084	-	-
RVT_3	-	-	16	0.01014	-	-
Acyl_transf_1	-	-	-	-	1	0.00209
Arginase	-	-	-	-	2	0.00406
BIR	-	-	-	-	3	0.01156
cEGF	-	-	-	-	5	0.02831
CUB	-	-	-	-	7	0.01029
EF_hand_5	-	-	-	-	8	0.03888
H_lectin	-	-	-	-	9	0.00419
ketoacyl-synt	-	-	-	-	11	0.00209
MAM	-	-	-	-	13	0.03520
Peptidase_S8	-	-	-	-	14	0.01706

Pearson's chi-squared test was performed to test for the enrichment of each unique HMM accession. If the expected count for any cell in the 2x2 contingency table was 5 or fewer, Fisher's exact test was performed instead. Significance p-values listed above were adjusted for the number of independent enrichment tests performed for each group of DEGs by the Bonferroni correction method. Pfam domains are ordered above by (1) rank order of significance among F3L_early DEGs, (2) rank order of significance among F3L_late DEGs, then (3) rank order of significance among GX_early DEGs.

Table 8. TribeMCL clusters enriched for differentially expressed genes (DEGs) from resistant (GX) and susceptible (F3L) oysters at early (1 and 5 d) and late (15 and 30 d) timepoints, with rank order of significance of enrichment and significance of enrichment (p-value)

Tribe MCL cluster	# contigs in cluster	Annotation	Rank order p-value F3L_early	Enrich. p-value F3L_early	Rank order p-value F3L_late	Enrich. p-value F3L_late	Rank order p-value GX_early	Enrich. p-value GX_early
15	214	c-type lectin	1	3.44E-28	17	2.49E-06	17	0.00713
33	150	c1q domain containing protein	2	3.70E-22	9	2.32E-08	-	-
19	196	unknown	3	3.14E-13	1	2.01E-14	-	-
35	144	fibrinogen domain-containing*	4	3.87E-11	2	7.92E-11	-	-
27	155	hemicentin/rhamnospondin/t hrombospondin*	5	6.43E-10	26	2.31E-05	5	0.00052
535	16	peritrophin*	6	7.87E-10	41	0.00064	-	-
1084	9	vdg3	7	7.93E-10	49	0.00163	-	-
493	17	dopamine beta hydroxylase	8	1.39E-09	4	1.32E-10	-	-
31	151	scavenger receptor cysteine-rich protein type 12/deleted in malignant brain tumors 1	9	2.81E-09	34	0.00038	-	-
179	40	c1q domain containing protein	10	1.36E-08	3	1.32E-10	-	-
532	16	nose resistant to fluoxetine family member (nrf-6)	11	1.78E-08	40	0.00064	-	-
165	43	unknown	12	3.13E-08	5	4.88E-09	-	-
32	151	unknown	13	3.63E-08	6	4.88E-09	-	-
1987	6	unknown	14	4.07E-08	-	-	-	-
325	24	c1q domain containing protein	15	6.03E-08	24	2.17E-05	-	-
55	106	IgGFc-binding protein-like / ig-like domain-containing / fc fragment of binding protein*	16	6.22E-08	52	0.00193	-	-
45	124	collagen alpha-1/3/4/5	17	1.51E-07	-	-	-	-
1416	7	unknown	18	2.15E-07	25	2.31E-05	10	0.00059

50	112	unknown	19	4.48E-06	8	2.32E-08	-	-
1716	6	unknown	20	4.76E-06	22	1.16E-05	6	0.00052
928	10	hypothetical protein BRAFLDRAFT_87756*	21	4.76E-06	-	-	-	-
67	89	short-chain collagen c4	22	4.76E-06	27	2.31E-05	-	-
22	181	von willebrand factor d and egf domain-containing	23	4.76E-06	15	1.49E-06	-	-
513	16	fibropellin	24	6.81E-06	39	0.00064	-	-
135	48	myc homolog	25	6.81E-06	10	1.51E-07	-	-
1457	7	unknown	26	1.30E-05	-	-	-	-
3508	4	unknown	27	1.82E-05	75	0.00455	-	-
3509	4	unknown	28	1.82E-05	14	1.25E-06	-	-
24	167	cytochrome p450	29	3.73E-05	-	-	-	-
142	47	unknown	30	3.94E-05	21	1.16E-05	-	-
36	139	polyprotein	31	5.97E-05	29	3.75E-05	-	-
1036	9	serine protease inhibitor cvsi-2	32	5.99E-05	-	-	1	8.10E-10
2141	5	unknown	33	7.50E-05	-	-	-	-
335	23	unknown	34	8.55E-05	13	1.25E-06	-	-
888	10	unknown	35	0.00011	18	3.87E-06	-	-
52	108	multicopper oxidase	36	0.00018	-	-	-	-
1604	6	unknown	37	0.00019	93	0.00904	-	-
1747	6	unknown	38	0.00019	-	-	-	-
1817	6	heat shock protein 60	39	0.00019	-	-	8	0.00052
463	18	loc571499 protein	40	0.00019	46	0.00096	-	-
109	58	tyrosinase/cre-tyr protein	41	0.00021	-	-	-	-
4174	3	unknown	42	0.00035	28	3.68E-05	-	-
4944	3	unknown	43	0.00035	-	-	-	-
5117	3	unknown	44	0.00035	-	-	-	-
103	62	serine protease*	45	0.00035	-	-	3	0.00037
658	13	unknown	46	0.00038	-	-	-	-
254	30	chitin binding domain	47	0.00042	-	-	16	0.00713

		protein*						
3	635	nacht and ankyrin domain containing protein / ankyrin repeat protein / ankyrin unc44*	48	0.00046	7	5.66E-09	-	-
97	65	c1q domain containing protein	49	0.00048	16	1.99E-06	-	-
39	137	transient receptor potential cation subfamily member*	50	0.00049	-	-	-	-
1200	8	unknown	51	0.00061	12	1.25E-06	-	-
1258	8	AF369699_1SHG	52	0.00061	-	-	-	-
341	23	unknown	53	0.00066	-	-	-	-
542	16	hemagglutinin/amebocyte aggregation factor/dermatopontin 2	54	0.00095	42	0.00064	15	0.00219
970	9	collagen*	55	0.00095	-	-	-	-
1031	9	protein-glutamine gamma-glutamyltransferase	56	0.00095	-	-	-	-
1079	9	glycoside hydrolase	57	0.00095	-	-	-	-
2593	4	unknown	58	0.00095	-	-	-	-
3367	4	unknown	59	0.00095	-	-	-	-
3386	4	unknown	60	0.00095	-	-	-	-
3460	4	unknown	61	0.00095	-	-	-	-
3921	4	unknown	62	0.00095	33	0.00012	-	-
30	152	gtpase imap family member	63	0.00095	19	4.38E-06	-	-
505	17	unknown	64	0.00117	30	5.17E-05	2	2.64E-06
929	10	epididymal secretory protein e1 precursor/niemann-pick type c2	65	0.00139	-	-	-	-
287	27	pancreatic lipase-related protein	66	0.00139	-	-	-	-
47	122	organic cation transporter	67	0.00145	-	-	-	-
840	11	beta-lactamase family protein*	68	0.00202	-	-	-	-
25	160	neurotransmitter transporter	69	0.00202	-	-	-	-
2108	5	unknown	70	0.00204	-	-	-	-
2140	5	c1q domain containing	71	0.00204	82	0.00665	4	0.00052

		protein						
71	85	von willebrand factor d and egf domain-containing*	72	0.00313	-	-	-	-
1615	6	myc homolog	73	0.00340	-	-	-	-
1642	6	unknown	74	0.00340	-	-	-	-
1656	6	monocarboxylate transporter	75	0.00340	-	-	-	-
1808	6	unknown	76	0.00340	-	-	-	-
1810	6	pe-pgrs family protein	77	0.00340	-	-	-	-
1873	6	unknown	78	0.00340	95	0.00904	-	-
1973	6	unknown	79	0.00340	-	-	-	-
1989	6	actin binding protein	80	0.00340	37	0.00054	-	-
654	13	cell adhesion molecule	81	0.00340	-	-	-	-
663	13	sialic acid binding lectin/c1q domain containing protein*	82	0.00340	-	-	-	-
669	13	paired box protein	83	0.00340	-	-	-	-
700	13	retinal-binding protein/sec14l1 protein	84	0.00340	-	-	-	-
68	88	peroxidase	85	0.00350	43	0.00065	-	-
29	153	serine threonine kinase 3/7	86	0.00429	35	0.00041	-	-
619	14	hypothetical protein BRAFLDRAFT_129074*	87	0.00442	36	0.00042	-	-
626	14	tissue inhibitor of metalloproteinase-3	88	0.00442	-	-	-	-
137	48	sodium myo-inositol cotransporter / sodium glucose cotransporter*	89	0.00472	-	-	-	-
1365	7	unknown	90	0.00529	-	-	-	-
1469	7	mgc154819	91	0.00529	-	-	-	-
552	15	sushi domain-containing protein 2/von willebrand factor type d domain protein	92	0.00555	-	-	-	-
570	15	hypothetical protein BRAFLDRAFT_117849	93	0.00555	-	-	-	-
294	26	jagged 1-like*	94	0.00661	-	-	-	-
514	16	serine protease inhibitor-11	95	0.00698	-	-	-	-

