Multi-species protein similarity clustering reveals novel expanded immune gene families in the Eastern oyster *Crassostrea virginica*

Ian C. McDowell, Tejashree H. Modak, Chris Lane, and Marta Gomez-Chiarri*

University of Rhode Island, Kingston, Rhode Island, USA

* Email: gomezchi@mail.uri.edu

ABSTRACT

Comparative genomics research in non-model species has highlighted how invertebrate hosts possess complex diversified repertoires of immune molecules. The levels of diversification in particular immune gene families appear to differ between invertebrate lineages and even between species within lineages, reflecting differences not only in evolutionary histories, but also in life histories, environmental niches, and pathogen exposures. The goal of this research was to identify immune-related gene families experiencing high levels of diversification in eastern oysters, *Crassostrea virginica*. Families containing 1) transcripts differentially expressed in eastern oysters in response to bacterial challenge and 2) a larger number of transcripts compared to other species included those coding for the C1q and C-type lectin domain containing proteins (C1qDC and CTLDC), GTPase of the immune-associated proteins (GIMAP), scavenger receptors (SR), fibrinogen-C domain containing proteins (also known as FREPs), dopamine beta-hydroxylase (DBH), interferon-inducible 44 (IFI44), serine protease inhibitors, apextrin, and dermatopontin. Phylogenetic analysis of two of the families significantly expanded in bivalves, IFI44 and GIMAP, showed a patchy distribution within both protostomes and deuterostomes, suggesting multiple independent losses and lineage-specific expansions. Increased availability of genomic information for a broader range of non-model species broadly distributed through vertebrate and invertebrate phyla will likely lead to improved knowledge on mechanisms of immune-gene diversification.

Abbreviations: C1qDC: C1q domain-containing; CTLDC: C-type lectin domain-containing; DBH: dopamine beta-hydroxylase; Fib_C D: fibrinogen C domain-containing; GIMAP: GTPase of the immunity-associated protein; IFI44: interferon-induced protein 44; SPI: serine protease inhibitor; SR: scavenger receptor cysteine-rich.

1. Introduction

In the last few decades, research into invertebrate immunity has led to the description of genes and mechanisms accounting for effective and flexible defense responses in the absence of adaptive immunity (Ghosh et al. 2011; Liberti et al. 2015; Loker et al. 2012; Messier-Solek et al. 2010). Despite lacking the immunoglobulin or variable lymphocyte receptor repertoires characteristic of adaptive immune systems, invertebrate hosts have developed complex diversified sets of immune receptors, regulators, and effectors. Similarly to vertebrates, strategies of diversification in immune molecules described in invertebrates include allelic diversity/SNPs/indels, alternative splicing, gene family expansion, somatic diversification, and post-translational modifications. These mechanisms can act in combination to lead to extreme levels of diversification (reviewed in Ghosh et al. 2011; Liberti et al. 2015; Loker et al. 2012; Messier-Solek et al. 2010).

The levels of diversification in particular immune gene families differ between invertebrate lineages and even between species within lineages, reflecting differences not only in evolutionary histories, but also in life histories and environmental niches among this large paraphyletic group of organisms and the strong selective forces exerted by pathogens on their hosts. For example, in Pacific oysters, a comprehensive analysis of transcriptomes in response of a variety of immune and stress stimuli revealed a complex repertoire of highly diversified gene families, including Toll-like receptors (TLRs) and tumor necrosis factor (TNF) among several others, that expanded through tandem gene duplications and lineage specific expansions (Zhang et al. 2015). Extensive expansions through tandem duplications of families of receptors binding to pathogen associated molecular patterns (PAMPs) have been observed in the genome of the sea urchin *Strongylocentrotus purpuratus* (Hibino et al. 2006). Fibrinogen-related proteins (FREPs), which participate in agglutination and phagocytosis in arthropods and molluscs (Hanington & Zhang 2011), appear to have attained diversity through multiple strategies of gene family expansion, alternative splicing, allelic diversity, and even somatic recombination (Ghosh et al. 2010; Hanington et al. 2010; Zhang et al. 2004). Other instances of innate immune diversification include the remarkable expansion of the C1q domain-containing (C1qDC) gene family in the mussel *Mytilus galloprovincialis* (Gerdol & Venier 2015) and the Pacific oyster (Zhang et al. 2015), the expansion of the Down-syndrome cell adhesion molecule (Dscam) gene

in mosquitos achieved through alternative splicing of single copy genes (Brites et al. 2013), the expansion through multiple mechanisms of *He185/333* – *Sp185/333* genes and scavenger receptor genes in sea urchins (Pancer 2000; Roth et al. 2014), the highly polymorphic chitin binding proteins in cephalochordates (Dishaw et al. 2011), and the expansion of C-type lectin domains (CTDL) in crustaceans, insects, and the nematode *Caenorhabditis elegans* (Pees et al. 2015). Exploration of immune repertoires in other invertebrate species may reveal additional examples of gene families with high levels of diversification resulting from the selective pressures provided by a large variety of environmental niches.

A recent investigation into the transcriptomic response of eastern oysters, *C. virginica*, in response to bacterial challenge demonstrated the importance of recognition and binding, signaling, cell adhesion, extracellular matrix remodeling, inflammation, detoxification, and apoptosis in the host defense response of the eastern oyster. Several large immune-related gene families enriched for differentially expressed transcripts upon bacterial challenge were identified using an enrichment analysis (McDowell et al. 2014). The goal of this research was to determine if any of these large gene families were significantly expanded in oysters as compared to other invertebrate species, suggesting an important role of these genes and associated pathways in immune responses in oysters. Sequences for the identified gene families from 13 evolutionarily diverse invertebrate species and 1 vertebrate species were clustered into multi-species gene families using the TribeMCL clustering pipeline in order to identify those gene families showing expansions in oysters. Numbers of genes in each family were compared between species. Expression of *C. virginica* gene family members following challenge by the bacterial pathogen *Alliostoeovarius crassostreae* was analyzed. Two families showing significant levels of gene expansion not previously reported in the literature, GIMAP and IFI44, were investigated in more depth phylogenetically and through network analysis.

2. Materials and Methods

2.1. Bacterial challenge of eastern oysters, cDNA sequencing, and assembly

This analysis used data from eastern oyster transcriptomes obtained in a previous study (McDowell et al. 2014). Briefly, juvenile eastern oysters from 2 families with known differential susceptibility to Roseovarius Oyster Disease (resistant and susceptible) were exposed by

immersion to the bacterial pathogen *Alliroseovarius crassostreae*, strain CV919-312^T (previously known as *Roseovarius crassostreae*; Boettcher et al. 2005). Total RNA from resistant, susceptible, and control oysters was collected at days 1, 5, 15, and 30 after challenge, from which cDNA libraries were prepared, sequenced on the Illumina GAIIx platform, and de-novo assembled using Trinity with default parameters. Families of immune-related genes or domains (as determined by Pfam) that were enriched for differentially expressed transcripts upon bacterial challenge were identified using an enrichment analysis (McDowell et al. 2014).

