Integrative and Comparative Biology

Integrative and Comparative Biology, volume 56, number 6, pp. 1092–1102
doi:10.1093/icb/icw043

Society for Integrative and Comparative Biology

SYMPOSIUM

Construction and Characterization of Two Novel Transcriptome Assemblies in the Congeneric Porcelain Crabs *Petrolisthes cinctipes* and *P. manimaculis*

Eric J. Armstrong[1,*,†] and Jonathon H. Stillman[*,†]

*Department of Integrative Biology, University of California, 3040 Valley Life Sciences Bldg, Berkeley, CA 94720, USA; †Romberg Tiburon Center for Environmental Studies, San Francisco State University, 3152 Paradise Drive, Tiburon, CA 94920, USA

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

[1]E-mail: armstrong@berkeley.edu

Synopsis Crustaceans have commonly been used as non-model systems in basic biological research, especially physiological regulation. With the recent and rapid adoption of functional genomic tools, crustaceans are increasingly becoming model systems for ecological investigations of development and evolution and for mechanistic examinations of genotype–phenotype interactions and molecular pathways of response to environmental stressors. Comparative transcriptomic approaches, however, remain constrained by a lack of sequence data in closely related crustacean taxa. We identify challenges in the use of functional genomics tools in comparative analysis among decapod crustacean in light of recent advances. We present RNA-seq data from two congeneric species of porcelain crabs (*Petrolisthes cinctipes* and *P. manimaculis*) used to construct two *de novo* transcriptome assemblies with ~194K and ~278K contigs, respectively. We characterize and contrast these assemblies and compare them to a previously generated EST sequence library for *P. cinctipes*. We also discuss the potential use of these data as a case-study system in the broader context of crustacean comparative transcriptomics.

Introduction

Crustaceans as emerging model systems in functional genomics

Functional genomics investigative approaches are rapidly transforming the field of comparative physiology, allowing investigators unparalleled ability to measure genotypic features and phenotypic responses in great detail (Feder and Mitchell-Olds 2003; Stillman and Armstrong 2015). In recognition of the potential of next-generation sequencing (NGS) enabled research, the National Science Foundation's Animal Genomes to Phenomes Research Coordination Network was established with the goal of unifying and improving NGS resources from non-model animals in order to address key issues in animal husbandry, disease, and mechanistic, evolutionary and ecological physiology including questions on adaptive evolution and non-linear or non-additive (e.g., synergistic, antagonistic) phenotypic responses to stress. Although the diverse range of organisms and stress exposures currently being investigated using NGS approaches has provided fruitful grounds for comparative research in non-model systems (see Stillman and Armstrong (2015) for a recent review), our ability to detect adaptive shifts and to address mechanistic questions regarding phenotypic plasticity in non-model organisms, including nearly all crustaceans, is ultimately limited by the availability of comparative sequence data from sufficiently closely related taxa.

As a group, crustaceans are highly valuable commercially and have been used for well over a century as ideal study subjects in the field of comparative embryology (Dana 1852). In recent years, crustaceans

Advanced Access publication July 3, 2016

have also begun to emerge as model systems in the fields of evolutionary developmental biology and ecological physiology for investigating mechanisms of evolution and adaptation (Zeng et al. 2011). Despite this growing interest in crustaceans as research systems, genomic resources for crustacean taxa are exceedingly few, with publicly available sequenced genome data existing for seven species, five of which were added to NCBI's GenBank within the past two years. Perhaps the most widely referenced genome among these repositories is that of the extensively studied water flea, *Daphnia pulex*. While ecogenomic studies of *D. pulex* have revealed valuable information regarding genome–phenome linkages and genomic signatures of selection (Miner et al. 2012), it remains to be seen how representative the *D. pulex* genome is of other crustacean lineages and thus how generalizable these results are across a group as functionally and ecologically diverse as the Crustacea. Because of this, there is a strong need to develop databases of genomic sequence in other ecologically and commercially important species, especially within the Decapoda, for comparative analyses.

Recent advances and challenges in NGS research using non-model crustacean systems

Although several EST projects have been conducted in crustacean systems, including the economically valuable shrimp *Litopenaeus vannamei* (~163K ESTs; Vega et al. 2008; Alcivar-Warren et al. 2009; Muller et al. 2010) and lobster *Homarus americanus* (~52K ESTs; Stepanyan et al. 2006; Verslycke et al. 2009), and the parasitic copepod *Lepeophtheirus salmonis* (~129K ESTs; Nilsen et al. 2010), emerging NGS technologies such as RNA- and RAD-seq have only recently begun to be used in earnest. While a sizeable microarray EST dataset does already exist for the porcelain crab *Petrolisthes cinctipes* (~98K ESTs; Tagmount et al. 2010), the sequences generated in this study via NGS approaches nearly triple this dataset (~194K ESTs) and adds ~279K sequences for the closely related species *Petrolisthes manimaculis*. Within just the last few years, the community of crustacean biologists have utilized NGS-powered functional genomics approaches to examine, among other topics, growth and development (lipid storage and diapause cues in *Calanus finmarchicus* copepods; Tarrant et al. (2014); comparison of eye-development among ecotypes in the freshwater isopod, *Asellus aquaticus*; Stahl et al. (2015)), immune and toxin responses (investigation of resistance to a delousing drug in *Caligus rogercresseyi*, Chávez-Mardones and Gallardo-Escárate (2015); identification of candidate genes involved in virus-induced immune response in *L. vannamei*; Robalino et al. (2007); elucidation of differences between sexes and populations in drug resistance in the parasitic copepod *L. salmonis*; Poley et al. (2015), regulation of molting (Y-organ molt gland profiling, Das et al. (2016); global expression changes throughout the molt cycle in *L. vannamei*, Gao et al. (2015), and responses to hypercapnia and hypoxia (characterization of a suite of hemocyanin genes involved in hypercapnic-hypoxia response in *L. vannamei*; Johnson et al. (2015), nitrate stress (contrasting acute and chronic exposure in the river prawn *Macrobrachium nipponense*; Xu et al. (2016), and thermal and osmotic challenges (acute and plastic heat shock responses in *P. cinctipes*; Teranishi and Stillman (2007); Stillman and Tagmount (2009); Ronges et al. (2012); responses to acute salinity stress in *L. vannamei*, Wang et al. (2015). Such approaches are useful in helping to develop mechanistic models of cellular processes accompanying phenotypic shifts in response to stress (phenotypic plasticity) and provide fruitful grounds for addressing hypotheses in the field of ecological physiology. An understanding of the evolutionary implications of phenotypic responses however requires that we extend this comparative transcriptomic framework to include larger-scale (e.g., between populations or ecotypes) and interspecific comparisons to elucidate shared and unique responses to environmental stressors.