541	16	t-box transcription factor	96	0.00698	-	-	-	-
58	98	unknown	97	0.00704	69	0.00447	-	-
1161	8	isopentenyl pyrophosphate:dimethylallyl pyrophosphate	98	0.00743	-	-	-	-
1191	8	endo-1'4'-beta-d-glucanase	99	0.00743	-	-	-	-
1226	8	BRAFLDRAFT_84494	100	0.00743	-	-	-	-
1249	8	unknown	101	0.00743	47	0.00115	-	-
489	17	unknown	102	0.00829	-	-	-	-
490	17	ring finger protein 213-like	103	0.00829	-	-	-	-
278	28	unknown	104	0.00845	-	-	-	-
174	41	tripartite motif-containing	105	0.00891	-	-	-	-
1042	9	endo-beta-1'4'-glucanase	106	0.00973	-	-	-	-
168	43	amino acid transporter	107	0.00973	-	-	-	-
416	19	receptor tyrosine kinase / insulin receptor*	108	0.00973	-	-	-	-
430	19	novel protein vertebrate egf-like repeats and discoidin i-like domains 3*	109	0.00973	-	-	-	-
4167	3	unknown	110	0.00973	54	0.00285	-	-
4249	3	pacifastin	111	0.00973	-	-	-	-
4254	3	collagen alpha-1 chain	112	0.00973	-	-	-	-
4314	3	unknown	113	0.00973	55	0.00285	-	-
4315	3	unknown	114	0.00973	-	-	-	-
4316	3	mitogen-activated protein kinase kinase kinase 7	115	0.00973	-	-	-	-
4360	3	unknown	116	0.00973	-	-	-	-
4509	3	unknown	117	0.00973	56	0.00285	-	-
4539	3	unknown	118	0.00973	-	-	-	-
4545	3	unknown	119	0.00973	-	-	-	-
4613	3	legumain	120	0.00973	-	-	-	-
4633	3	unknown	121	0.00973	57	0.00285	-	-
4715	3	unknown	122	0.00973	-	-	-	-

4736	3	peptidoglycan-binding domain 1 protein	123	0.00973	-	-	-	-
4786	3	unknown	124	0.00973	58	0.00285	-	-
5277	3	unknown	125	0.00973	-	-	-	-
5296	3	predicted protein	126	0.00973	60	0.00285	-	-
5743	3	unknown	127	0.00973	-	-	-	-
5832	3	unknown	128	0.00973	-	-	-	-
5912	3	unknown	129	0.00973	-	-	-	-
6037	3	unknown	130	0.00973	-	-	-	-
6056	3	unknown	131	0.00973	-	-	-	-
6151	3	unknown	132	0.00973	61	0.00285	-	-
6333	3	unknown	133	0.00973	-	-	-	-
6428	3	unknown	134	0.00973	62	0.00285	-	-
6438	3	unknown	135	0.00973	-	-	-	-
6600	3	unknown	136	0.00973	63	0.00285	-	-
6626	3	unknown	137	0.00973	-	-	-	-
163	44	rapunzel 4/5*	138	0.00978	11	9.26E-07	-	-
404	20	apextrin-like	-	-	20	7.96E-06	-	-
76	84	neoverrucotoxin	-	-	23	1.59E-05	-	-
1010	9	unknown	-	-	31	6.64E-05	-	-
234	32	peptidoglycan-binding lysin domain*	-	-	32	9.77E-05	-	-
51	111	cell adhesion*	-	-	38	0.00064	-	-
92	68	unknown	-	-	44	0.00076	-	-
61	92	unknown	-	-	45	0.00085	-	-
450	19	unknown	-	-	48	0.00115	-	-
370	22	unknown	-	-	50	0.00193	-	-
372	22	unknown	-	-	51	0.00193	-	-
838	11	interferon-induced very large gtpase 1	-	-	53	0.00285	-	-
4912	3	unknown	-	-	59	0.00285	-	-
299	26	unknown	-	-	64	0.00303	-	-

703	12	unknown	-	-	65	0.00306	-	-
753	12	unknown	-	-	66	0.00306	-	-
157	45	predicted protein(gi 156221710 gb ED O42562.1)*	-	-	67	0.00308	-	-
263	29	unknown	-	-	68	0.00435	-	-
2783	4	unknown	-	-	70	0.00455	-	-
2857	4	unknown	-	-	71	0.00455	-	-
2901	4	legumain	-	-	72	0.00455	-	-
2902	4	unknown	-	-	73	0.00455	-	-
3158	4	unknown	-	-	74	0.00455	-	-
3606	4	unknown	-	-	76	0.00455	-	-
3827	4	creb-binding protein	-	-	77	0.00455	-	-
3873	4	unknown	-	-	78	0.00455	-	-
124	52	interferon-induced protein 44	-	-	79	0.00503	-	-
567	15	aromatic amino acid decarboxylase	-	-	80	0.00503	-	-
230	33	hypothetical protein BRAFLDRAFT_82912	-	-	81	0.00597	-	-
2238	5	unknown	-	-	83	0.00665	-	-
2460	5	unknown	-	-	84	0.00665	-	-
2463	5	tissue inhibitor of metalloproteinase 3	-	-	85	0.00665	-	-
2487	5	unknown	-	-	86	0.00665	-	-
2520	5	unknown	-	-	87	0.00665	-	-
2521	5	unknown	-	-	88	0.00665	-	-
74	84	SINV_05289	-	-	89	0.00746	-	-
454	18	unknown	-	-	90	0.00771	-	-
104	62	cell cycle checkpoint protein rad17	-	-	91	0.00904	-	-
1575	6	unknown	-	-	92	0.00904	-	-
1646	6	receptor for egg jelly	-	-	94	0.00904	-	-
1962	6	antileukoproteinase	-	-	96	0.00904	-	-

411	20	gm2 ganglioside activator	-	-	97	0.00975	-	-
1781	6	organic solute transporter subunit alpha	-	-	-	-	7	0.00052
1369	7	unknown	-	-	-	-	9	0.00059
1119	8	melatonin receptor 1a-like	-	-	-	-	11	0.00071
1026	9	unknown	-	-	-	-	12	0.00083
919	10	unknown	-	-	-	-	13	0.00096
99	64	fibropellin	-	-	-	-	14	0.00210
48	117	monocarboxylate transporter	-	-	-	-	18	0.00912
192	37	proprotein convertase subtilisin kexin*	-	-	-	-	19	0.00922

Pearson's chi-squared test was performed to test for the enrichment of each group of DEGs for each TribeMCL cluster. If the expected count for any cell in the 2x2 contingency table was 5 or fewer, Fisher's exact test was performed instead. Significance p-values listed above were adjusted for the number of independent enrichment tests performed for each group of DEGs by the False Discovery Rate correction method. Enriched TribeMCL clusters are ordered above by (1) rank order of significance of enrichment for F3L_early DEGs, (2) rank order of significance of enrichment for F3L_late DEGs, then (3) rank order of significance of enrichment for GX_early DEGs. Enriched TribeMCL clusters were annotated with a protein name if more than half of the contigs had identical or nearly identical BLASTX best hits and the remainder of contigs had no significant hits, or if $\geq 80\%$ of the contigs had identical or nearly identical BLASTX best hits and the remainder of contigs had dissimilar best BLASTX hits (which may be the case in the sharing of domains, e.g., neurotrypsin and trypsin). In the cases in which contigs that composed a TribeMCL cluster had no best BLASTX hits or had a number of dissimilar hits that if counted conjointly did not represent a majority of the contigs, then that TribeMCL cluster was not annotated, that is, was named "unknown." * Enriched TribeMCL clusters were reservedly annotated if less than half of the contigs had identical or nearly identical BLASTX best hits and composed a plurality while the remainder of contigs had no significant hits or a small number of dissimilar but non-repeating hits (frequency equal to one), or if a number of dissimilar hits of similar frequency that if counted conjointly composed $\geq 50\%$ of the contigs in the cluster (e.g., "hemicentin/rhamnospondin/thrombospondin*").

Table 9. Comparison of select “diversified” groups of genes in the Eastern oyster to the Pacific oyster *C.gigas* genome and differential expression

Select Gene Groups of Interest	SPs	SPIs	FREPs	C1qDC proteins	CTLDC proteins	DMBT1/SRCR type 12	IFI44	GIMAP proteins
# of contigs	112	99	180	492	404	187	88	210
# of nonredundant contigs (reduced by 95% similarity)	73	81	128	391	276	122	59	173
# of nonredundant contigs that map to <i>C. gigas</i> genome	70	73	115	323	220	109	45	158
# of unique <i>C. gigas</i> loci to which contigs map	51	61	74	149	140	61	22	33
# of genes in <i>C.gigas</i> genome	22 ^a	40 ^a	190 ^b	321 ^b	266 ^b	43 ^a	27 ^a	19 ^a
#DEGs GX_early	7	8	0	5	6	3	2	3
#DEGs F3L_early	13	22	34	78	56	21	5	18
#DEGs F3L_late	2	9	27	54	25	11	6	18
#DEGs total	14/112	24/99	41/180	90/492	65/404	22/187	9/88	21/210

a. Number of genes in the *C. gigas* genome determined by number of *C. gigas* GLEAN gene model peptide sequences with reciprocal blast hits (e-value $\leq 1e-05$) to transcript nucleotide sequence in each group of interest (TBLASTN, BLASTX) and with best blast hits with the select gene group of interest.

b. Numbers of genes in the *C. gigas* genome provided by Zhang et al. 2012.