2.2. Peptide similarity clustering

Transcripts were translated into peptides using prot4EST. A multi-species sequence similarity graph was constructed using the TribeMCL clustering pipeline (Enright 2002) using: translated *C. virginica* transcripts (McDowell et al. 2014) along with translated ESTs from the Pacific oyster *C. gigas*, the gastropods *Biomphalaria glabrata*, *Lottia gigantea*, and *Aplysia californica*, and protein sequences (translated gene models) from the pearl oyster *Pinctada fucata*, the gastropod *L. gigantea*, the basal eumetazoan *Nematostella vectensis*, the deuterostome *Strongylocentrotus purpuratus*, the proto-chordates *Branchiostoma floridae* and *Ciona intestinalis*, the hemichordate *Saccoglossus kowalevskii*, the basal vertebrate *Petromyzon marinus*, and the ecdysozoans *Drosophila melanogaster* and *Daphnia pulex* (Table S1). For codon usage bias training sets, all full-length Bivalvia Uniref100 proteins were used for *C. virginica* and *C. gigas*, and all full-length Gastropoda Uniref100 proteins were used for *A. californica*, *B. glabrata*, and *L. gigantea*. 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 (Fischer et al. 2011). Proteins were compared in a parallel run of all-versus-all BLASTP, retaining hits with e-value $\leq 1e-10$ (Altschul et al. 1997). The orthomcl utility orthomclBlastParser was used to calculate the percent identity of each hit. Negative log₁₀-transformed e-values above 99 were set to 99 so that subsequent edge weights for MCL clustering were not skewed (Frech & Chen 2010). The Markov cluster (MCL) algorithm was set on the default inflation index, 1.5, and was used to dissolve less reliable edges, which may result from sequence similarity errors or from shared protein domains across families (Enright et al. 2002).

2.3. Selection and manual annotation of clusters of interest

Among the clusters created using the TribeMCL clustering pipeline, several gene families of interest were chosen for further study based on the following criteria: (1) the presence of transcripts that were differentially expressed in either resistant or susceptible *C. virginica* upon bacterial challenge and corresponded to immune-related Pfam families and/or domains that were enriched for differentially expressed genes as determined in (McDowell et al. 2014); and (2) the relatively abundance of oyster transcripts present in the cluster as compared to number of transcripts or genes from other species. Each of the sequences in these gene families was manually annotated to confirm assignation to the gene family. Manual annotation was done by BLASTP to reference sequences taken from UniprotKB or Pfam representing the gene families of interest (Table S2) and to Uniref90. Criteria used to assign a sequence to a particular gene family included: (1) having a BLASTP hit with e-value $\leq 1e-05$ and percent identity $\geq 60\%$ with at least one other member of the family; (2) not having a best-scoring BLASTP hit to Uniref90 conflicting with the gene family of interest (e.g. "Agrin" is inconsistent with the C1qDC family); and (3) for gene families containing characteristic domains (e.g. C-type lecting domain) having a significant relevant hit to PatternScan, HMMPfam, HMMSmart, FPrintScan, HMMPanther, or superfamily, through InterProScan search (version 4.8; Table S3). Those translated transcripts and gene models that passed the above criteria were clustered into non-redundant sets using CD-HIT, on settings "-G 0 -aS 0.60 -c 0.95," which reduced the sets of sequences by 95% redundancy for an alignment length at least 60% of the length of the smaller sequences of the clustered set (Li & Godzik 2006). Sequences were clustered by CD-HIT only within species for proper enumeration of gene family members by species.

2.4. Differential Gene Expression in selected gene families

For each of the *C. virginica* transcripts included in the multi-species clusters selected for further study, as described above, levels of differential gene expression in each of the treatment groups (resistant or susceptible at 1, 5, 15, or 30 days after bacterial challenge) was calculated as compared to the control, non-challenged resistant oysters as previously described (McDowell et al. 2014). Transcript abundances were estimated using RSEM (RNA-Seq by Expectation Maximization) through the Trinity plug-in, run_RSEM.pl (Grabherr et al. 2011) with normalization factors calculated using the trimmed mean of M values (TMM) method (Robinson and Oshlack 2010). Read abundances for contigs in each of the samples were compared to read

abundances in the control pool using edgeR (Robinson & Oshlack 2010). Significance values yielded by hypergeometric test were adjusted using the False Discovery Rate (FDR) correction and contigs were considered differentially expressed (DE) with an FDR-adjusted p-value ≤ 0.05 (Benjamini and Hochberg 1995).

2.5. Estimation of gene models by mapping of transcripts to *C. gigas* genome

All non-redundant *C. virginica* and *C. gigas* transcripts from the gene families of interest were mapped to the genome of the closest available relative, *C. gigas* (Zhang et al. 2012), using sim4db (Walenz & Florea 2011) with mappings retained where gene model length was ≥ 50 bp and identity was $\geq 60\%$ and $\geq 80\%$ for *C. virginica* (cross-species) and *C. gigas*, respectively. Only those mRNA models with at least 50 identically mapped nucleotides were retained. Only the best mRNA model (greatest number of identically mapped nucleotides) for each transcript was retained. The number of gene models or transcripts for each species was Z-score centered by gene family using the program R (Becker et al. 1988).

2.6. Phylogenetic analysis of selected gene families

Two families, GIMAP and IFI44, were selected for further study due to the fact that they had not been previously described as expanded in oysters. These sequences were searched for in all the genomes and transcriptomes in table S1 as well as in additional selected species (Figs. S1 and S2). Non-redundant gene family sequences and selected Swiss-Prot sequences were iteratively aligned at the nucleotide level, independently for each gene family (Figs. S1 and S2). The iterative alignment process involved alignment by MAFFT-EINSI (Katoh et al. 2009) and manual curation in JalView (Waterhouse et al. 2009) whereby insertions were removed where $\geq 75\%$ of sequences had gaps. The final IFI44 alignment totalled 780 bp of the most conserved region of the alignment, which included all IFI44-associated GTP binding motifs (Fig S1). Only sequences ≥ 585 bp in length (at least 75% occupancy of alignment) were retained. The final GIMAP alignment totalled 380 bp of the most conserved region of the alignment, which included the characteristic G1 through G4 GTP-binding motifs (Fig S2). Only sequences ≥ 341 bp in length (at least 90% occupancy of alignment) were retained. The final curated nucleotide alignments were confirmed to be codon-aligned by extracting selected nucleotide sequences from the multiple sequence alignment and aligning these to their corresponding Swiss-Prot

amino acid sequences using GeneWise (Birney et al. 2004). Amino acid alignments were generated from the translated nucleotide alignments. The models of amino acid substitution that best fit the multiple sequence alignments were found using ProtTest (version 3, Abascal et al. 2005). Phylogenetic trees were constructed for GIMAP and IFI44 alignments by partitioned analysis with the first and second coding positions as one partition, the third coding position as another partition, and the amino acid alignment as the final partition. Maximum likelihood (ML) tree searching and bootstrapping was performed with RAxML (version 7.5.4) using the GTR model and CAT approximation of rate heterogeneity for nucleotide partitions. The LG substitution model (Le & Gascuel, 2008) and CAT approximation (Stamatakis 2006) were used for the amino acid partition (for both GIMAP and IFI44) and fixed empirical amino acid frequencies were also used for the GIMAP amino acid partition. The best tree from 1000 ML tree search replicates was bootstrapped with RAxML standard bootstrap search replicates using the MRE-based bootstrapping criterion

Figure 1. Multi-species sequence similarity graph for IFI44 and GIMAP. The graphs, built using TribeMCL, show relationships between (A) IFI44 and (B) GIMAP transcripts from the organisms listed in the tree adjacent to the graph. The colored circles indicate nodes (sequences) and the yellow lines indicate connections between nodes. Black nodes mark *C. virginica* transcripts that showed up-regulation in response to bacterial challenge, a white node a down-regulated transcript.