Crustaceans are an ideal group in which to further develop such interspecific comparative transcriptomic approaches because they are functionally diverse and have long been used as interesting model systems for understanding animal evolution and physiology. Because of these qualities, there has been a surge in the number of genomics-based investigations utilizing crustaceans study systems, several of which have attempt to integrate across multiple levels of organization from the ecological to the mechanistic in an effort to develop general principles of stress response in arthropods (Stillman et al. 2008). Recognizing this potential for advancing the field of comparative physiology, there is currently a strong push amongst the community of crustacean biologists to generate additional NGS genomic sequence resources. Here, we present the results of our efforts to generate two novel porcelain crab sequence databases from congeneric species for analysis within this larger comparative framework.

Porcelain crabs as a comparative case-study system

Porcelain crabs in the genus *Petrolishthes* (Decapoda: Anomura: Porcellanidae) are a highly diverse group of

crustaceans which inhabit intertidal zone and shallow benthic habitats worldwide (Haig 1960). This group comprises over 100 species with highest diversity in the Pacific and spans a wide latitudinal range from temperate to the tropical habitats along both Eastern and Western Pacific shorelines (Stillman and Somero 1996; Stillman 2002, 2004). Within a given region, individual *Petrolisthes* species segregate into distinct vertical zones from subtidal to upper intertidal zone habitats. This vertical zonation is tied to differences in thermal performance between species, and suggests that porcelain crabs are well adapted the thermal environments they inhabit (Stillman and Reeb 2001). In the temperate Eastern Pacific, this vertical zonation pattern is observed between the species *P. cinctipes* and *P. manimaculis* with the more heat-tolerant *P. cinctipes* dominating in the upper intertidal zone and the heat-sensitive *P. manimaculis* restricted to the lower intertidal and subtidal zones (Stillman and Reeb 2001). Because of their demonstrated differences in thermal sensitivity, *P. cinctipes* and *P. manimaculis* represent an exciting model system for gene expression profiling.

Here, we present two novel gene expression profiles generated via Illumina NGS technologies for the congeneric porcelain crabs *P. cinctipes* and *P. manimaculis*. These profiles build on the wealth of phylogenetic and comparative thermal physiology data presently available for these species (Stillman and Somero 1996, 2000; Stillman 2004; Carter et al. 2013; Ceballos-Osuna et al. 2013; Paganini et al. 2014) and add to the repository of sequence data generated previously in *P. cinctipes* under other stress conditions (Stillman et al. 2006; Teranishi and Stillman 2007; Tagmount et al. 2010). Further development of this genomic resource in these closely related congeners will prove especially useful for comparative functional genomic analyses of mechanisms of thermal adaptation (i.e., adaptive evolution) and of patterns of response (i.e., phenotypic plasticity) to thermal and acidification stress within the Crustacea.

Methods

Collection of specimens and construction of cDNA libraries

Specimen collection and experimental stress treatments

Our cDNA libraries were constructed from specimens of *P. cinctipes* ($n = 22$) and *P. manimaculis* ($n = 24$) collected from beneath rubble in the upper and lower intertidal zones in Pacifica, California (38.5143°N, 123.2438°W) during low tide in September 2013. Crabs were held under common conditions in the lab for a 10-day acclimation period prior to the experiment before being placed in experimental aquaria and exposed to two levels of pH and temperature variability: low-variability control conditions of constant temperature (11 ± 0.5 °C) and pH (8.1 ± 0.1) or high-variability conditions where specimens experienced a daily temperature spike to 28 °C during low tide emersion (ramp rate of 3.4 °C h^{-1} over 5 h, 11:00–16:00) and a daily pH drop to 7.55 ± 0.1 during high tide (2:00–9:00) as described previously (Paganini et al. 2014). Tissue samples for RNA-seq analysis were collected on Days 1 and 14 of the experiment following the pH-drop (9:00) and were immediately processed as described below. Samples were sequenced individually, but all generated sequences were combined in the final assembly regardless of experimental treatment condition.