FIGURES

Figure 1. Cumulative percent mortality in two families of oysters challenged with the bacterial pathogen *Roseovarius crassostreae* (F3L and GX) and in unchallenged controls (CF3L and CGX).

The cumulative mortality is shown over the course of the 93-day bacterial challenge for both families (F3L and GX) and both challenge and control oysters, the latter of which is indicated by the prefix “C” for “control.” Time 0 is the moment of exposure by bath to *Roseovarius crassostreae*. A Pearson’s chi-squared test of significance performed for each pairwise comparison between groups F3L, CGX, and GX at day 28, two days before the final RNA sample collection, and at day 93, the final timepoint of the bacterial challenge. Significance values (p-values) were adjusted independently at each timepoint by the Bonferroni method to account for the multiple comparisons. The four arrows indicate days 1, 5, 15, and 30 at which timepoints RNA was isolated from CGX, F3L, and GX, for cDNA synthesis and sequencing.

Figure 2. Heatmap of all differentially expressed genes (DEGs) in two oyster families experimentally challenged with the bacterial pathogen *Roseovarius crassostreae* (GX and F3L) and unchallenged controls

For all genes differentially expressed in any one of the treatments GX-d1, GX-d5, F3L-d1, F3L-d5, F3L-d15, and F3L-d30, the Z-score centered $\log_2$ -transformed RPKM for each gene in each of eight treatment-days (including control, CGX-d15 and CGX-d30) is shown. Genes were hierarchically clustered using Euclidean distance and complete linkage of the Z-score-transformed gene expression. Sample groups were clustered using the complete linkage Euclidean distance of the Spearman correlation of the Z-score-transformed gene expression. Clustering and visualization were performed using the fastcluster and gplots packages, respectively, in the R programming environment.

Figure 3. Numbers of differentially expressed contigs (DEGs) shared and unique between GX_early (resistant oysters – days 1 and 5), F3L_early (susceptible oysters – days 1 and 5), and F3L_late (susceptible oysters – days 15 and 30)

Figure 4. Functionally enriched Gene Ontology terms among DEGs from resistant (GX) and susceptible (F3L) oysters at early (1 and 5 d) and late (15 and 30 d) time points compared to control (CGX) oysters mapped by semantic similarity using SimRel method and REVIGO

Functionally enriched Gene Ontology (GO) terms for each category of differentially expressed genes, GX_early, F3L_early, F3L_late, and for the three highest-level categories of the Gene Ontology hierarchy are displayed in SimRel semantic mapping space (Schlicker et al. 2006) by modifying the output of the REVIGO server (Supek et al. 2011) in the R programming environment and plotting using ggplot2. Semantic mapping space is

equivalent within the same broad GO term category, e.g. a x,y-coordinate in GX_early biological process space would have the same identity as the same x,y-coordinate in F3L_early biological process space. The color of nodes from cool (green) to warm (red) signifies increasing significance of enrichment as indicated in the legend. The size of nodes reflects whether the GO term is enriched among upregulated DEGs (large) or downregulated DEGs (small), while a GO term enriched among both upregulated and downregulated DEGs is represented by a medium size node and can be further identified by its unique, square shape. To enhance readability, overlapping nodes were sometimes labeled conjointly. This was done in a manual manner by selecting a term name which properly described the conjointly labeled terms. These cases were noted by the addition of the suffix “-related.”

Figure 5. Numbers of non-redundant transcripts/proteins and gene models in selected putative diversified gene families in multiple organisms from diverse taxa

In each cell is shown side-by-side: 1) the numbers of non-redundant transcripts/proteins, as determined by similarity clustering and 2) numbers of gene models (see Methods). A species tree is reproduced beside the matrix to emphasize evolutionarily relationships between the featured organisms. The number of gene models (the second number in each cell) was Z-score centered by gene family and the magnitude of this Z-score was assigned a color according to the color gradient indicated in the key.

Figure 6. *C. virginica* GIMAP translated transcripts clustered by similarity with color and size reflective of differential expression

C. virginica GIMAP translated transcripts represented in Cytoscape 2.8 and yFiles organic layout. An edge corresponds to a significant similarity (e-value $\leq 1e-05$, hit identity $\geq 20\%$), and the wider the edge, the higher the hit identity percentage.

Figure 7. *C. virginica* IFI44 translated transcripts clustered by similarity with color and size reflective of differential expression

C. virginica IFI44 translated transcripts represented in Cytoscape 2.8 and yFiles organic layout. An edge corresponds to a significant similarity (e-value $\leq 1e-05$, hit identity $\geq 20\%$), and the wider the edge, the higher the hit identity percentage.

Figure 8. Maximum likelihood phylogenetic tree of non-redundant GTPase of the immunity associated protein (GIMAP) transcripts/proteins from multiple organisms with leaves colored by organism

A manually curated multiple alignment of GTPase of immunity associated protein (GIMAP) translated transcripts/proteins from multiple organisms in diverse taxa was used generate the above maximum likelihood phylogenetic tree using FastTree2 assuming a WAG model and hybrid CAT/gamma approximations. The tree is represented as a circular cladogram and bootstrap support from 1,000 replicates is indicated as a percentage next to each tree node. Leaves of the tree are colored according to the species to which the sequence belongs, as specified in the legend.

Figure 9. Maximum likelihood phylogenetic tree of non-redundant GTPase of the immunity associated protein (GIMAP) transcripts/proteins from multiple organisms with major groupings highlighted and summarized

A manually curated multiple alignment of GTPase of immunity associated protein (GIMAP) translated transcripts/proteins from multiple organisms in diverse taxa was used generate the above maximum likelihood phylogenetic tree using FastTree2 assuming a WAG model and hybrid CAT/gamma approximations (as in Fig. 8). The tree is represented as a circular cladogram. Bootstrap support from 1,000 replicates is indicated for the only node used to define the three major groupings indicated by the solid or dotted arcs that circumscribe the tree. (The group indicated by the dotted arc is dotted because it was defined negatively by exclusion.) The numbers of sequences for each species in each of the major groupings are listed in parentheses beside the species abbreviation. Some details are provided on the *C. gigas* genomic loci to which the *C. virginica* sequences mapped.

Figure 10. Maximum likelihood phylogenetic tree of non-redundant interferon-induced protein 44 (IFI44) transcripts/proteins with leaves colored by organism

A manually curated multiple alignment of interferon-induced protein 44 (IFI44) translated transcripts/proteins from multiple organisms in diverse taxa was used to generate the above maximum likelihood phylogenetic tree using FastTree2 assuming a WAG model and hybrid CAT/gamma approximations. The tree is represented as a circular cladogram and bootstrap support from 1,000 replicates is indicated as a percentage next to each tree node. Leaves of the tree are colored according to the species to which the sequence belongs, as specified in the legend.

Figure 11. Maximum likelihood phylogenetic tree of non-redundant interferon-induced protein 44 (IFI44) transcripts/proteins with major groupings highlighted and summarized

A manually curated multiple alignment of interferon-induced protein 44 (IFI44) translated transcripts/proteins from multiple organisms in diverse taxa was used to generate the above maximum likelihood phylogenetic tree using FastTree2 assuming a WAG model and hybrid CAT/gamma approximations (as in Fig. 10). The tree is represented as a circular cladogram. Bootstrap support from 1,000 replicates is indicated for only those nodes used to define the major groupings indicated by the solid arcs that circumscribe the tree. The major groupings of the tree that contained bivalve-only sequences were indicated by red coloration in the inter-branch space and the major grouping that contained sequences from diverse organisms was indicated by light blue coloration in the inter-branch space. The numbers of sequences for each species in each of the major groupings are listed in parentheses beside the species abbreviation. Some details are provided on the *C. gigas* genomic loci to which the *C. virginica* sequences mapped.

APPENDIX

A. Tables

Table S1. Sources and type of sequences downloaded for multiple organisms for multi-species similarity clustering

Organism	Type of seq.	Source	Notes	Date down-loaded	Name of seq. file
<i>L. stagnalis</i>	ESTs	http://www.ncbi.nlm.nih.gov/nucest	-	2/28/12	-
<i>B. glabrata</i>	ESTs	http://www.ncbi.nlm.nih.gov/nucest	-	2/28/12	-
<i>A. californica</i>	ESTs	http://www.ncbi.nlm.nih.gov/nucest	EST assembly*	2/28/12	-
<i>C. gigas</i>	ESTs	http://public-contigbrowser.sigenae.org:9090/Crassostrea_gigas/index.html	version 8	2/28/12	cgigas_all_contigs.tfa
<i>B. floridae</i>	Protein	http://www.uniprot.org/uniref/ , Suzek et al. 2007	Uniprot reference proteome	2/28/12	-
<i>D. melanogaster</i>	Protein	http://www.uniprot.org/uniref/ , Suzek et al. 2007	Uniprot reference proteome	2/28/12	-
<i>D. pulex</i>	Protein	http://www.uniprot.org/uniref/ , Suzek et al. 2007	Uniprot reference proteome	2/28/12	-
<i>N. vectensis</i>	Protein	http://www.uniprot.org/uniref/ , Suzek et al. 2007	Uniprot reference proteome	2/28/12	-
<i>C. intestinalis</i>	Protein	http://www.uniprot.org/uniref/ , Suzek et al. 2007	Uniprot reference proteome	2/28/12	-
<i>S. purpuratus</i>	Protein	http://sugp.caltech.edu/SPBase/download/	translations of gene models	2/28/12	SPU_peptide.fasta
<i>P. marinus</i>	Protein	ftp://ftp.ensembl.org/pub/release-67/fasta/petromyzon_marinus/pep/	translations of gene models	2/28/12	Petromyzon_marinus.Pmarinus_7.0.66.pep.all.fasta
<i>S. kowalevskii</i>	Protein	http://www.ncbi.nlm.nih.gov/RefSeq/	translations of gene models	2/28/12	-
<i>L. gigantea</i>	Protein	http://genome.jgi-psf.org/Lotgi1/Lotgi1.download.ftp.html	translations of gene models	2/28/12	Lotgi1_GeneModels_FilteredModels1_aa.fasta
<i>P. fucata</i>	Protein	http://marinegenomics.oist.jp/genomes/download?project_id=20	translations of gene	2/28/12	pfu_aug1.0_Pall.fasta

		Takeuchi et al. 2012	models		
--	--	----------------------	--------	--	--

* ESTs (255,605) were assembled using CLC Genomics Workbench (version 4.8) on default settings into 70,053 sequences (including contigs and singletons).