(GIMAP: 576 replicates, IFI44: 192 replicates). Bootstrap support was optimized by pruning “rogue” sequences (7 sequences pruned from GIMAP, 1 pruned from IFI44; Aberer et al. 2013).

3. Results

3.1. Multi-species gene cluster analysis identifies gene families expanded in molluscs.

Several immune-related gene families containing genes differentially expressed in resistant and susceptible oysters in response to bacterial challenge and showing expansion in transcript or gene model numbers as compared to other species included in the analysis were identified using multi-species cluster analysis. Representative images of the multi-family clusters used to identify and select gene families for further study are shown in Fig. 1. These clusters show a relative dominance of eastern oyster transcripts (in grey), including some differentially expressed in resistant oysters upon bacterial challenge (in white or black), suggesting a potential role in

immune responses. Expanded gene families in oysters included those coding for the C1q and C-type lectin domain containing proteins (C1qDC and CTLDC), GTPase of the immunity-associated protein (GIMAP), scavenger receptors (SR), fibrinogen-C domain containing proteins (Fib-CDC), dopamine beta-hydroxylase (DBH), interferon-inducible

44 (IFI44), apextrin, and dermatopontin-2 (Fig. 2, Table S4). Several of these gene families (C1qDC, CTLDC, DBH, GIMAP, and IFI44) showed higher levels of expansion in the bivalve species (pearl oysters *P. fucata*, and the eastern and Pacific oysters), than in the 3 gastropods or

most other species included in the analysis. Very few IFI44 transcripts were found in non-

Figure 2. Expansion of selected immune-relevant gene families in diverse species. The number of sequences in each family for each of the species was Z-score centered within gene family and the magnitude of this Z-score was assigned a color according to the color gradient indicated in the key, with warmth of color positively associated with number of sequences. A species tree is reproduced beside the matrix to emphasize evolutionarily relationships between the featured organisms, with shading to separate major phylogenetic groups. Green: gastropod species; light red: bivalves; lilac: deuterostomes. For numbers of genes/transcripts, see table S4

bivalve invertebrate species (Fig. 1B, Fig. 2). Other families also showed evidence of gene expansion in selected species randomly distributed through the different evolutionary lineages, suggesting independent gene expansions within different lineages. For example, a remarkable expansion of C1qDC, with more than 100 genes in each species, was observed in bivalves and in the arthropod *Da. pulex* (Fig. 2). Large gene-family expansions were observed for CTLDC and Fib_C DC in bivalves and the deuterostome *B. floridae*, and for SR in sea urchins, and, to a lesser extent, in bivalves and the gastropod *L. gigantea*. The GIMAP family had a high relative abundance of sequences in bivalves, the gastropod *L. gigantea*, and the deuterostome *B. floridae* (Figs. 1A, Fig. 2).

3.2. Differential gene expression in *C. virginica* transcripts in selected expanded gene families

Several transcripts in these gene families were differentially expressed in *C. virginica* upon bacterial challenge (Fig. 3A). In all gene families, there were a greater number of differentially expressed transcripts in susceptible oysters than in resistant oysters (Fig. 3B). Within gene family and among differentially expressed transcripts, magnitude of fold change of expression in

Figure 3. Differential gene expression in oysters in response to bacterial challenge. (A) Proportion of transcripts in each gene family showing differential gene expression in oysters in response to bacterial challenge (black). The size of each pie chart is proportional to the total number of transcripts in each gene family. (B) Number of differentially expressed transcripts in each experimental group. Light blue: resistant oysters, 1 and 5 dpc; dark blue: resistant oysters, 15 and 30 dpc; pink: susceptible oysters, 1 and 5 dpc; red: susceptible oysters, 15 and 30 dpc. For example, for apextrin, 1 transcript was differentially expressed (DE) in susceptible oysters at 1 and 5 dpc, 2 transcripts were DE in susceptible oysters at all times, and 1 transcript was DE in all treatments (C) Fold change in gene expression at 1, 5, 15, and 30 dpc. Values represent the mean and standard deviation for all transcripts DE in each treatment. Negative values indicate downregulation. Blue: resistant oysters; red: susceptible oysters.

challenged oysters over non-challenged oysters was more variable in susceptible than resistant oysters. In general, expression of transcripts within a gene family differed between resistant and susceptible oysters (Fig. 3C). Within SRs, for example, multiple transcripts were highly up-regulated throughout the course of the bacterial challenge in susceptible oysters while only one transcript was up-regulated in resistant oysters and only at 1 d after challenge (Fig. 3C). For GIMAP, most differentially expressed transcripts were detected in susceptible oysters, showing down-regulation. For IFI44, differentially expressed transcripts were mostly either down-regulated in susceptible oysters or up-regulated in resistant oysters (Fig. 3C).

3.3. Estimation of gene models oysters

Table 1. Mapping of oyster (*C. virginica* and *C. gigas*) transcripts to the *C. gigas* genome.

Gene Family	Species	Number of Transcripts ^a	Number (percent) of mapped transcripts ^b	Number of genes ^c
Apextrin	<i>C. virginica</i>	14	12 (86%)	8-9
	<i>C. gigas</i>	10	9 (90%)	5
C1qDC	<i>C. virginica</i>	154	124 (81%)	80-107
	<i>C. gigas</i>	208	190 (91%)	125-153
CTLDC	<i>C. virginica</i>	130	90 (69%)	55-61
	<i>C. gigas</i>	144	138 (96%)	82-94
DBH	<i>C. virginica</i>	32	23 (72%)	15
	<i>C. gigas</i>	27	26 (96%)	21-23
Dermatopontin	<i>C. virginica</i>	5	5 (100%)	3
	<i>C. gigas</i>	6	6 (100%)	4
Fib_C DC	<i>C. virginica</i>	47	40 (85%)	35-43
	<i>C. gigas</i>	50	46 (92%)	42
GIMAP	<i>C. virginica</i>	101	83 (82%)	26-28
	<i>C. gigas</i>	27	18 (67%)	14-17
IFI44	<i>C. virginica</i>	39	17 (44%)	9-13
	<i>C. gigas</i>	38	35 (92%)	16-18
SPI	<i>C. virginica</i>	8	8 (100%)	7
	<i>C. gigas</i>	20	20 (100%)	10
SR	<i>C. virginica</i>	47	42 (89%)	27-33
	<i>C. gigas</i>	29	27 (93%)	22-24

^aTranscripts are non-redundant (by CD-HIT) and translate to amino acid sequences ≥ 100 residues. ^b Mapping of *C. virginica* and *C. gigas* transcripts to the *C. gigas* genome was performed using sim4db (sim4cc) using percent identity ≥ 60 and ≥ 80 , respectively, with individual exon length ≥ 50 bp. As a further filter, only those mRNA models with at least 200 identically mapped nucleotides were retained. ^cGenes by mapping was determined by extending the span of mapped genomic target coordinates 5 and

10 Kb up- and down-stream for each accepted hit (corresponding to high estimate and low estimate of the range given above, respectively), then merging all overlapping target regions and enumerating the total number of non-overlapping genomic target spans.