Total RNA extraction and cDNA library preparation

Total RNA was extracted by homogenizing vivisected gill tissue (7.9 ± 3.0 mg average wet mass) in 0.5 mL Trizol at room temperature using a QIAgen TissueLyser (QIAgen, Venio, Netherlands). Homogenates were then mixed with 100 μL chloroform by vortex for 30 s, and incubated at room temperature for 2 min. Samples were then centrifuged for 15 min at 12,000 *g* and 4 °C. The clear, aqueous top layer of each sample was carefully removed to fresh microcentrifuge tubes, mixed with 0.5 mL chilled isopropanol, and incubated at −20 °C overnight. Precipitated RNA was pelleted by centrifugation at 12,000 *g* for 10 min at 4 °C. The supernatant was then discarded and pellets were washed twice with 1 mL 75% EtOH and centrifuged at 12,000 *g* for 5 min at 4 °C. Pellets were resuspended in 1 mM Na citrate, pH 6.5. Total RNA concentration and purity were determined spectrophotometrically using a Nanodrop®, Invitrogen Qubit® fluorometer and BioAnalyzer (Agilent Technologies, Santa Clara, CA). In general, A_{260}/A_{280} ratios were ≥ 1.84, with an average concentration of 309 ± 192 ng RNA/μL. Each individual RNA extract ($n = 5$–6 per treatment; 46 total RNA-seq samples) was used as a separate sample for cDNA library construction (i.e., samples were not pooled). We constructed all cDNA libraries using Illumina Tru-Seq® RNA sample preparation kits (set A: Fc 122–1001; set B: Fc 12–1002). Library construction was conducted following the manufacturers protocol, as in Benner et al. (2013).

Sequencing, assembly, and annotation of the *Petrolisthes* transcriptomes

All cDNA libraries were paired-end (PE) sequenced (100 bp reads) in the Vincent J. Coates Functional Genomics Laboratory at UC Berkeley on an Illumina HiSeq 2000. Two lanes of sequencing were

Table 1 Summary statistics of *Petrolisthes* and reference transcriptome assemblies

	P. cinctipes	*P. manimaculis*	*L. vannamei*[a]
Raw reads	4.25×10^8	4.82×10^8	1.4×10^8
Contigs	194,105	278,989	52,190
Mean length (bp)	1435	1065	870
N50 (bp)	2964	1980	1680
RMBT (%)	85.51%	82.43%	96.03%
GC content (%)	42.83%	41.30%	–

Note: [a]Data from (Johnson et al. 2015).

performed with 23 samples multiplexed per lane. We sequenced a total of 425M reads (Table 1), yielding sequencing depths of ~17.7M reads/sample and 21.9M reads/sample (post-quality control described below) for *P. cinctipes* and *P. manimaculis*, respectively.

All 100 bp PE raw sequences generated were analyzed using scripts modified from those created by Dr. Scott Fay available on GitHub (https://github.com/safay/RNA_seq/tree/master/blacklight_pipeline) and bioinformatic pipeline processes were conducted on the Pittsburg Supercomputer Center "Blacklight" as follows. First, library sequences were trimmed to remove Illumina adaptors (stringency = 1) and bases with a Phred quality score under 20 using Trim_Galore! (V 0.3.0; Krueger et al. 2012). For both assemblies, more than 95% of the raw reads passed these quality control metrics and were therefore used in the following *de novo* assembly protocol. FLASH (V 1.2.8) was used to construct longer reads for *de novo* assembly by joining of overlapping paired end reads and to avoid double counting of read overlap regions (Magoč and Salzberg 2011). Assembly of the *de novo* transcriptomes for each species was performed using Trinity (v2.1.0; Grabherr et al. 2011) with a minimum kmer coverage = 2. The resulting assemblies contained 194K and 279K contigs for *P. cinctipes* and *P. manimaculis* with average reads mapped back to transcriptome assembly values (RMBT%) of 85.5% and 82.4%, respectively (Table 1). The mean contig length was similar for both assemblies, although slightly higher in *P. cintipes* (1.4 versus 1.1 kb; Fig. 1). Contig N50 lengths differed significantly between assemblies, with *P. cinctipes* exhibiting a higher proportion of large contigs than *P. manimaculis* (N50 of 2.96 versus 1.98 kb, respectively). FASTA files of all assembly products are freely available from the authors upon request.

Transcriptome annotation was performed using Trinotate (v 2.0.1), and transcripts were grouped by gene ontology (GO) terms assigned from Blast2GO analysis (Conesa et al. 2005) of BLASTx homology search results against the Swiss-Prot UniProt database. All BLAST searches were conducted using BLAST v2.2.31 with a modified expect score cutoff threshold (expect < 2e−5). BLAST analyses were conducted on the Pittsburg Supercomputer Center "Greenfield". Comparison of assemblies was carried out using three, two-way reciprocal BLAST analyses (Fig. 2) between the three unique gene sets—*P. cinctipes* and *P. manimaculis* RNA-seq assemblies and the existing *P. cinctipes* cDNA microarray sequence database (Tagmount et al. 2010).

Results and Discussion

Analysis of sequencing depth of the transcriptomes

In order to assess the effects of our experimental design on our ability to detect differentially expressed genes in downstream analyses, we performed a power analysis using the online statistical software tool Scotty (Busby et al. 2013; http://scotty.genetics.utah.edu/). The Scotty Power Analysis software utilizes user-generated pilot data to calculate the predicted number of differentially-expressed genes that are likely to be detected when a transcript is represented by at least 10 reads as a function of sequencing depth and number of biological replicates in an RNA-seq experiment. The number of differentially-expressed genes detected is expected to increase with a greater number of replicates and a higher sequencing depth per replicate. In practice, these two parameters are often limited by the experimental budget, leading to tradeoffs in sequencing depth versus replication. ENCODE (Encyclopedia of DNA Elements) consortium best practices recommends sequencing depths of at least 30M reads/sample for analysis of differential gene expression (https://genome.ucsc.edu/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf). However, a number of recent studies have suggested that for typical differential expression analysis of RNA-seq data in non-model species, increasing sequencing depth per sample beyond a certain threshold yields diminishing returns for detection of truly differentially-expression genes (Liu et al. 2014, Harvid and Santos 2016). For such species however, at all sequencing depths, increasing the number of biological replicates consistently increases detection of differentially expressed genes. In light of this, we chose to utilize higher biological replication ($n = 5$ individuals per treatment) in this study at the expense of sequencing depth (~17.7M reads/sample and ~21.9M reads/sample as compared to