Table S2. Estimation of total number of gene models in the genomes of multiple organisms

Organism	Number of gene models	Source
<i>B. glabrata</i>	21900	transferred from http://genome.jgi-psf.org/Lotgi1/Lotgi1.info.html
<i>A. californica</i>	25000	(roughly estimated here)
<i>C. gigas</i>	28027	Zhang et al. 2012
<i>C. virginica</i>	28027	transferred from Zhang et al. 2012
<i>B. floridae</i>	21900	Dishaw et al. 2012
<i>D. melanogaster</i>	14442	Hahn et al. 2007
<i>D. pulex</i>	30907	Colbourne et al. 2012
<i>N. vectensis</i>	18000	Putnam et al. 2007
<i>C. intestinalis</i>	16000	Dehal et al. 2002
<i>S. purpuratus</i>	28944	http://www.spbase.org/SpBase/resources/index.php
<i>P. marinus</i>	10402	http://useast.ensembl.org/Petromyzon_marinus/Info/Annotation/#genebuild
<i>S. kowalevskii</i>	20000	http://www.hgsc.bcm.tmc.edu/project-species-o-Acorn%20worm.hgsc
<i>L. gigantea</i>	23851	http://genome.jgi-psf.org/Lotgi1/Lotgi1.info.html
<i>P. fucata</i>	23257	Takeuchi et al. 2012

Table S3. Unique annotations in the set of GX_early DEGs not in F3L DEGs

Contig	Reg. GX 1d	Reg. GX 5d	logFC	log CPM	p-value	Best blastx hit to nr db	Hit accession	Hit e- value
comp12125_c0_seq3	NA	up	2.76	0.95	0.000178	notch 2	XP_858190	3.06E-07
comp1285_c1_seq3	NA	down	-3.54	3.50	9.56E-05	arginase ii	XP_002130834	6.53E-12
comp1285_c1_seq8	NA	down	-3.90	3.84	2.38E-05	arginase type i-like protein	AEB70965	5.87E-28
comp18757_c0_seq1	NA	down	-7.14	0.09	0.000236	hla-b associated transcript 1	XP_003217350	6.91E-21
comp18902_c0_seq1	NA	down	-3.43	1.62	0.000649	rho gtpase	XP_002739105	1.37E-58
comp24124_c0_seq1	NA	up	4.49	2.59	9.31E-09	ched related family member (ptr-19)	XP_002734100	1.39E-138
comp24124_c0_seq4	NA	up	2.71	1.85	0.000232	ched related family member (ptr-19)	XP_002734100	1.20E-07
comp2906_c1_seq15	up	NA	2.62	1.79	0.000334	acetyl-carboxylase	AAF22966	9.83E-44
comp4626_c0_seq4	up	NA	3.34	2.55	3.32E-06	probable alpha-ketoglutarate-dependent hypophosphite dioxxygenase-like	XP_002944900	1.40E-10
comp6161_c0_seq10	up	NA	3.17	2.50	2.16E-05	proprotein convertase subtilisin kexin type	CBY34171	4.38E-13
comp6161_c0_seq11	up	NA	3.10	3.16	1.59E-05	furin (paired basic amino acid cleaving enzyme)	AAA49718	1.42E-22
comp6161_c0_seq5	up	NA	3.49	3.24	1.57E-06	type 2 proinsulin processing endopeptidase	2206277A	2.33E-42
comp664_c0_seq4	up	NA	2.42	4.67	0.000621	unc-5 homolog b	XP_001642030	1.34E-14
comp6837_c0_seq1	up	NA	3.15	4.45	1.16E-05	interleukin 17d	A9XE49	1.05E-56
comp6837_c0_seq2	up	NA	3.02	4.92	2.41E-05	interleukin 17-like	A9XE49	7.77E-66
comp688_c0_seq1	up	NA	2.52	7.53	0.000343	erythrocyte membrane-associated giant protein antigen 332	XP_002167006	6.21E-09
comp7972_c0_seq1	NA	down	-8.09	1.08	7.32E-06	cubilin	XP_002734392	0
comp7972_c0_seq4	NA	down	-3.11	4.03	0.000313	cubilin-like	XP_002612977	0
comp7992_c0_seq1	up	NA	2.52	4.11	0.000383	aac4 protein	XP_797207	2.33E-60

Annotations unique to DEGs in GX days 1 and 5 are shown. Significance of differential is expressed as FDR-adjusted p-value, magnitude of differential expression is expressed as log₁₀ fold change over control pool (logFC) and abundance is expressed as log₁₀ counts per million (logCPM). Regulation, or “Reg.”, (up- or down-regulated) is shown for each contig.

Table S4. Numbers of single nucleotide polymorphisms (SNPs)

	Number of Contigs	Number of Contigs with SNPs	Total Number of SNPs	Number of Synonymous SNPs	Number of Non- synonymous SNPs	mean dN/dS
Total Transcriptome	363173	69711	185024	NA	NA	NA
Within determined coding region, no indels allowed	272760	48115	123987	86439	37548	ND
DEG groups						
GX_early DEGs unique	201	56	148	99	49	0.20
GX_early DEGs shared with F3L_early and/or F3L_late	357	105	283	174	109	0.26
Diversified Groups of Interest						
Serine Proteases	73	42	81	56	25	0.07
Serine Protease Inhibitors	81	32	90	28	62	0.19
C1qDC proteins	391	108	251	128	123	0.35
FREPs	276	34	90	58	32	0.22
IAN/GIMAP proteins	173	27	62	36	26	0.25
IFI44	59	11	29	14	15	0.48
CTLDC proteins	276	105	300	209	91	0.19
DMBT1/SRCR type 12	122	56	206	150	56	0.22

B. Supplementary Figures

Figure S1. Sequence read processing statistics in numbers of reads and length of reads.

BIBLIOGRAPHY

- Abascal, F., R. Zardoya, and D. Posada. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*. 21:2104-2105.
- Alexa, A., J.r. Rahnenfuhrer, and T. Lengauer. 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*. 22:1600-1607.
- Allam, B., K.A. Ashton-Alcox, and S.E. Ford. 2001. Haemocyte parameters associated with resistance to brown ring disease in *Ruditapes* spp. clams. *Developmental & Comparative Immunology*. 25:365-375.
- Allam, B., and S.E. Ford. 2006. Effects of the pathogenic *Vibrio tapetis* on defence factors of susceptible and non-susceptible bivalve species: I. Haemocyte changes following in vitro challenge. *Fish & Shellfish Immunology*. 20:374-383.
- Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*. 25:3389-3402.
- Anderson, R.S., G. Ozbay, D.H. Kingsley, and M.A. Strauss. 2011. Oyster Hemocyte Mobilization and Increased Adhesion Activity After β -Glucan Administration. *Journal of Shellfish Research*. 30:635-641.

- Apte, S.S. 2004. A disintegrin-like and metalloprotease (reprolysin type) with thrombospondin type 1 motifs: the ADAMTS family. *The international journal of biochemistry & cell biology*. 36:981-985.
- Azumi, K., R. De Santis, A. De Tomaso, I. Rigoutsos, F. Yoshizaki, M.R. Pinto, R. Marino, K. Shida, M. Ikeda, M. Arai, Y. Inoue, T. Shimizu, N. Satoh, D.S. Rokhsar, L. Du Pasquier, M. Kasahara, M. Satake, and M. Nonaka. 2003. Genomic analysis of immunity in a Urochordate and the emergence of the vertebrate immune system: "waiting for Godot". *Immunogenetics*. 55:570-581.
- Barber, B.J., C.V. Davis, and M.A. Crosby. 1998. Cultured oysters, *Crassostrea virginica*, genetically selected for fast growth in the Damariscotta River, Maine, are resistant to mortality caused by Juvenile Oyster Disease (JOD). *Journal of Shellfish Research*. 17:1171-1176.
- Benjamini, Y., and Y. Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*. 57:289-300.
- Bennett, V., and A.J. Baines. 2001. Spectrin and ankyrin-based pathways: metazoan inventions for integrating cells into tissues. *Physiological reviews*. 81:1353-1392.
- Berridge, M.J. 1984. Inositol trisphosphate and diacylglycerol as second messengers. *Biochemical Journal*. 220:345.
- Boardman, C.L., A.P. Maloy, and K.J. Boettcher. 2008. Localization of the

bacterial agent of juvenile oyster disease (*Roseovarius crassostreae*) within affected eastern oysters (*Crassostrea virginica*). *J Invertebr Pathol.* 97:150-158.