Gene models for each diversified family was calculated in oysters *C. gigas* and *C. virginica* by mapping transcripts to the *C. gigas* genome (Table 1). The number of transcripts in each family ranged from 5 – 6 for dermatopontin-2 to 158 – 208 for C1qDC. The percent of oyster transcripts mapping to the Pacific oyster genome ranged from 44% for IFI44 to 100% for SPI and dermatopontin. Most families showed a similar number of gene models in both oyster species. Some exceptions to this were the families of C1qDC, CTLDC, and DBH, where more gene models were found in *C. gigas* than *C. virginica*. In contrast, more gene models were found in *C. virginica* for GIMAP and apextrin. Differences in the number of gene models for each species could be due to the number of transcripts detected in each species (which would depend on the sets of experimental conditions used) and divergence in sequence between oysters species within a family, which may prevent mapping using the relatively stringent parameters used in our study.

3.4. Phylogenetic analysis of IFI44 and GIMAP gene families

Of the species included in this study, IFI44 was present in the basal anthozoan *N. vectensis* and appears to have been lost in the deuterostome *S. purpuratus*, the protostomes *D. pulex* and *D. melanogaster*, and the two gastropods, *B. glabrata* and *A. californica* (Fig. 1A). The maximum likelihood tree of the IFI44 family separated sequences into two clades with strong bootstrap support, with one containing the lophotrochozoan sequences and the other containing the protochordate sequences, vertebrate sequences, and a sequence from the anthozoan *N. vectensis* (Fig. 4).

[insert figure 4 here]

Based on estimates of gene numbers in the species included in this study, GIMAP genes were not found in the deuterostomes *C. intestinalis*, *S. purpuratus*, and *P. marinus* and the protostomes *Da. pulex* and *Dr. melanogaster* (Fig. 1B). Phylogenetic analysis of GIMAP sequences included the related *Arabidopsis thaliana* GIMAP/AIG1 sequences showed a clade of plant-only sequences that was basal to all animal sequences in the maximum likelihood tree (not shown). A phylogenetic tree constructed with just animal sequences, including human GIMAPs,

showed a clade including the sequences from humans and some pearl oyster *P. fucata* sequences, another clade with bivalve-only sequences, and a more speciose cluster of mostly lineage-specific sequences from bivalves, gastropods, and two proto-chordates. Many of the subclades within this cluster show expansions of bivalve sequences (Fig. 5).

[insert figure 5 here]

4. Discussion

Recent comparative genomics studies in invertebrates have led to the identification of immune-related gene families with remarkable levels of sequence diversity within particular species or evolutionary lineages (e.g., Gerdol & Venier 2015; Liberti et al. 2015; Loker 2012; Messier-Solek et al. 2010). This study, focused on the analysis of immune-related transcripts in the eastern oyster *C. virginica*, identified several gene families that are significantly more diverse in oysters as compared to other invertebrate species. These results are consistent with several other recent studies showing lineage-specific major expansions of certain immune-related gene families, including C1qDC, CTLDC, and Fib_C DC in bivalve species (Gerdol & Venier 2015; Zhang et al. 2014; Zhang et al. 2015). We also report here several gene families that, to our knowledge, had not been previously described to be expanded in bivalves, such as GIMAP, DBH, and dermatopontin. The GIMAP and IFI44 families were studied in more depth because of the pattern seen in the multi-species protein clustering suggesting independent gene losses and lineage-specific expansions. The gene families identified in this research are prime targets for further research into the mechanisms driving diversity in innate immune responses in invertebrate species.

A novel finding of this research is the high and unprecedented diversity of GIMAP transcripts (101) seen in the transcriptomes of eastern oysters in response to bacterial challenge, which is remarkable considering that GIMAP genes are completely absent in many vertebrate and invertebrate species. GTPase of the immunity-associated proteins (GIMAPs), also known as immune-associated-nucleotide-binding proteins (IANs), are G proteins with one or two GTPase domain (Ciucci & Bosselut 2014; Heinonen et al. 2015). These proteins were first identified in plants as responsible for disease resistance to bacterial infection (Reuber & Ausubel 1996). In vertebrates, GIMAPs are involved in the development of thymocytes and the regulation of

apoptosis (Ciucci & Bosselut 2014; Filen and Lahesmaa 2010). Several (4) GIMAP transcripts were significantly upregulated in the amebocyte-producing organ (the site of hemocyte production) of the snail *B. glabrata* in response to immune stimuli (Zhang et al. 2016). In our study, GIMAP transcripts were mostly down-regulated in *C. virginica* following bacterial challenge and may promote the survival of hemocytes by negatively regulating apoptosis as they have been proposed to do in humans (Webb et al. 2016). In addition to GIMAPs, an expanded repertoire of apoptosis-related proteins, including tumor necrosis factor receptors, inhibitor of apoptosis proteins, and caspases, were found to be expanded in the Pacific oyster *C. gigas* and other bivalves (Zhang et al. 2012; Zhang et al. 2015).

As a family, GIMAPs are at least as ancient as the divergence of plants and animals (Filen & Lahesmaa 2010; Wang & Li 2009). While no GIMAP proteins were found in the two protostomes herein studied, *Dr. melanogaster* and *Da. pulex*, many GIMAP sequences were found in molluscs and especially bivalves. Our results are consistent with recent findings of a study in the coral *Acropora millepora*, which identified GIMAP-like transcripts in corals exposed to immune challenge (Weiss et al. 2013). Phylogenetic analysis including coral GIMAP sequences showed a patchy distribution suggesting multiple gene losses and independent lineage-specific expansions occurring within protostomes (including a lophotrochozoan species, mainly gastropods) and within deuterostomes (including several proto-chordate and chordate species) (Weiss et al. 2013). Our study adds bivalves to the list of species showing a remarkable level of diversity in GIMAP genes, comparable to what has been observed in the zebrafish *Danio rerio* (Weiss et al. 2013). There are 26 GIMAP genes annotated in the *C. gigas* genome (Zhang et al. 2012), which agrees well with the 26-28 unique *C. gigas* genomic loci to which the *C. virginica* transcripts in the present study mapped and the 27 GIMAP gene models in the *P. fucata* predicted proteome. These results suggest that the diversification of the GIMAP family in bivalves appears to have occurred by tandem gene duplication, further amplified through allelic variation. Access to the *C. virginica* genome (in progress, Gomez-Chiarri et al. 2015) will allow future research into the mechanisms of GIMAP diversification in oysters, and to determine if GIMAP loci are clustered in the oyster genomes, as reported for corals, humans, and plants (Wang & Li 2009; Weiss et al. 2013).