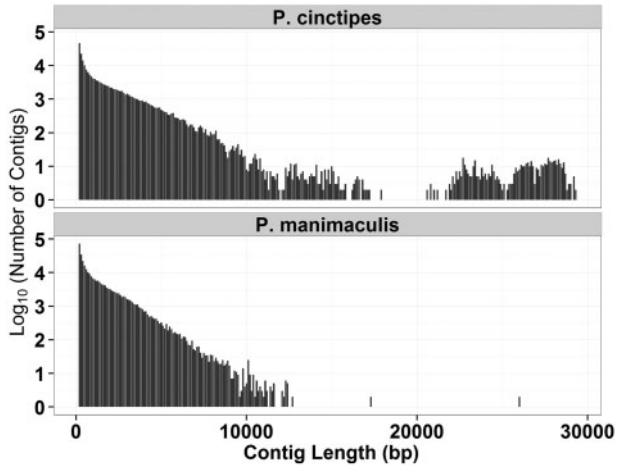

Fig. 1 Length distributions of all trimmed reads >200 bp in length from both *P. cinctipes* and *P. manimaculis* assemblies. The majority of assembled contigs for both species fell in the range of 200–800 bp (107453 sequences, 55.4% of total; 167936 sequences, 60.2% of total) with mean lengths of 1435 bp and 1065 bp for the *P. cinctipes* and *P. manimaculis* assemblies, respectively. For the *P. cinctipes* assembly, a small fraction of the assembled contigs (578 sequences, 0.29% of the total) fell in the range of > 20 kb thereby increasing the mean contig length and N50 values of this assembly relative to the *P. manimaculis* assembly.

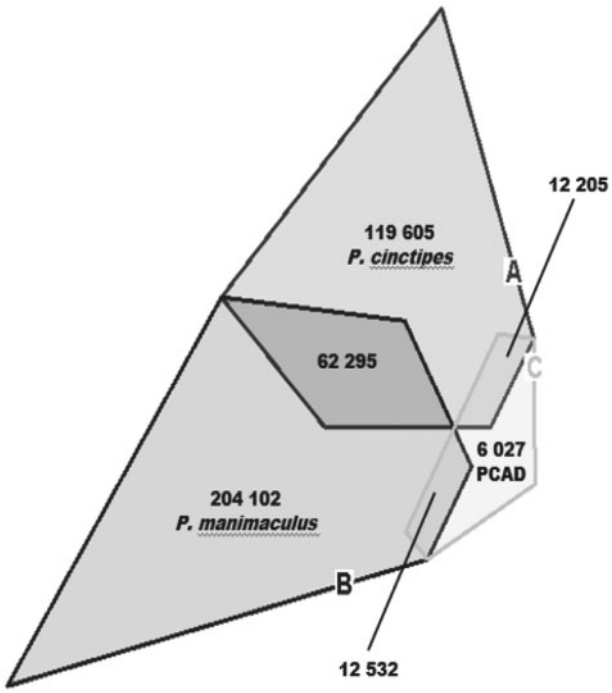

Fig. 2 Scaled Venn diagram (Euler 3 Java Applet software) displaying the results of three, two-way reciprocal BLASTx results between novel *P. cinctipes* (A) and *P. manimaculis* (B) RNA-seq generated transcriptomes and an existing *P. cinctipes* microarray sequence database (C). The area of each region is proportionally scaled to the number of contigs in that dataset and regions of overlap represent number of shared sequences between assemblies.

30M reads/sample ENCODE recommendation). For our assemblies, Scotty Power analysis suggests that, at sequencing depths < 150M total reads, we predict very little difference in gene detection rate between ENCODE recommendations and our experimental design (Fig. 3). Indeed, for our *P. cinctipes* assembly, power analysis predicts a slightly higher gene detection rate with the greater biological replication and shallower sequencing utilized in this study as compared to ENCODE best practice recommendations carried out on fewer replicates (Fig. 3, y-intercepts; ~133k versus ~114k genes, respectively).

To assess how deeply we had sequenced the transcripts in each of the species-specific cDNA pools, we utilized a progressive BLASTx search approach described previously for an RNA-seq assembly in the amphipod *Parhyale hawaiensis* (Zeng et al. 2011). Briefly, we created progressively larger BLAST query subassemblies from random subsets of the total number of reads generated for each species and assessed whether adding sequence data improved new gene discovery (Fig. 4). For the *P. manimaculis* assembly, the number of BLAST hits (performed against the Swiss-Prot database) consistently increased across all fractions of the assembly queried. This suggests that adding sequencing data for the same samples (i.e., increasing sequencing depth) would likely have led to even greater gene discovery in this species. For the *P. cinctipes* assembly, gene discovery rate shows a binary distribution pattern with low discovery rates at sampling percentages < 50% of the total and high, but relatively constant rates beyond that threshold. This suggests that for this species, gene discovery was saturated at sequencing depths > 50% of the total assembly. This pattern likely results from increased representation among the high length (> 20 kb), contigs present in assembly subsamples which contain > 50% of the total sequences assembled.