Boettcher, K.J., B.J. Barber, and J.T. Singer. 1999. Use of antibacterial agents to elucidate the etiology of juvenile oyster disease (JOD) in *Crassostrea virginica* and numerical dominance of an alpha-proteobacterium in JOD-affected animals. *Appl Environ Microbiol.* 65:2534-2539.

Boettcher, K.J., K.K. Geaghan, A.P. Maloy, and B.J. Barber. 2005. *Roseovarius crassostreae* sp. nov., a member of the Roseobacter clade and the apparent cause of juvenile oyster disease (JOD) in cultured Eastern oysters. *International Journal of Systematic and Evolutionary Microbiology.* 55:1531-1537.

Boldrick, J.C., A.A. Alizadeh, M. Diehn, S. Dudoit, C.L. Liu, C.E. Belcher, D. Botstein, L.M. Staudt, P.O. Brown, and D.A. Relman. 2002. Stereotyped and specific gene expression programs in human innate immune responses to bacteria. *Proceedings of the National Academy of Sciences.* 99:972-977.

Bolker, B., L. Bonebakker, R. Gentleman, W. Huber, L. A., T. Lumley, M. Maechler, A. Magnusson, S. Moeller, M. Schwartz, and B. Venables. 2010. gplots: Various R programming tools for plotting data. R package version 2.8.0.

Borrego, J.J., A. Luque, D. Castro, J.A. Santamaria, and E. Martinez-

- Manzanares. 1996. Virulence factors of *Vibrio* P1, the causative agent of brown ring disease in the Manila clam, *Ruditapes philippinarum*. *Aquat Living Resour.* 9:125-136.
- Brew, K., D. Dinakarpandian, and H. Nagase. 2000. Tissue inhibitors of metalloproteinases: evolution, structure and function. *Biochimica et biophysica acta.* 1477:267.
- Bricelj, V.M., Ford, S. E., Borerro, F. J., Perkins, F. O., Rivara, G., Hillman, R. E., , and R.A.a.C.J. Elston. 1992. Unexplained mortalities of hatchery-reared, juvenile oysters, *Crassostrea virginica* (Gmelin). *Journal of Shellfish Research.* 11:331-347.
- Brulle, F., F. Jeffroy, S.p. Madec, J.-L. Nicolas, and C. Paillard. 2012. Transcriptomic analysis of *Ruditapes philippinarum* hemocytes reveals cytoskeleton disruption after in vitro *Vibrio tapetis* challenge. *Developmental & Comparative Immunology.* 38:368-376.
- Burke, R.D., M. Lail, and Y. Nakajima. 1998. The apical lamina and its role in cell adhesion in sea urchin embryos. *Cell Communication and Adhesion.* 5:97-108.
- Cancela, M.L., L. Bargelloni, P. Boudry, V. Boulo, J. Dias, A. Huvet, V. Laizé, S. Lapègue, R. Leite, S. Mira, E.E. Nielsen, J.V. Planas, N. Roher, E. Sarropoulou, and F.A.M. Volckaert. 2010. Genomic Approaches in Aquaculture and Fisheries, In: Introduction to Marine Genomics. Vol. 1. J.M. Cock, K. Tessmar-Raible, C. Boyen, and F. Viard, editors. Springer

Netherlands. 213-286.

- Chang, C.I., J.C. Liao, and L. Kuo. 1998. Arginase modulates nitric oxide production in activated macrophages. *Am J Physiol.* 274:H342-348.
- Chen, F., A.J. Mackey, J.K. Vermunt, and D.S. Roos. 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One.* 2:e383.
- Chen, M., H. Yang, B. Xu, F. Wang, and B. Liu. 2008. Catecholaminergic responses to environmental stress in the hemolymph of Zhikong scallop *Chlamys farreri*. *Journal of Experimental Zoology Part A: Ecological Genetics and Physiology.* 309:289-296.
- Clark, M.S., M.A.S. Thorne, F.A. Vieira, J.C.R. Cardoso, D.M. Power, and L.S. Peck. 2010. Insights into shell deposition in the Antarctic bivalve *Laternula elliptica*: gene discovery in the mantle transcriptome using 454 pyrosequencing. *BMC Genomics.* 11:362.
- Colombatti, A., P. Bonaldo, and R. Doliana. 1993. Type A modules: interacting domains found in several non-fibrillar collagens and in other extracellular matrix proteins. *Matrix.* 13:297-306.
- Conesa, A., S. Goetz, J.M. Garcia-Gomez, J. Terol, M. Talon, and M. Robles. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics.* 21:3674-3676.
- Cook, T., M. Folli, J. Klinck, S. Ford, and J. Miller. 1998. The Relationship Between Increasing Sea-surface Temperature and the Northward

- Spread of *Perkinsus marinus* (Dermo) Disease Epizootics in Oysters. *Estuarine, Coastal and Shelf Science*. 46:587-597.
- Davis, C.V., and B.J. Barber. 1999. Growth and survival of selected lines of eastern oysters, *Crassostrea virginica* (Gmelin 1791) affected by juvenile oyster disease. *Aquaculture*. 178:253-271.
- de Lorgeril, J., R. Zenagui, R.D. Rosa, D. Piquemal, and E. Bachere. 2011. Whole Transcriptome Profiling of Successful Immune Response to *Vibrio* Infections in the Oyster *Crassostrea gigas* by Digital Gene Expression Analysis. *PLoS One*. 6:e23142.
- Denison, M.S., and J.P. Whitlock Jr. 1995. Xenobiotic-inducible transcription of cytochrome P450 genes. *Journal of Biological Chemistry*. 270:18175-18178.
- Donaghy, L., C. Lambert, K.S. Choi, and P. Soudant. 2009. Hemocytes of the carpet shell clam *Ruditapes decussates* and the Manila clam *Ruditapes philippinarum*: Current knowledge and future prospects. *Aquaculture*. 297:10-24.
- Dower, K., D.K. Ellis, K. Saraf, S.A. Jelinsky, and L.L. Lin. 2008. Innate immune responses to TREM-1 activation: overlap, divergence, and positive and negative cross-talk with bacterial lipopolysaccharide. *The Journal of Immunology*. 180:3520-3534.
- Du Pasquier, L. 2005. Immunology. Insects diversify one molecule to serve two systems. *Science*. 309:1826-1827.

- Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics*. 14:755-763.
- Egging, D., I. van Vlijmen-Willems, T. van Tongeren, J. Schalkwijk, and A. Peeters. 2007. Wound healing in tenascin-X deficient mice suggests that tenascin-X is involved in matrix maturation rather than matrix deposition. *Connective tissue research*. 48:93-98.
- Enright, A.J., S. Van Dongen, and C.A. Ouzounis. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*. 30:1575-1584.
- Faisal, M., E. MacIntyre, K. Adham, B. Tall, M. Kothary, and J. La Peyre. 1998. Evidence for the presence of protease inhibitors in eastern *Crassostrea virginica* and Pacific *Crassostrea gigas* oysters. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*. 121:161-168.
- Feinberg, T., and S.J. Weiss. 2009. Developmental ECM sculpting: laying it down and cutting it up. *Dev Cell*. 17:584-586.
- Filen, S., and R. Lahesmaa. 2010. GIMAP Proteins in T-Lymphocytes. *Journal of signal transduction*. 2010.
- Finn, R.D., J. Clements, and S.R. Eddy. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*. 39:W29-37.
- Flajnik, M.F., and L. Du Pasquier. 2004. Evolution of innate and adaptive immunity: can we draw a line? *Trends Immunol*. 25:640-644.
- Fleury, E., and A. Huvet. 2012. Microarray analysis highlights immune

- response of Pacific oysters as a determinant of resistance to summer mortality. *Marine biotechnology*. 14:203-217.
- Flierl, M.A., D. Rittirsch, B.A. Nadeau, J.V. Sarma, D.E. Day, A.B. Lentsch, M.S. Huber-Lang, and P.A. Ward. 2009. Upregulation of phagocyte-derived catecholamines augments the acute inflammatory response. *PLoS One*. 4:e4414.
- Ford, S., and R. Smolowitz. 2007. Infection dynamics of an oyster parasite in its newly expanded range. *Marine Biology*. 151:119-133.
- Ford, S.E., and F.J. Borrero. 2001. Epizootiology and pathology of juvenile oyster disease in the Eastern oyster, *Crassostrea virginica*. *J Invertebr Pathol*. 78:141-154.
- Ford, S.E., and M.M. Chintala. 2006. Northward expansion of a marine parasite: Testing the role of temperature adaptation. *Journal of Experimental Marine Biology and Ecology*. 339:226-235.
- Ford, S.E., and M.R. Tripp. 1996. Disease and defense mechanisms. In The eastern oyster: *Crassostrea virginica*. V.S. Kennedy, R.I.E. Newell, and A.F. Eble, editors. Maryland Sea Grant Press, College Park, MD.
- Gomez-Leon, J., L. Villamil, S.A. Salger, R.H. Sallum, A. Remacha-Trivio, D.F. Leavitt, and M. Gomez-Chiarri. 2008. Survival of eastern oysters *Crassostrea virginica* from three lines following experimental challenge with bacterial pathogens. *Diseases of Aquatic Organisms*. 79:95.
- Genard, B., D. Moraga, F. Pernet, E. David, P. Boudry, and R. Tremblay.