Another gene family proposed here as a novel diversified mediator of immune responses in oysters is the IFI44 gene family. In vertebrates, interferon-induced protein 44 is inducible by interferon- α and is implicated in antiviral host defense (Kitamura et al. 1994), showing antiproliferative activity possibly through cell cycle arrest (Hallen et al. 2007). IFI44-like transcripts are upregulated by a pathogenic *Vibrio* in *C. gigas* (de Lorgeril et al. 2011), and 8 IFI44 transcripts show variability in basal gene expression in *C. gigas* hemocytes from oysters resistant and susceptible to bacterial challenge (Rosa et al. 2012). Moreover, expression of an IFI44 transcript in *C. gigas* can be primed by exposure to poly I:C (Green & Montagnani 2013). These results support a role of IFI44 in immune responses in oysters in response to a variety of immune challenges. In our study, we detected 39 IFI44-like transcripts differentially expressed in *C. virginica* in response to bacterial challenge. This is comparable to the diverse set of up to 24 identified IFI44 transcripts are up-regulated in *C. gigas* in response to challenge by the virus OsHV-1 (Renault et al. 2011; He et al. 2015). Phylogenetic analysis of IFI44 sequences supports the hypothesis of a lineage-specific expansion through gene duplication that appears, so far, to be unique to bivalves, and gene losses in a variety of species within protostomes and deuterostomes.

The relative expansion in diversity of dopamine beta-hydroxylase genes in bivalves is another novel finding in our study. DBH catalyzes the conversion of dopamine to norepinephrine, which are both signaling catecholamines shown to be important in the neuroendocrine-immune regulatory system in vertebrates (Sanders & Straub 2002) and in immunity in several invertebrate phyla (Ottaviani et al. 1996). Dopamine beta-hydroxylase modulates phagocytosis in *C. gigas* (Lacoste et al. 2001). The expression of DBH is induced in the scallop *Chlamys farreri* in response to LPS (Zhou et al. 2011) and heat-killed *Vibrio anguillarum* (Zhou et al. 2012). Consistent with these reports, DBH transcripts were up-regulated in eastern oysters in response to bacterial challenge. The number of non-redundant DBH sequences found in each oyster species (27-32) is comparable in number to the 29 DBH gene models found in *B. floridae*. Presently, only 10-12 DBH genes are annotated in the *C. gigas* genome, but this may be an underestimation, given that *C. gigas* transcripts mapped to 21-23 unique genomic loci in this study. Our findings indicate that much of the diversity in this gene family resulted from gene duplication.

The finding of an expansion of C1qDC and CTLDC gene families in *C. virginica* is consistent with previous reports in other bivalve species (Gerdol et al. 2011; Gerdol & Venier 2015). C1q domain-containing proteins, through self and non-self binding, function in a variety of processes such as agglutination, cell adhesion, inflammation, and clearance of apoptotic bodies (Kishore et al. 2004). The expansion of the C1qDC family is particularly remarkable in mussels, with more than 1200 putative loci detected in the draft genome of the mussel *M. galloprovincialis* (Gerdol & Venier 2015). In addition to massive sequence diversity the C1qDC family in the mussel shows high tissue-specific expression in hemocytes and other tissues, and its members are induced by bacterial challenge (Gerdol et al. 2011). Our finding that *C. virginica* expresses a great number of C1qDC proteins in response to bacterial challenge confirms that many of these proteins likely play a role in host defense against gram-negative bacteria in bivalves.

The C-type lectin domain-containing (CTLDC) family was the second largest immune gene family, after C1qDC in the bivalve species included in our analysis. CTLDC proteins are extracellular proteins that contain conserved carbohydrate recognition, CTL or CRD, domains and function in processes as diverse as fertilization, cell adhesion, endocytosis, activation of antimicrobials, and pathogen recognition and agglutination (Gerdol & Venier 2015). It has been suggested that the more than 200 CTLDC genes in the worm *Caenorhabditis elegans* represent a diversified immune gene family conferring specificity to different pathogens since they show pathogen-induced differential gene expression and, more importantly, some of these genes are required for disease resistance against bacterial infection as demonstrated by RNAi studies (Pees et al. 2015). The high diversity of CTLDC containing proteins in bivalves and gastropods is well known (Dheilly et al. 2015; Gerdol & Venier 2015; Song et al. 2015). In mussels, 154 C-type lectin protein (containing a single CTL domain) transcripts have been annotated in a digestive gland transcriptomes. There are 350 CTLDC genes annotated in the *C. gigas* genome (Gerdol & Venier, 2015; Zhang et al. 2012), although only roughly half that number were found here in the transcriptomes of *C. gigas* and *C. virginica*, suggesting patterns of specific gene expression stimulated by experimental conditions not included in the oyster transcriptomes studies performed so far. Due to the diversity of functions in of CTLDC proteins, further research will be needed to determine which members of the CTLDC family contribute to recognition and/or elimination of pathogens in bivalves.

Diversification in Fib_C DC or FREPs has been described in the gastropod *B. glabrata* and other molluscan species, including bivalves (Gerdol & Venier 2015; Gordy et al. 2015; Hanington and Zhang 2011, Guo et al. 2015). Fibrinogen-related proteins are of particular interest because they can diversify upon pathogen exposure and create unique populations of hemocytes (Hanington et al. 2010). There are indications from gastropods of diversification through somatic diversification and hypermutation (Zhang et al. 2004). Here, many fewer FREP transcripts were found in the *C. virginica* transcriptome than are annotated in the *C. gigas* genome, and the vast majority of differentially expressed FREPs in the eastern oyster were down-regulated upon bacterial challenge. We hypothesize that either FREPs are not important in the response of *C. virginica* to *R. crassostreae* or that FREPs are expressed at time points other than the ones included in this analysis (i.e. less than 24-hr after challenge).

Although not to the level of the ones discussed above, other immune-related gene families showing increased diversity of transcripts in *C. virginica* include apextrin, serine protease inhibitors, and dermatopontin. A similar diversity of apextrin and serine protease transcripts has been observed in the mussel *M. galloprovincialis* (Gerdol & Venier 2015; Zhang et al. 2016). Apextrin, a complement-related protein containing a membrane-attack complex/perforin domain (Miller et al. 2007), and a CELIII type lectin have been identified as putative immune-related genes in the coral *A. millepora* (Puill-Stephan et al. 2012). Interestingly, apextrin is another immune-related gene family that shows a patchy phylogenetic distribution (Dheilly et al. 2011; Miller et al. 2007). Experimental evidence supporting the role of dermatopontin in immune responses in bivalves is available (Martin-Gomez et al. 2014; Wang et al. 2013).

In conclusion, a multi-species similarity cluster analysis focus on transcripts identified in eastern oysters in response to bacterial challenge has allowed for the identification of both known and novel diversified gene families with potential roles in immunity in oysters. A limitation of this study is the paucity of genomes for non-model invertebrate species spanning a diversity of phyla, since identification of expanded families based solely on transcripts and the mapping of transcripts to genomes of related species may lead to inflated estimates of diversity. Furthermore, lack of reference genomes hinders the determination of underlying mechanisms of diversification. The constantly growing wealth of sequence information on non-model species that span the phylogenetic diversity of invertebrate phyla should confirm the identification of

gene families reported here and other studies and uncover additional independent gene family expansions. The elucidation of mechanisms of diversification in specific gene families requires: 1) availability of well-assembled reference genomes allowing comparative genomics studies between related species with divergent histories of pathogen exposures; 2) access to a thorough catalog of transcriptomes from a comprehensive set of different developmental stages, tissue and cell types exposed to targeted and relevant immune challenges; and 3) a focus on diversity of responses that span from the level of individual cells (such as hemocytes) and organisms to the level of populations. The exponential decrease in sequencing costs allowing for extensive resequencing projects will likely pave the way for the discovery of novel mechanisms of immune diversification. Mechanisms of diversification that require further study include gene editing, which has shown to contribute to diversity in genes involved in specialized functions in cephalopods (e.g. Albertin et al. 2015) and somatic rearrangements, which may be responsible for FREP diversity as shown in gastropods (e.g. Zhang et al. 2004). In particular, sequencing of mRNA repertoires in individual hemocytes would allow determining the role of these mechanisms on immune gene diversification and immune responses.