BLAST mapping of assembly sequences

The BLAST hit rate of the *P. cinctipes* and *P. manimaculis* assemblies were 31.1% and 22.9%, respectively (Table 2). Both of these hit ratios fell within the general range of values reported for other arthropod *de novo* transcriptome assemblies though there is considerable variation in hit rate reported among species (~10–50% depending on stringency criteria; Hahn et al. 2009; Roeding et al. 2009; Zeng et al. 2011; Johnson et al. 2015). The significantly higher BLAST hit rate for *P. cinctipes* is surprising given the close phylogenetic relationship of these two congeners. The comparatively low sequence

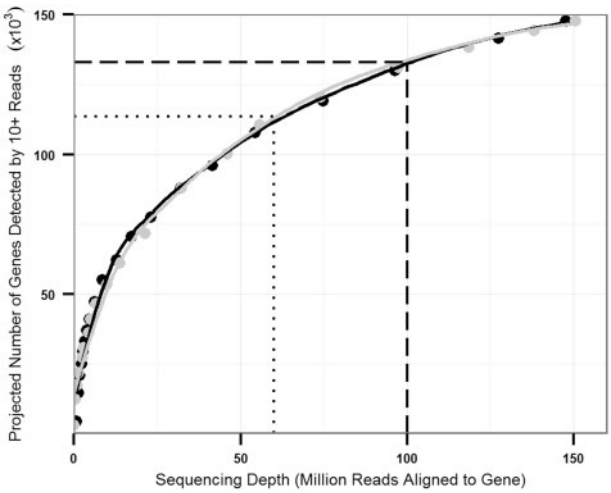

Fig. 3 Analysis of the effects of read depth and increased biological replication on gene detection rate. Depicted are the expected number of differentially-expressed genes detected by at least 10 reads as a function of increasing sequencing depth utilized in an RNA-seq analysis as calculated using the RNA-seq Power Analysis software tool Scotty. Solid grey and black lines represent Loess curves fit to mean predicted gene-detection data generated from the full *P. cinctipes* assembly for experimental designs of $n = 2$ and $n = 5$ biological replicates, respectively. Dotted and dashed lines represent hypothetical scenarios with sequencing depths of 30M reads per sample for $n = 2$ biological replicates (ENCODE consortium recommendation, 60M reads total; dotted line) or 20M reads per sample for $n = 5$ replicates (approximating our experimental design, 100M reads total; dashed line). At sequencing depths $<$150M total reads, Scotty analysis predicts very little difference in detection rate of differentially-expressed genes between these experimental designs. This implies that for our *P. cinctipes* assembly, gene detection rate was primarily influenced by total number of sequences generated, not by the number of sequences generated per sample.

similarity of *P. manimaculis* to other organisms relative to *P. manimaculis* may indicate proportionally higher sequence divergence in this species. Given the significant differences in tolerance thresholds between these crabs (Stillman 2002), it is tempting to speculate that at least a portion of this seemingly unique sequence in the more sensitive *P. manimaculis* may play a role in setting physiological limits in this species. However, our ability to detect such ecologically relevant adaptive signatures is currently limited by a lack of decapod genome sequences on which to map this transcriptomic data.

In general, the large number of contigs from both species which lacked high confidence BLAST hits (~69% in *P. cinctipes* and ~77% in *P. manimaculis*) could indicate the presence of a large number of *Petrolisthes*-specific gene sequences. The significantly lower annotation rate (~2%) among genes shared between the two *Petrolisthes* assemblies relative to the full transcriptomes supports this possibility, suggesting that the genes shared by the *Petrolisthes* congeners show higher sequence dissimilarity to other animal species than do the species-specific contigs. A general lack high-confidence BLAST hits has also been reported for at least two other crustacean species, *D. pulex* and the amphipod *P. hawiensis* (Zeng et al. 2011). It has been noted that even when compared to other arthropod references, the *D. pulex* genome contains a large proportion (~36%) of seemingly *Daphnia*-specific genes (Colbourne et al. 2011). The consistent reporting of low BLAST hit proportions among crustacean transcriptomes is tantalizing, potentially suggesting a greater number of clade-specific genes among this phylum relative to other animal lineages.

However, the lower proportion of BLAST hits might also be reflective of the lack of crustacean sequences in the major reference sequence databases. Although the Nr database does contain *D. pulex* sequence data, the Swiss-Prot database does not contain sequences from any crustacean. Further, only ~2% of the sequences contained in the Swiss-Prot database are of arthropod origin and the vast majority of these are from one species, *Drosophilla melanogaster*. It is therefore difficult at present to determine whether this pattern of low sequence similarity is the result of an increased prevalence of "private" genes in crustacean lineages or simply a lack of crustacean genomic resources for bioinformatic comparison. Furthermore, without an annotated reference genome for either species used in this study, it is impossible to accurately estimate the number of genes in *P. cinctipes* or *P. manimaculis* for comparison with *D. pulex*. There is thus a strong need to generate additional high-quality sequence data from a more diverse array of crustacean lineages to address these questions.

Characterization of transcriptomes and comparison to other crustacean assemblies

After BLAST hits had been compiled, Blast2GO (Conesa et al. 2005) was used to obtain GO terms for the top 10 BLAST hits for each sequence in the *Petrolisthes* assemblies. For *P. cinctipes*, of the 60,295 sequences with BLAST hits assigned at an expect value threshold of $e \leq 2e{-}5$, 43,707 sequences (72.5%) had associated GO terms. The *P. manimaculis* assembly had significantly higher GO annotation, with 59,796 of the 63,796 total sequences with BLAST hits returning associated GO terms (93.7%). Blast2GO results for both assemblies were categorized by select cellular component (Fig. 5(A)) and