2012. Expression of candidate genes related to metabolism, immunity and cellular stress during massive mortality in the American oyster *Crassostrea virginica* larvae in relation to biochemical and physiological parameters. *Gene*. 499:70-75.
- Genard, B., F. Pernet, K. Lemarchand, P. Boudry, D. Moraga, and R. Tremblay. 2011. Physiological and biochemical changes associated with massive mortality events occurring in larvae of American oyster (*Crassostrea virginica*). *Aquat Living Resour.* 24:247.
- Gerdol, M., C. Manfrin, G. De Moro, A. Figueras, B. Novoa, P. Venier, and A. Pallavicini. 2011. The C1q domain containing proteins of the Mediterranean mussel *Mytilus galloprovincialis*: A widespread and diverse family of immune-related molecules. *Dev Comp Immunol.*
- Ghai, R., P. Waters, L.T. Roumenina, M. Gadjeva, M.S. Kojouharova, K. Reid, R.B. Sim, and U. Kishore. 2007. C1q and its growing family. *Immunobiology.* 212:253-266.
- Ghosh, J., K.M. Buckley, S.V. Nair, D.A. Raftos, C. Miller, A.J. Majeske, T. Hibino, J.P. Rast, M. Roth, and L.C. Smith. 2010. Sp185/333: a novel family of genes and proteins involved in the purple sea urchin immune response. *Dev Comp Immunol.* 34:235-245.
- Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol.* 11:725-736.

- Grabherr, M.G., B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B.W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotech.* 29:644-652.
- Griffiths, G., B. Hoflack, K. Simons, I. Mellman, and S. Kornfeld. 1988. The mannose 6-phosphate receptor and the biogenesis of lysosomes. *Cell.* 52:329-341.
- Gueguen, Y., J.P. Cadoret, D. Flament, C. Barreau-Roumiguere, A.L. Girardot, J. Garnier, A. Hoareau, E. Bachere, and J.M. Escoubas. 2003. Immune gene discovery by expressed sequence tags generated from hemocytes of the bacteria-challenged oyster, *Crassostrea gigas*. *Gene.* 303:139-145.
- Guo, X., Y. Wang, L. Wang, and J.H. Lee. 2008. Oysters. *Genome Mapping and Genomics in Fishes and Aquatic Animals*:163-175.
- Hallen, L., Y. Burki, M. Ebeling, C. Broger, F. Siegrist, K. Oroszlan-Szovik, B. Bohrmann, U. Certa, and S. Foser. 2007. Antiproliferative activity of the human IFN- α -inducible protein IFI44. *Journal of Interferon & Cytokine Research.* 27:675-680.
- Hanington, P.C., and S.M. Zhang. 2011. The primary role of fibrinogen-related proteins in invertebrates is defense, not coagulation. *Journal of innate*

immunity. 3:17-27.

- Hansen, K.D., S.E. Brenner, and S. Dudoit. 2010. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* 38:e131.
- Haskins, K.A., J.F. Russell, N. Gaddis, H.K. Dressman, and A. Aballay. 2008. Unfolded Protein Response Genes Regulated by CED-1 Are Required for *Caenorhabditis elegans* Innate Immunity. *Dev Cell.* 15:87-97.
- Haviv, F., M.F. Bradley, D.M. Kalvin, A.J. Schneider, D.J. Davidson, S.M. Majest, L.M. McKay, J. Catherine, R.L. Bell, and B. Nguyen. 2005. Thrombospondin-1 mimetic peptide inhibitors of angiogenesis and tumor growth: design, synthesis, and optimization of pharmacokinetics and biological activities. *Journal of medicinal chemistry*. 48:2838-2846.
- Hertel, L.A., C.M. Adema, and E.S. Loker. 2005. Differential expression of FREP genes in two strains of *Biomphalaria glabrata* following exposure to the digenetic trematodes *Schistosoma mansoni* and *Echinostoma paraensei*. *Developmental & Comparative Immunology*. 29:295-303.
- Hibino, T., M. Loza-Coll, C. Messier, A.J. Majeske, A.H. Cohen, D.P. Terwilliger, K.M. Buckley, V. Brockton, S.V. Nair, K. Berney, S.D. Fugmann, M.K. Anderson, Z. Pancer, R.A. Cameron, L.C. Smith, and J.P. Rast. 2006. The immune gene repertoire encoded in the purple sea urchin genome. *Dev Biol.* 300:349-365.
- Huang, S., X. Wang, Q. Yan, L. Guo, S. Yuan, G. Huang, H. Huang, J. Li, M.

- Dong, S. Chen, and A. Xu. 2011. The evolution and regulation of the mucosal immune complexity in the basal chordate amphioxus. *J Immunol.* 186:2042-2055.
- Huson, D., D. Richter, C. Rausch, T. DeZulian, M. Franz, and R. Rupp. 2007. Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics.* 8:460.
- Ingham, P.W. 1998. The patched gene in development and cancer. *Current opinion in genetics & development.* 8:88-94.
- Jenny, M.J., A.H. Ringwood, E.R. Lacy, A.J. Lewitus, J.W. Kempton, P.S. Gross, G.W. Warr, and R.W. Chapman. 2002. Potential indicators of stress response identified by expressed sequence tag analysis of hemocytes and embryos from the American oyster, *Crassostrea virginica*. *Mar Biotechnol (NY).* 4:81-93.
- Johansson, M.E.V., J.M.H. Larsson, and G.C. Hansson. 2011. The two mucus layers of colon are organized by the MUC2 mucin, whereas the outer layer is a legislator of host-microbial interactions. *Proceedings of the National Academy of Sciences.* 108:4659-4665.
- Johansson, M.W. 1999. Cell adhesion molecules in invertebrate immunity. *Developmental & Comparative Immunology.* 23:303-315.
- Jorieux, S., C. Gaucher, J. Goudemand, and C. Mazurier. 1998. A novel mutation in the D3 domain of von Willebrand factor markedly decreases its ability to bind factor VIII and affects its multimerization. *Blood.*

92:4663-4670.

Katoh, K., and H. Toh. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform.* 9:286-298.

Kennedy, V.S., R.I.E. Newell, and A.F. Eble. 1996. The Eastern Oyster: *Crassostrea virginia*. University of Maryland Sea Grant Press, College Park, MD.

Kent, W.J. 2002. BLAT, the BLAST-like alignment tool. *Genome research.* 12:656-664.

Kim, T., and Y.J. Kim. 2005. Overview of innate immunity in *Drosophila*. *J Biochem Mol Biol.* 38:121-127.

Kishore, U., C. Gaboriaud, P. Waters, A.K. Shrive, T.J. Greenhough, K. Reid, R.B. Sim, and G.J. Arlaud. 2004. C1q and tumor necrosis factor superfamily: modularity and versatility. *Trends in immunology.* 25:551-561.

Klose, R.J., E.M. Kallin, and Y. Zhang. 2006. JmjC-domain-containing proteins and histone demethylation. *Nat Rev Genet.* 7:715-727.

Kolls, J.K., and A. Linden. 2004. Interleukin-17 family members and inflammation. *Immunity.* 21:467-476.

Kong, P., H. Zhang, L. Wang, Z. Zhou, J. Yang, Y. Zhang, L. Qiu, and L. Song. 2010. AiC1qDC-1, a novel gC1q-domain-containing protein from bay scallop *Argopecten irradians* with fungi agglutinating activity. *Developmental & Comparative Immunology.* 34:837-846.

- Kong, Y. 2011. Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. *Genomics*. 98:152-153.
- Kuchel, R.P., A. McCarthy, and D.A. Raftos. 2012. Phenoloxidase activity as an indicator of stress in the silver-lip pearl oyster, *Pinctada maxima*. *Aquaculture*.
- La Peyre, J.F., Q.G. Xue, N. Itoh, Y. Li, and R.K. Cooper. 2010. Serine protease inhibitor cvSI-1 potential role in the eastern oyster host defense against the protozoan parasite *Perkinsus marinus*. *Developmental & Comparative Immunology*. 34:84-92.
- Lacoste, A., F. Jalabert, S.K. Malham, A. Cueff, and S.A. Poulet. 2001. Stress and stress-induced neuroendocrine changes increase the susceptibility of juvenile oysters (*Crassostrea gigas*) to *Vibrio splendidus*. *Applied and Environmental Microbiology*. 67:2304-2309.
- Lamitina, T., and S. Cherry. 2008. Dangerous liaisons: the apoptotic engulfment receptor CED-1 links innate immunity to the unfolded protein response. *Dev Cell*. 15:3-4.
- Lang, R.P., C.J. Bayne, M.D. Camara, C. Cunningham, M.J. Jenny, and C.J. Langdon. 2009. Transcriptome profiling of selectively bred Pacific oyster *Crassostrea gigas* families that differ in tolerance of heat shock. *Marine biotechnology*. 11:650-668.
- Langdon, C., F. Evans, D. Jacobson, and M. Blouin. 2003. Yields of cultured Pacific oysters *Crassostrea gigas* Thunberg improved after one

- generation of selection. *Aquaculture*. 220:227-244.
- Langmead, B. 2010. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics*. Chapter 11:Unit 11 17.
- Langmead, B., and S.L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 9:357-359.
- Langmead, B., C. Trapnell, M. Pop, and S.L. Salzberg. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 10:R25.
- Leonard, P.M., C.M. Adema, S.M. Zhang, and E.S. Loker. 2001. Structure of two FREP genes that combine IgSF and fibrinogen domains, with comments on diversity of the FREP gene family in the snail *Biomphalaria glabrata*. *Gene*. 269:155-165.
- Lewis, E.J. 2001. Juvenile oyster disease (JOD) and management strategies: a review. *Bull Natl Res Inst Aquac*:101–109.
- Li, B., and C.N. Dewey. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 12:323.
- Li, L., C.J. Stoeckert, and D.S. Roos. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research*. 13:2178-2189.
- Li, M.O., M.R. Sarkisian, W.Z. Mehal, P. Rakic, and R.A. Flavell. 2003. Phosphatidylserine receptor is required for clearance of apoptotic cells.