6. Acknowledgements

We are grateful to all Gomez-Chiarri lab members who have helped contribute to this project including J. Piesz, M. Karim, and S. B. Sohn. We are grateful to S. Halary (Université Pierre et Marie Curie) and I. Misner (University of Rhode Island) for their helpful clustering pipeline script. We also thank the staff at Brown University's Center for Computing and Visualization, especially L. Dong and M. Howison for technical support and D. Proestou (USDA ARS), M. Hare (Cornell U), and 2 anonymous reviewers for helpful suggestions on the manuscript. This research has been funded by USDA CSREES Award 2009-38925-19971, USDA award RI00AH-884 from the Rhode Island Agricultural Experimental Station and Specific Cooperative Agreement 5358-31000-003-02S between USDA ARS and University of Rhode Island. This research was possible thanks to the Rhode Island Genomics and Sequencing Center, supported in part by the National Science Foundation under EPSCoR Grants No. 0554548 and EPS-1004057.

7. References

Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21, 2104–2105. doi:10.1093/bioinformatics/bti263

476 Aberer, A.J., Krompass, D., Stamatakis, A., 2013. Pruning rogue taxa improves phylogenetic
477 accuracy: an efficient algorithm and webservice. *Syst. Biol.* 62, 162–166.
478 doi:10.1093/sysbio/sys078

479 Albertin, C.B., Simakov, O., Mitros, T., Wang, Z.Y., Pungor, J.R., Edsinger-Gonzales, E.,
480 Brenner, S., Ragsdale, C.W., Rokhsar, D.S., 2015. The octopus genome and the evolution of
481 cephalopod neural and morphological novelties. *Nature* 524, 220–224. doi:10.1038/nature14668

482 Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.,
483 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.
484 *Nucleic acids research* 25, 3389–3402.

485 Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) *The New S Language*. Wadsworth &
486 Brooks/Cole.

487 Birney, E., Clamp, M., Durbin, R., 2004. GeneWise and Genomewise. *Genome Res.* 14, 988–
488 995. doi:10.1101/gr.1865504

489 Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful
490 approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*
491 289–300.
492

493 Boettcher, K.J., Geaghan, K.K., Maloy, A.P., Barber, B.J., 2005. *Roseovarius crassostreae* sp.
494 nov., a member of the Roseobacter clade and the apparent cause of juvenile oyster disease (JOD)
495 in cultured Eastern oysters. *International journal of systematic and evolutionary microbiology*
496 55, 1531–1537.

497 Brites, D., Brena, C., Ebert, D., Du Pasquier, L., 2013. More Than One Way to Produce Protein
498 Diversity: Duplication and Limited Alternative Splicing of an Adhesion Molecule Gene in Basal
499 Arthropods. *Evolution* 67, 2999–3011. doi:10.1111/evo.12179

500 Ciucci, T., Bosselut, R., 2014. Gimap and T cells: A matter of life or death. *Eur. J. Immunol.* 44,
501 348–351. doi:10.1002/eji.201344375

502 Dheilly, N.M., Duval, D., Mouahid, G., Emans, R., Allienne, J.-F., Galinier, R., Genthon, C.,
503 Dubois, E., Du Pasquier, L., Adema, C.M., Grunau, C., Mitta, G., Gourbal, B., 2015. A family of
504 variable immunoglobulin and lectin domain containing molecules in the snail *Biomphalaria*
505 *glabrata*. *Developmental & Comparative Immunology* 48, 234–243.
506 doi:10.1016/j.dci.2014.10.009

507 De Lorgeril, J., Zenagui, R., Rosa, R.D., Piquemal, D., Bachère, E., 2011. Whole Transcriptome
508 Profiling of Successful Immune Response to *Vibrio* Infections in the Oyster *Crassostrea gigas* by
509 Digital Gene Expression Analysis. *PLoS ONE* 6, e23142. doi:10.1371/journal.pone.0023142

510 Dheilly, N.M., Haynes, P.A., Bove, U., Nair, S.V., Raftos, D.A., 2011. Comparative proteomic
511 analysis of a sea urchin (*Heliocidaris erythrogramma*) antibacterial response revealed the
512 involvement of apextrin and calreticulin. *Journal of Invertebrate Pathology* 106, 223–229.
513 doi:10.1016/j.jip.2010.09.008

Dishaw, L.J., Giacomelli, S., Melillo, D., Zucchetti, I., Haire, R.N., Natale, L., Russo, N.A., De
 Santis, R., Litman, G.W., Pinto, M.R., 2011. A role for variable region-containing chitin-binding
 proteins (VCBPs) in host gut–bacteria interactions. *Proc Natl Acad Sci U S A* 108, 16747–
 16752. doi:10.1073/pnas.1109687108

Enright, A.J., Van Dongen, S., Ouzounis, C.A., 2002. An efficient algorithm for large-scale
 detection of protein families. *Nucleic Acids Res* 30, 1575–1584.

Filen, S., Lahesmaa, R., 2010. GIMAP Proteins in T-Lymphocytes. *Journal of Signal
 Transduction* 2010, e268589. doi:10.1155/2010/268589

Fischer, S., Brunk, B.P., Chen, F., Gao, X., Harb, O.S., Iodice, J.B., Shanmugam, D., Roos, D.S.,
 Stoeckert, C.J., 2011. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster
 proteomes into new ortholog groups. *Curr Protoc Bioinformatics* Chapter 6, Unit 6.12.1–19.
 doi:10.1002/0471250953.bi0612s35

Frech, C., Chen, N., 2010. Genome-Wide Comparative Gene Family Classification. *PLoS One* 5.
 doi:10.1371/journal.pone.0013409

Gerdol, M., Manfrin, C., De Moro, G., Figueras, A., Novoa, B., Venier, P., Pallavicini, A., 2011.
 The C1q domain containing proteins of the Mediterranean mussel *Mytilus galloprovincialis*: A
 widespread and diverse family of immune-related molecules. *Developmental & Comparative
 Immunology* 35, 635–643. doi:10.1016/j.dci.2011.01.018

Gerdol, M., Venier, P., 2015. An updated molecular basis for mussel immunity. *Fish & Shellfish
 Immunology*. doi:10.1016/j.fsi.2015.02.013

Gerdol, M., Venier, P., Pallavicini, A., 2015. The genome of the Pacific oyster *Crassostrea gigas*
 brings new insights on the massive expansion of the C1q gene family in Bivalvia.
Developmental & Comparative Immunology 49, 59–71. doi:10.1016/j.dci.2014.11.007

Ghosh, J., Lun, C.M., Majeske, A.J., Sacchi, S., Schrankel, C.S., Smith, L.C., 2011. Invertebrate
 immune diversity. *Dev. Comp. Immunol.* 35, 959–974. doi:10.1016/j.dci.2010.12.009

Gómez-Chiarri, M., Warren, W.C., Guo, X., Proestou, D., 2015. Developing tools for the study
 of molluscan immunity: The sequencing of the genome of the eastern oyster,
Crassostrea virginica. *Fish & Shellfish Immunology, SI: Molluscan Immunity* 46, 2–4.
 doi:10.1016/j.fsi.2015.05.004

Gordy, M.A., Pila, E.A., Hanington, P.C., 2015. The role of fibrinogen-related proteins in the
 gastropod immune response. *Fish & Shellfish Immunology, SI: Molluscan Immunity* 46, 39–49.
 doi:10.1016/j.fsi.2015.03.005

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X.,
 Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N.,
 di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011.
 Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature
 Biotechnology* 29, 644–652.