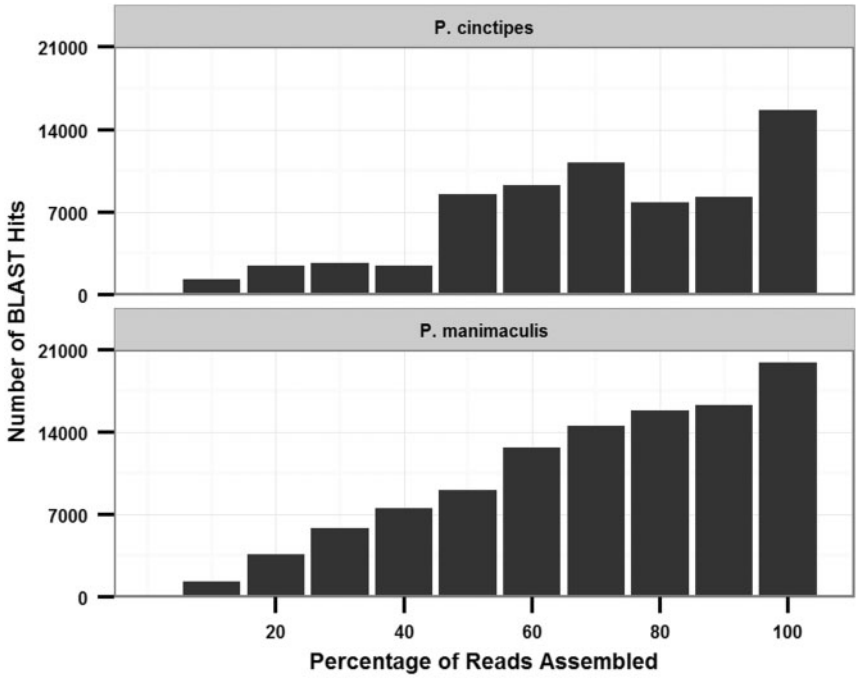

Fig. 4 Assessment of depth of assembly sequencing. Randomly sampled subsets of reads were generated in increments of 10% of the total reads for each species and were used to generate progressively larger sub-assemblies. An increase in the number returned of BLAST hits against the Swiss-Prot database with increasing proportion of the assembly utilized in the search suggests that adding sequencing depth improves gene discovery (e.g., for *P. manimaculis* above). A plateau in number of BLAST hits as a function of assembly fraction suggests saturation of gene discovery at those sampling depths (e.g., for *P. cinctipes*).

Table 2 Results of BLASTx analysis for each assembly and comparison to crustacean references

	Database	
	Nr[a]	Swiss-Prot[b]
P. cinctipes (194,105 total)		
Strong match[c]	–	60,295 (31.06%)
P. manimaculis (278,989 total)		
Strong match[c]	–	63,796 (22.87%)
P. hawaiensis		
Strong match[d]	10.8%	–
L. vannamei		
Weak match[e]	38.15%	–
Cherax quadricarnitus		
Strong match[f]	37%	–

[a] *Note:* Nr: all non-redundant GenBank CDS (Coding Sequence) translations.
[b] Swiss-Prot: January 30, 2016 release of the Swiss-Prot protein sequence database.
[c] Strong Match: expect $< 2e^{-5}$.
[d] Strong Match: expect $< 1e^{-10}$ used in this study (Zeng et al. 2011).
[e] Weak Match: expect $< 1e^{-3}$ used in this study (Johnson et al. 2015).
[f] Strong Match: expect $< 1e^{-10}$ used in this study (Ali et al. 2015).

molecular function (Fig. 5(B)) GO terms. Some contigs had GO terms for more than one cellular component or molecular function and were therefore counted in both categories. Of the selected terms for *P. cinctipes*, the majority of annotated transcripts were associated with the cytoplasm (6.7%), nucleus (5.1%), and mitochondrion (4.4%) with a smaller number of sequences assigned to the endoplasmic reticulum (1.0%). In *P. manimaculis*, annotated sequences were similarly distributed with cytoplasm (5.6%), nucleus (5.0%), and mitochondrion-associated (4.6%) transcripts dominating GO term matches. For both *Petrolishtes* species, there were a significantly lower number of transcripts associated with the cytoskeleton and ribosome compared with other crustacean assemblies.

To determine whether major functional categories of genes were missing or underrepresented in our assembly, we compared our GO term distribution with those generated from recent assemblies in other crustacean species (Fig. 5) including hepatopancreas tissue in the shrimp *L. vannamei* (Johnson et al. 2015), whole embryos of the amphipod *P. hawaiensis* and predicted transcripts from the sequenced genome of the water-flea *D. pulex* (Zeng et al. 2011). Comparison of our gill tissue transcriptome assemblies with these assemblies revealed significant overlap in functional gene categories of annotated genes, suggesting that overall our *de novo* transcriptomes do not lack major functional categories. However, the proportional representation