Science. 302:1560-1563.

Li, W., and A. Godzik. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.

Bioinformatics. 22:1658-1659.

Liu, L., J. Yang, L. Qiu, L. Wang, H. Zhang, M. Wang, S. Vinu, and L. Song. 2011. A novel scavenger receptor-cysteine-rich (SRCR) domain containing scavenger receptor identified from mollusk mediated PAMP recognition and binding. *Developmental & Comparative Immunology*. 35:227-239.

Luna-Acosta, A., H. Thomas-Guyon, M. Amari, E. Rosenfeld, P. Bustamante, and I. Fruitier-Arnaudin. 2011. Differential tissue distribution and specificity of phenoloxidases from the Pacific oyster *Crassostrea gigas*. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*. 159:220-226.

Maloy, A.P., B.J. Barber, and K.J. Boettcher. 2005. A PCR-based diagnostic assay for the detection of *Roseovarius crassostreae* in *Crassostrea virginica* affected by juvenile oyster disease (JOD). *Dis Aquat Organ*. 67:155-162.

Massault, C., H. Bovenhuis, C. Haley, and K.D.J. de. 2008. QTL mapping designs for aquaculture. *Aquaculture*. 285:23-29.

McClure, S.J., and P.J. Robinson. 1996. Dynamin, endocytosis and intracellular signalling (Review). *Molecular membrane biology*. 13:189-

215.

- Messier-Solek, C., K.M. Buckley, and J.P. Rast. 2010. Highly diversified innate receptor systems and new forms of animal immunity. *Semin Immunol.* 22:39-47.
- Montagnani, C., F. Le Roux, F. Berthe, and J.M. Escoubas. 2001. Cg-TIMP, an inducible tissue inhibitor of metalloproteinase from the Pacific oyster *Crassostrea gigas* with a potential role in wound healing and defense mechanisms. *FEBS letters.* 500:64-70.
- Morga, B., I. Arzul, B. Chollet, and T. Renault. 2009. Infection with the protozoan parasite *Bonamia ostreae* modifies in vitro haemocyte activities of flat oyster *Ostrea edulis*. *Fish & Shellfish Immunology.* 26:836-842.
- Morga, B., I. Arzul, N. Faury, A. Segarra, B. Chollet, and T. Renault. 2011. Molecular responses of *Ostrea edulis* haemocytes to an in vitro infection with *Bonamia ostreae*. *Developmental & Comparative Immunology.* 35:323-333.
- Mortazavi, A., B.A. Williams, K. McCue, L. Schaeffer, and B. Wold. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods.* 5:621-628.
- National Marine Fisheries Service (NMFS). 2011. Annual Commercial Landing Statistics, Fisheries Statistics.
- Nitta, T., and Y. Takahama. 2007. The lymphocyte guard-IANs: regulation of

- lymphocyte survival by IAN/GIMAP family proteins. *Trends in immunology*. 28:58-65.
- Nitta, T., and Y. Takahama. 2007. The lymphocyte guard-IANs: regulation of lymphocyte survival by IAN/GIMAP family proteins. *Trends in immunology*. 28:58-65.
- Ordas, M.C., A. Ordas, C. Beloso, and A. Figueras. 2000. Immune parameters in carpet shell clams naturally infected with *Perkinsus atlanticus*. *Fish & Shellfish Immunology*. 10:597-609.
- Perez-Morga, D., B. Vanhollebeke, F. Paturiaux-Hanocq, D.P. Nolan, L. Lins, F. Homble, L. Vanhamme, P. Tebabi, A. Pays, and P. Poelvoorde. 2005. Apolipoprotein LI promotes trypanosome lysis by forming pores in lysosomal membranes. *Science*. 309:469-472.
- Paillard, C., K.A. Ashton-Alcox, and S.E. Ford. 1996. Changes in bacterial densities and hemocyte parameters in eastern oysters, *Crassostrea virginica*, affected by juvenile oyster disease. *Aquat Living Resour*. 9:145-158.
- Pancer, Z. 2000. Dynamic expression of multiple scavenger receptor cysteine-rich genes in coelomocytes of the purple sea urchin. *Proceedings of the National Academy of Sciences*. 97:13156-13161.
- Parks, W.C., C.L. Wilson, and Y.S. Lopez-Boado. 2004. Matrix metalloproteinases as modulators of inflammation and innate immunity. *Nature Reviews Immunology*. 4:617-629.

- Philipp, E.E., L. Kraemer, F. Melzner, A.J. Poustka, S. Thieme, U. Findeisen, S. Schreiber, and P. Rosenstiel. 2012. Massively parallel RNA sequencing identifies a complex immune gene repertoire in the lophotrochozoan *Mytilus edulis*. *PLoS One*. 7:e33091.
- Prado-Alvarez, M., C. Gestal, B. Novoa, and A. Figueras. 2009. Differentially expressed genes of the carpet shell clam *Ruditapes decussatus* against *Perkinsus olseni*. *Fish & Shellfish Immunology*. 26:72-83.
- Price, M.N., P.S. Dehal, and A.P. Arkin. 2010. FastTree 2, Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One*. 5:e9490.
- Rawlings, N.D., and A.J. Barrett. 1994. Families of serine peptidases. *Methods in enzymology*. 244:19-61.
- Renault, T., N. Faury, V. Barbosa-Solomieu, and K. Moreau. 2011. Suppression subtractive hybridisation (SSH) and real time PCR reveal differential gene expression in the Pacific cupped oyster, *Crassostrea gigas*, challenged with Ostreid herpesvirus 1. *Developmental & Comparative Immunology*. 35:725-735.
- Roberts, S., Y. Gueguen, J. de Lorgeril, and F. Goetz. 2008. Rapid accumulation of an interleukin 17 homolog transcript in *Crassostrea gigas* hemocytes following bacterial exposure. *Dev Comp Immunol*. 32:1099-1104.
- Robinson, M.D., D.J. McCarthy, and G.K. Smyth. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene

- expression data. *Bioinformatics*. 26:139-140.
- Robinson, M.D., and A. Oshlack. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11:R25.
- Romero, A., N. Estevez-Calvar, S. Dios, A. Figueras, and B. Novoa. 2011. New insights into the apoptotic process in mollusks: characterization of caspase genes in *Mytilus galloprovincialis*. *PLoS One*. 6:e17003.
- Rosa, R.D., J. de Lorgeril, P. Tailliez, R. Bruno, D. Piquemal, and E. Bachere. 2012. A hemocyte gene expression signature correlated with predictive capacity of oysters to survive *Vibrio* infections. *BMC Genomics*. 13:252.
- Ross, J., H. Jiang, M.R. Kanost, and Y. Wang. 2003. Serine proteases and their homologs in the *Drosophila melanogaster* genome: an initial analysis of sequence conservation and phylogenetic relationships. *Gene*. 304:117-131.
- Sato, H., T. Kinoshita, T. Takino, K. Nakayama, and M. Seiki. 1996. Activation of a recombinant membrane type 1-matrix metalloproteinase (MT1-MMP) by furin and its interaction with tissue inhibitor of metalloproteinases (TIMP)-2. *FEBS letters*. 393:101-104.
- Sayers, E.W., T. Barrett, D.A. Benson, E. Bolton, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. Dicuccio, S. Federhen, M. Feolo, I.M. Fingerman, L.Y. Geer, W. Helmberg, Y. Kapustin, S. Krasnov, D. Landsman, D.J. Lipman, Z. Lu, T.L. Madden, T. Madej, D.R. Maglott, A. Marchler-Bauer, V. Miller, I. Karsch-Mizrachi, J. Ostell, A. Panchenko,

- L. Phan, K.D. Pruitt, G.D. Schuler, E. Sequeira, S.T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T.A. Tatusova, L. Wagner, Y. Wang, W.J. Wilbur, E. Yaschenko, and J. Ye. 2012. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 40:D13-25.
- Schlicker, A., F.S. Domingues, J. Rahnenfuhrer, and T. Lengauer. 2006. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics.* 7:302.
- Schulenburg, H., M.P. Hoepfner, J. Weiner III, and E. Bornberg-Bauer. 2008. Specificity of the innate immune system and diversity of C-type lectin domain (CTLD) proteins in the nematode *Caenorhabditis elegans*. *Immunobiology.* 213:237-250.
- Shapiro, S.D. 1998. Matrix metalloproteinase degradation of extracellular matrix: biological consequences. *Current opinion in cell biology.* 10:602-608.
- Silke, J., and D.L. Vaux. 2001. Two kinds of BIR-containing protein-inhibitors of apoptosis, or required for mitosis. *Journal of cell science.* 114:1821-1827.
- Simonet, G., I. Claeys, V. Franssens, A. De Loof, and J.V. Broeck. 2003. Genomics, evolution and biological functions of the pacifastin peptide family: a conserved serine protease inhibitor family in arthropods. *Peptides.* 24:1633-1644.