551 Green, T.J., Montagnani, C., 2013. Poly I:C induces a protective antiviral immune response in
552 the Pacific oyster (*Crassostrea gigas*) against subsequent challenge with Ostreid herpesvirus
553 (OsHV-1 μ var). *Fish & Shellfish Immunology* 35, 382–388. doi:10.1016/j.fsi.2013.04.051

554 Hallen, L.C., Burki, Y., Ebeling, M., Broger, C., Siegrist, F., Oroszlan-Szovik, K., Bohrmann,
555 B., Certa, U., Foser, S., 2007. Antiproliferative activity of the human IFN- α -inducible
556 protein IFI44. *J. Interferon Cytokine Res.* 27, 675–680. doi:10.1089/jir.2007.0021

557 Hanington, P.C., Forys, M.A., Dragoo, J.W., Zhang, S.-M., Adema, C.M., Loker, E.S., 2010.
558 Role for a somatically diversified lectin in resistance of an invertebrate to parasite infection.
559 *PNAS* 107, 21087–21092. doi:10.1073/pnas.1011242107

560 Hanington, P.C., Zhang, S.-M., 2011. The Primary Role of Fibrinogen-Related Proteins in
561 Invertebrates Is Defense, Not Coagulation. *Journal of Innate Immunity* 3, 17–27.
562 doi:10.1159/000321882

563 He, Y., Jouaux, A., Ford, S.E., Lelong, C., Sourdain, P., Mathieu, M., Guo, X., 2015.
564 Transcriptome analysis reveals strong and complex antiviral response in a mollusc. *Fish*
565 *Shellfish Immunol.* 46, 131–144. doi:10.1016/j.fsi.2015.05.023

566 Heinonen, M.T., Laine, A.-P., Söderhäll, C., Gruzdeva, O., Rautio, S., Melén, E., Pershagen, G.,
567 Lähdesmäki, H.J., Knip, M., Ilonen, J., Henttinen, T.A., Kere, J., Lahesmaa, R., Registry,
568 T.F.P.D., Knip, M., Groop, P.-H., Ilonen, J., Lappi, A., Otonkoski, T., Saha, M.-T., Simell, O.,
569 Talvitie, T., Vaarala, O., Veijola, R., Aito, H., Bondestam, J., Dahllund, T., Granvik, J.,
570 Haapalehto-Ikonen, M., Hämäläinen, A.-M., Huopio, H., Johansson, C., Kinnala, A., Korhonen,
571 J., Korpela, P., Korteniemi, M., Lautala, P., Lindström, K., Miettinen, P., Mustila, T., Nuuja, A.,
572 Nykänen, P., Ojanperä, J., Putto-Laurila, A., Saha, M.-T., Sankila, J., Suomi, A.-M., Tenhola, S.,
573 Varimo, P., Veijola, R., Virransalo, R., Vuolukka, P., Ylitalo, S., 2015. GIMAP GTPase Family
574 Genes: Potential Modifiers in Autoimmune Diabetes, Asthma, and Allergy. *J Immunol* 194,
575 5885–5894. doi:10.4049/jimmunol.1500016

576 Hibino, T., Loza-Coll, M., Messier, C., Majeske, A.J., Cohen, A.H., Terwilliger, D.P., Buckley,
577 K.M., Brockton, V., Nair, S.V., Berney, K., Fugmann, S.D., Anderson, M.K., Pancer, Z.,
578 Cameron, R.A., Smith, L.C., Rast, J.P., 2006. The immune gene repertoire encoded in the purple
579 sea urchin genome. *Developmental Biology* 300, 349–365.

580 Huang, B., Zhang, L., Li, L., Tang, X., Zhang, G., 2015. Highly diverse fibrinogen-related
581 proteins in the Pacific oyster *Crassostrea gigas*. *Fish & Shellfish Immunology* 43, 485–490.
582 doi:10.1016/j.fsi.2015.01.021

583 Katoh, K., Asimenos, G., Toh, H., 2009. Multiple alignment of DNA sequences with MAFFT.
584 *Methods Mol. Biol.* 537, 39–64. doi:10.1007/978-1-59745-251-9_3

585 Kitamura, A., Takahashi, K., Okajima, A., Kitamura, N., 1994. Induction of the human gene for
586 p44, a hepatitis-C-associated microtubular aggregate protein, by interferon- α /beta. *Eur. J.*
587 *Biochem.* 224, 877–883.

588 Lacoste, A., Jalabert, F., Malham, S.K., Cueff, A., Poulet, S.A., 2001. Stress and stress-induced
589 neuroendocrine changes increase the susceptibility of juvenile oysters (*Crassostrea gigas*) to
590 *Vibrio splendidus*. *Applied and environmental microbiology* 67, 2304–2309.

591 Le, S.Q., Gascuel, O., 2008. An improved general amino acid replacement matrix. *Mol. Biol.*
592 *Evol.* 25, 1307–1320. doi:10.1093/molbev/msn067

593 Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of
594 protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659.
595 doi:10.1093/bioinformatics/btl158

596 Liberti, A., Leigh, B., Santis, R.D., Pinto, M.R., Cannon, J.P., Dishaw, L.J., Litman, G.W., 2015.
597 An Immune Effector System in the Protochordate Gut Sheds Light on Fundamental Aspects of
598 Vertebrate Immunity, in: Hsu, E., Pasquier, L.D. (Eds.), *Pathogen-Host Interactions: Antigenic*
599 *Variation v. Somatic Adaptations, Results and Problems in Cell Differentiation*. Springer
600 International Publishing, pp. 159–173.

601 Loker, E.S., 2012. Macroevoolutionary Immunology: A Role for Immunity in the Diversification
602 of Animal life. *Front Immunol* 3. doi:10.3389/fimmu.2012.00025

603 Martín-Gómez, L., Villalba, A., Carballal, M.J., Abollo, E., 2014. Molecular characterisation of
604 TNF, AIF, dermatopontin and VAMP genes of the flat oyster *Ostrea edulis* and analysis of their
605 modulation by diseases. *Gene* 533, 208–217. doi:10.1016/j.gene.2013.09.085

606 McDowell, I.C., Nikapitiya, C., Aguiar, D., Lane, C.E., Istrail, S., Gomez-Chiarri, M., 2014.
607 Transcriptome of American Oysters, *Crassostrea virginica*, in Response to Bacterial Challenge:
608 Insights into Potential Mechanisms of Disease Resistance. *PLoS ONE* 9, e105097.
609 doi:10.1371/journal.pone.0105097

610 Messier-Solek, C., Buckley, K.M., Rast, J.P., 2010. Highly diversified innate receptor systems
611 and new forms of animal immunity. *Seminars in Immunology* 22, 39–47.