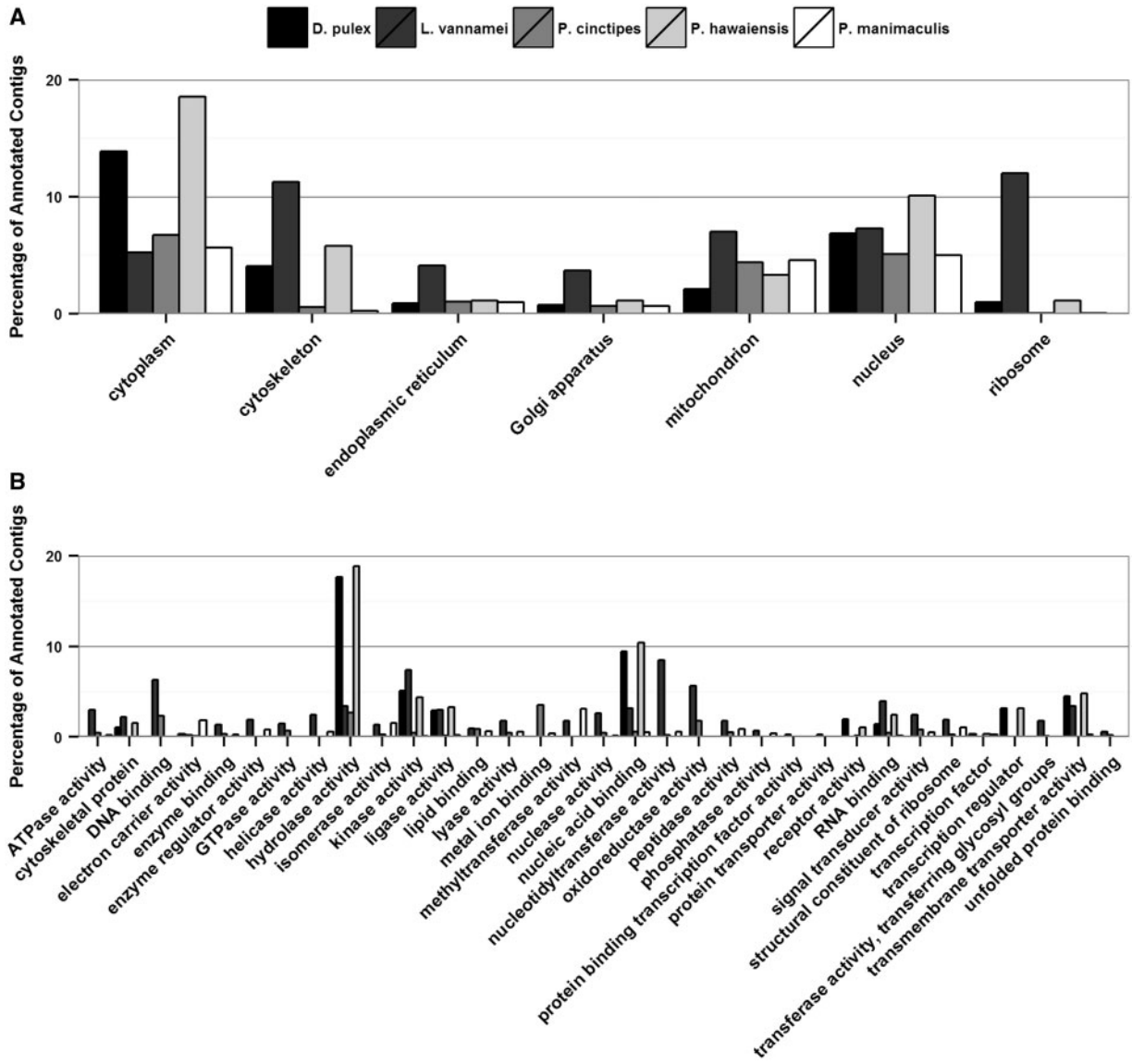

Fig. 5 Categorization of GO terms for cellular component (A) and molecular function (B) of top BLASTx hits for several recent crustacean assemblies. For comparison, GO term distributions from the shrimp *L. vannamei*, the amphipod *P. hawaiensis*, and the water-flea *D. pulex* (GO term predictions from sequenced genome) are shown. Categories with missing data for a species represent unreported numbers for that assembly.

of selected GO categories for molecular functions were significantly lower in *Petrolishtes* relative to other arthropods (Fig. 5(B)). This could be a result of low representation in the *Petrolisthes* assemblies among the functional categories selected for analysis (chosen based on previously reported categories for other arthropods). An alternative explanation is that our *Petrolisthes* assemblies display a more functionally diverse gene expression profile than the comparison assemblies, thereby reducing the proportional representation of any one functional GO category within the assembly. Given that our *Petrolishtes* assemblies were constructed from individuals exposed to a temporally variable, multi-stressor experimental regime, a greater diversity of expressed functional gene categories might be expected.

Expanding our search outside of the selected, comparative, GO terms revealed that for *P. cinctipes*, the largest represented functional gene categories were for metal ion binding (3.5%), hydrolase activity (2.7%), Adenosine triphosphate (ATP) binding (2.6%), DNA binding (2.3%), nucleotide binding (2.3%), and protein binding (2.2%). *P. manimaculis* exhibited a similar spread among the highest-represented functional categories of metal ion binding (3.1%), DNA binding (1.8%), and hydrolase

activity (1.5%). These results are consistent with functional GO category analyses of the *P. cinctipes* microarray sequence data carried out previously (Tagmount et al. 2010). Similar patterns were also observed within the arthropod reference assemblies mentioned above, with hydrolase activity and binding (in particular DNA binding) among the most highly represented functional gene categories across all assemblies (Fig. 5(B)). Although such broad-scale comparative analyses provide some useful insights into shared expression profiles across these taxa, functional interpretation of this data remains challenging. Ideally, NGS-generated genomic data would empower comparative functional analyses to identify shared response pathways to ecologically relevant stimuli. For example, similar proportional gene expression profiles in functional genetic categories under similar environmental regimes could highlight shared stress response pathways and generate new hypotheses for further mechanistic investigation. However, such detailed comparative analyses of functional gene categories are at present difficult to conduct for crustaceans given the lack of available EST data. Equally challenging is the inherent complexity of response that is recorded using NGS approaches such as RNA-seq. The significant differences in expression profiles between the two closely related congeners used in this study represent a significant challenge in the use of RNA-seq data for elucidating common functional responses. All of these challenges highlight the need to develop databases of Decapod genomic sequence in a diverse set of ecologically and evolutionarily informative species.