- Smit, A.F.A., R. Hubley, and P. Green. 1996-2010. RepeatMasker Open-3.0.
- Snyder, M.J. 2000. Cytochrome P450 enzymes in aquatic invertebrates: recent advances and future directions. *Aquatic Toxicology*. 48:529-547.
- Sokolova, I. 2009. Apoptosis in molluscan immune defense. *Invertebrate Surviv J*. 6:49-58.
- Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 22:2688-2690.
- Stamatakis, A., P. Hoover, and J. Rougemont. 2008. A rapid bootstrap algorithm for the RAxML Web servers. *Syst Biol*. 57:758-771.
- Supek, F., M. Bosnjak, N. Skunca, and T. Smuc. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One*. 6:e21800.
- Suzuki, E., and M. Nakayama. 2007. The mammalian Ced-1 ortholog MEGF10/KIAA1780 displays a novel adhesion pattern. *Experimental cell research*. 313:2451-2464.
- Terahara, K., K.G. Takahashi, and K. Mori. 2005. Pacific oyster hemocytes undergo apoptosis following cell-adhesion mediated by integrin-like molecules. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*. 141:215-222.
- Troemel, E.R., S.W. Chu, V. Reinke, S.S. Lee, F.M. Ausubel, and D.H. Kim. 2006. p38 MAPK regulates expression of immune response genes and contributes to longevity in *C. elegans*. *PLoS Genet*. 2:e183.

- Tucker, R., K. Drabikowski, J. Hess, J. Ferralli, R. Chiquet-Ehrismann, and J. Adams. 2006. Phylogenetic analysis of the tenascin gene family: evidence of origin early in the chordate lineage. *BMC Evolutionary Biology*. 6:60.
- Tudor, J.E., P.K. Pallaghy, M.W. Pennington, and R.S. Norton. 1996. Solution structure of ShK toxin, a novel potassium channel inhibitor from a sea anemone. *Nat Struct Biol*. 3:317-320.
- Vanhollebeke, B., and E. Pays. 2006. The function of apolipoproteins L. *Cellular and molecular life sciences*. 63:1937-1944.
- Verhagen, A.M., E.J. Coulson, and D.L. Vaux. 2001. Inhibitor of apoptosis proteins and their relatives: IAPs and other BIRPs. *Genome biology*. 2:reviews3009.
- Villamil, L., J. Gomez-Leon, and M. Gomez-Chiarri. 2007. Role of nitric oxide in the defenses of *Crassostrea virginica* to experimental infection with the protozoan parasite *Perkinsus marinus*. *Dev Comp Immunol*. 31:968-977.
- Vogel, B.E., and E.M. Hedgecock. 2001. Hemicentin, a conserved extracellular member of the immunoglobulin superfamily, organizes epithelial and other cell attachments into oriented line-shaped junctions. *Development*. 128:883-894.
- Wang, S., E. Peatman, H. Liu, D. Bushek, S.E. Ford, H. Kucuktas, J. Quilang, P. Li, R. Wallace, Y. Wang, X. Guo, and Z. Liu. 2010. Microarray

- analysis of gene expression in eastern oyster (*Crassostrea virginica*) reveals a novel combination of antimicrobial and oxidative stress host responses after dermo (*Perkinsus marinus*) challenge. *Fish Shellfish Immunol.* 29:921-929.
- Wang, X., Z. Wu, and X. Zhang. 2010. Isoform abundance inference provides a more accurate estimation of gene expression levels in RNA-seq. *J Bioinform Comput Biol.* 8 Suppl 1:177-192.
- Wang, Z., and X. Li. 2009. IAN/GIMAPs are conserved and novel regulators in vertebrates and angiosperm plants. *Plant signaling & behavior.* 4:165-167.
- Wasmuth, J., and M. Blaxter. 2004. prot4EST: Translating Expressed Sequence Tags from neglected genomes. *BMC Bioinformatics.* 5:187.
- Waterhouse, A.M., J.B. Procter, D.M. Martin, M. Clamp, and G.J. Barton. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics.* 25:1189-1191.
- Watson, F.L., R. Puttmann-Holgado, F. Thomas, D.L. Lamar, M. Hughes, M. Kondo, V.I. Rebel, and D. Schmucker. 2005. Extensive diversity of Ig-superfamily proteins in the immune system of insects. *Science.* 309:1874-1878.
- Weis, W.I., M.E. Taylor, and K. Drickamer. 2006. The C-type lectin superfamily in the immune system. *Immunological reviews.* 163:19-34.
- Wouters, M.A., I. Rigoutsos, C.K. Chu, L.L. Feng, D.B. Sparrow, and S.L.

- Dunwoodie. 2005. Evolution of distinct EGF domains with specific functions. *Protein Sci.* 14:1091-1103.
- Xue, Q.G., G.L. Waldrop, K.L. Schey, N. Itoh, M. Ogawa, R.K. Cooper, J.N. Losso, and J.F. La Peyre. 2006. A novel slow-tight binding serine protease inhibitor from eastern oyster (*Crassostrea virginica*) plasma inhibits perkinsin, the major extracellular protease of the oyster protozoan parasite *Perkinsus marinus*. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*. 145:16-26.
- Yang, J., L. Wang, H. Zhang, L. Qiu, H. Wang, and L. Song. 2011. C-type lectin in *Chlamys farreri* (CfLec-1) mediating immune recognition and opsonization. *PLoS One*. 6:e17089.
- Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586-1591.
- Yu, H., Y. He, X. Wang, Q. Zhang, Z. Bao, and X. Guo. 2011. Polymorphism in a serine protease inhibitor gene and its association with disease resistance in the eastern oyster (*Crassostrea virginica* Gmelin). *Fish & Shellfish Immunology*. 30:757-762.
- Zamocky, M., C. Jakopitsch, P.G. Furtmuller, C. Dunand, and C. Obinger. 2008. The peroxidase-cyclooxygenase superfamily: reconstructed evolution of critical enzymes of the innate immune system. *Proteins: Structure, Function, and Bioinformatics*. 72:589-605.
- Zhang, G., X. Fang, X. Guo, L. Li, R. Luo, F. Xu, P. Yang, L. Zhang, X. Wang,

- H. Qi, Z. Xiong, H. Que, Y. Xie, P.W.H. Holland, J. Paps, Y. Zhu, F. Wu, Y. Chen, J. Wang, C. Peng, J. Meng, L. Yang, J. Liu, B. Wen, N. Zhang, Z. Huang, Q. Zhu, Y. Feng, A. Mount, D. Hedgecock, Z. Xu, Y. Liu, T. Domazet-Loso, Y. Du, X. Sun, S. Zhang, B. Liu, P. Cheng, X. Jiang, J. Li, D. Fan, W. Wang, W. Fu, T. Wang, B. Wang, J. Zhang, Z. Peng, Y. Li, N. Li, J. Wang, M. Chen, Y. He, F. Tan, X. Song, Q. Zheng, R. Huang, H. Yang, X. Du, L. Chen, M. Yang, P.M. Gaffney, S. Wang, L. Luo, Z. She, Y. Ming, W. Huang, S. Zhang, B. Huang, Y. Zhang, T. Qu, P. Ni, G. Miao, J. Wang, Q. Wang, C.E.W. Steinberg, H. Wang, N. Li, L. Qian, G. Zhang, Y. Li, H. Yang, X. Liu, J. Wang, Y. Yin, and J. Wang. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*. 490:49-54.
- Zhang, H., L. Wang, L. Song, X. Song, B. Wang, C. Mu, and Y. Zhang. 2009. A fibrinogen-related protein from bay scallop *Argopecten irradians* involved in innate immunity as pattern recognition receptor. *Fish & Shellfish Immunology*. 26:56-64.
- Zhang, S.M., C.M. Adema, T.B. Kepler, and E.S. Loker. 2004. Diversification of Ig superfamily genes in an invertebrate. *Science Signalling*. 305:251.
- Zhao, X., H. Yu, L. Kong, and Q. Li. 2012. Transcriptomic Responses to Salinity Stress in the Pacific Oyster *Crassostrea gigas*. *PLoS One*. 7:e46244.
- Zhou, Z., D. Ni, M. Wang, L. Wang, X. Shi, F. Yue, R. Liu, and L. Song. 2012.

The phenoloxidase activity and antibacterial function of a tyrosinase from scallop *Chlamys farreri*. *Fish & Shellfish Immunology*.

Zhou, Z., L. Wang, X. Shi, H. Zhang, Y. Gao, M. Wang, P. Kong, L. Qiu, and L. Song. 2011. The modulation of catecholamines to the immune response against bacteria *Vibrio anguillarum* challenge in scallop *Chlamys farreri*. *Fish & Shellfish Immunology*.

Zhou, Z., L. Wang, J. Yang, H. Zhang, P. Kong, M. Wang, L. Qiu, and L. Song. 2011. A dopamine beta hydroxylase from *Chlamys farreri* and its induced mRNA expression in the haemocytes after LPS stimulation. *Fish & Shellfish Immunology*. 30:154-162.