612 Miller, D.J., Hemmrich, G., Ball, E.E., Hayward, D.C., Khalturin, K., Funayama, N., Agata, K.,
613 Bosch, T.C., 2007. The innate immune repertoire in Cnidaria - ancestral complexity and
614 stochastic gene loss. *Genome Biol* 8, R59. doi:10.1186/gb-2007-8-4-r59

615 Ottaviani, E., Franceschi, C., 1996. The neuroimmunology of stress from invertebrates to man.
616 *Progress in Neurobiology* 48, 421–440. doi:10.1016/0301-0082(95)00049-6

617 Pancer, Z., 2000. Dynamic expression of multiple scavenger receptor cysteine-rich genes in
618 coelomocytes of the purple sea urchin. *Proceedings of the National Academy of Sciences* 97,
619 13156–13161.

620 Pees, B., Yang, W., Zarate-Potes, A., Schulenburg, H., Dierking, K., 2015. High Innate Immune
621 Specificity through Diversified C-Type Lectin-Like Domain Proteins in Invertebrates. *Journal of*
622 *Innate Immunity*. doi:10.1159/000441475

623 Puill-Stephan, E., Seneca, F.O., Miller, D.J., van Oppen, M.J.H., Willis, B.L., 2012. Expression
624 of Putative Immune Response Genes during Early Ontogeny in the Coral *Acropora millepora*.
625 *PLoS One* 7. doi:10.1371/journal.pone.0039099

Renault, T., Faury, N., Barbosa-Solomieu, V., Moreau, K., 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. *Dev. Comp. Immunol.* 35, 725–735. doi:10.1016/j.dci.2011.02.004

Reuber, T.L., Ausubel, F.M., 1996. Isolation of Arabidopsis genes that differentiate between resistance responses mediated by the RPS2 and RPM1 disease resistance genes. *Plant Cell* 8, 241–249. doi:10.1105/tpc.8.2.241

Robinson, M.D., Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 11, R25.

Roth, M.O., Wilkins, A.G., Cooke, G.M., Raftos, D.A., Nair, S.V., 2014. Characterization of the Highly Variable Immune Response Gene Family, He185/333, in the Sea Urchin, *Heliocidaris erythrogramma*. *PLoS One* 9. doi:10.1371/journal.pone.0062079

Sanders, V.M., Straub, R.H., 2002. Norepinephrine, the β -Adrenergic Receptor, and Immunity. *Brain, Behavior, and Immunity* 16, 290–332. doi:10.1006/brbi.2001.0639

Song, L., Wang, L., Zhang, H., Wang, M., n.d. The immune system and its modulation mechanism in scallop. *Fish & Shellfish Immunology*. doi:10.1016/j.fsi.2015.03.013

Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–2690. doi:10.1093/bioinformatics/btl446

Walenz, B., Florea, L., 2011. Sim4db and Leaf: utilities for fast batch spliced alignment and sequence indexing. *Bioinformatics* 27, 1869–1870. doi:10.1093/bioinformatics/btr285

Wang, G., Zhang, Z., Lin, S., Zhang, L., Wang, B., Wang, S., Wang, Y., 2013. A homologue of dermatopontin from *Haliotis diversicolor* and its response to pathogenic infection. *Aquaculture Research* n/a–n/a. doi:10.1111/are.12305

Wang, Z., Li, X., 2009. IAN/GIMAPs are conserved and novel regulators in vertebrates and angiosperm plants. *Plant Signaling & Behavior* 4, 165–167. doi:10.4161/psb.4.3.7722

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., Barton, G.J., 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189–1191. doi:10.1093/bioinformatics/btp033

Webb, L.M.C., Datta, P., Bell, S.E., Kitamura, D., Turner, M., Butcher, G.W., 2016. GIMAP1 Is Essential for the Survival of Naive and Activated B Cells In Vivo. *J Immunol* 196, 207–216. doi:10.4049/jimmunol.1501582

Weiss, Y., Forêt, S., Hayward, D.C., Ainsworth, T., King, R., Ball, E.E., Miller, D.J., 2013. The acute transcriptional response of the coral *Acropora millepora* to immune challenge: expression of GiMAP/IAN genes links the innate immune responses of corals with those of mammals and plants. *BMC Genomics* 14, 400. doi:10.1186/1471-2164-14-400

662 Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., Yang, P., Zhang, L., Wang, X., Qi, H.,
663 2012. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*
664 490, 49–54.

665 Zhang, L., Li, L., Guo, X., Litman, G.W., Dishaw, L.J., Zhang, G., 2015. Massive expansion and
666 functional divergence of innate immune genes in a protostome. *Scientific Reports* 5, 8693.
667 doi:10.1038/srep08693

668 Zhang, L., Li, L., Zhu, Y., Zhang, G., Guo, X., 2014. Transcriptome Analysis Reveals a Rich
669 Gene Set Related to Innate Immunity in the Eastern Oyster (*Crassostrea virginica*). *Mar*
670 *Biotechnol* 16, 17–33. doi:10.1007/s10126-013-9526-z

671 Zhang, S.-M., 2004. Diversification of Ig Superfamily Genes in an Invertebrate. *Science* 305,
672 251–254. doi:10.1126/science.1088069

673 Zhang, S.-M., Loker, E.S., Sullivan, J.T., 2016. Pathogen-associated molecular patterns activate
674 expression of genes involved in cell proliferation, immunity and detoxification in the amebocyte-
675 producing organ of the snail *Biomphalaria glabrata*. *Developmental & Comparative Immunology*
676 56, 25–36. doi:10.1016/j.dci.2015.11.008

677 Zhou, Z., Wang, L., Shi, X., Yue, F., Wang, M., Zhang, H., Song, L., 2012. The expression of
678 dopa decarboxylase and dopamine beta hydroxylase and their responding to bacterial challenge
679 during the ontogenesis of scallop *Chlamys farreri*. *Fish & Shellfish Immunology* 33, 67–74.
680 doi:10.1016/j.fsi.2012.04.002

681 Zhou, Z., Wang, L., Yang, J., Zhang, H., Kong, P., Wang, M., Qiu, L., Song, L., 2011. A
682 dopamine beta hydroxylase from *Chlamys farreri* and its induced mRNA expression in the
683 haemocytes after LPS stimulation. *Fish & Shellfish Immunology* 30, 154–162.
684 doi:10.1016/j.fsi.2010.09.020

685 **Supplemental Information**

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

688 Supporting Table S2. Gene family annotation reference sequences search terms.

689 Supporting Table S3. Relevant InterProScan entries for selected gene families of interest.

690 Supporting Table S4. Number of gene models for each gene family.

691 Supplemental Figure S1. Alignment of IFI44 sequences from selected organisms.

692 Supplemental Figure S2. Alignment of GIMAP sequences from selected organisms.