Summary

Although we have only just begun to fully employ NGS empowered approaches in non-model crustacean systems, the utility of these technologies in comparative physiological studies is considerable. As more functional genomic datasets become available, our ability to characterize the role of differential regulation in stress responsiveness and phenotypic plasticity will improve and may lead to the generation of common, broad-scale, hypothesis regarding genome to phenome linkages across arthropod taxa. We have generated two novel transcriptomes for the congeneric porcelain crabs *P. cinctipes* and *P. manimaculis*. The new sequence data generated for *P. cinctipes* nearly triples that previously available for this species (Tagmount et al. 2010). Surprisingly, the *de novo* transcriptomes generated for these two closely related species were significantly different even though they were constructed from the same number of individuals, tissues (gill), same acclimation treatments, sequencing library construction protocol, sequencing method, and sequence data analyses. Those differences emphasize the need for genome sequences on which to map-back transcriptomic data, and the inherent challenges with use of *de novo* transcriptomes for functional interpretation of differential gene expression studies using RNA-seq. In both *Petrolisthes* species, the low similarity to other arthropod sequence resources may suggest a greater prevalence of species-specific genes in crustacean lineages. However, because the vast majority of arthropod genomic resources currently available are for insects, it is at present difficult to confirm this speculation.

Further analysis of species-specific stress-induced expression profiles are planned for these assemblies and have the potential to elucidate regulatory processes that result in differential sensitivity to stress and which underlie or limit phenotypic plasticity in these closely related congeners. Although, the present lack of crustacean genomic resources poses significant challenges for developing this group as model systems in ecological and evolutionary biology, the rapid adoption of high-throughput NGS technologies in crustacean systems will likely make such investigation tractable in the near future. As more crustacean resources become available, we expect these functional genomic data to be of increasing value as a comparative genomics tool for biologists working within a wide variety of disciplines including physiology, ecology, and phylogeography.

Acknowledgments

We would also like to thank the NSF (BIO IOS-1551003 to D. Mykles, D. Durica, K. Burnett, and J.H.S.), the Society of Integrative and Comparative Biology, Division of Comparative Physiology and Biochemistry (SICB-DCPB), and The Crustacean Society for their support of this symposium and travel for E.J.A. at the 2015 SICB meeting in Portland, Oregon.

Funding

This work was supported by National Science Foundation funding [NSF BIO: MCB-1041225 to J.H.S.] and conducted with US Government support awarded by the Department of Defense, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship [32 CFR 168a to E.J.A.]. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by National

Institutes of Health (NIH S10 Instrumentation Grants S10RR029668 and S10RR027303). Computing on Pittsburgh Supercomputer Center was funded through XSEDE grant DEB140012 to J.H.S.

References

Alcivar-Warren A, Delaney M, Meehan-Meola D, Alcivar M, Warren W. 2009. Expressed sequence tags from cadmium-exposed postlarvae stage 42 of cultured, specific pathogen-free Pacific whiteleg shrimp, Litopennaeus vannamei. NCBI.

Ali MY, Pavasovic A, Mather PB, Prentis PJ. 2015. Analysis, characterisation and expression of gill-expressed carbonic anhydrase genes in the freshwater crayfish *Cherax quadricarinatus*. Gene 564:176–87.

Benner I, Diner RE, Lefebvre SC, Li D, Komada T, Carpenter EJ, Stillman JH 2013. Emiliania huxleyi increases calcification but not expression of calcification-related genes in long-term exposure to elevated temperature and p CO 2. Philos Trans R Soc Lond B Biol Sci 368:20130049.

Busby MA, Stewart C, Miller CA, Grzeda KR, Marth GT. 2013. Scotty: A web tool for designing RNA-Seq experiments to measure differential gene expression. Bioinformatics 29:656–7.

Carter H. a, Ceballos-Osuna L, Miller N. a, Stillman JH. 2013. Impact of ocean acidification on metabolism and energetics during early life stages of the intertidal porcelain crab *Petrolisthes cinctipes*. J Exp Biol 216:1412–22.

Ceballos-Osuna L, Carter H. a, Miller N. a, Stillman JH. 2013. Effects of ocean acidification on early life-history stages of the intertidal porcelain crab *Petrolisthes cinctipes*. J Exp Biol 216:1405–11.

Chávez-Mardones J, Gallardo-Escárate C. 2015. Next-Generation transcriptome profiling of the salmon louse *Caligus rogercresseyi* exposed to deltamethrin (AlphaMax^{TM}): Discovery of relevant genes and sex-related differences. Mar Biotech 17:793–810.

Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK et al. 2011. The ecoresponsive genome of *Daphnia pulex*. Science (New York, NY) 331:555–61.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. 2005. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–6.

Dana, J. 1852. Crustacea. In Wilkes C, Sherman C, editors. US Exploring Edition during the Years 1838-1842, Vol. XII. Philadelphia, PA, C. Sherman.

Das S, Pitts NL, Mudron MR, Durica DS, Mykles DL. 2016. Transcriptome analysis of the molting gland (Y-organ) from the blackback land crab, Gecarcinus lateralis. Comp Biochem Physiol Part D: Gen Prot 17:26–40.

Mitchell-Olds T. 2003. Evolutionary and ecological functional genomics. Nature reviews. Genetics 4:651–7.

Feder M, Mitchell-Olds T. 2003. Evolutionary and ecological functional genomics. Nature Reviews Genetics 4:651–7.

Gao Y, Zhang X, Wei J, Sun X, Yuan J, Li F, Xiang J. 2015. Whole transcriptome analysis provides insights into molecular mechanisms for molting in *Litopenaeus vannamei*. PLos One 10:e0144350.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotech 29:644–52.

Hahn D. a, Ragland GJ, Shoemaker DD, Denlinger DL. 2009. Gene discovery using massively parallel pyrosequencing to develop ESTs for the flesh fly Sarcophaga crassipalpis. BMC Gen 10:234.

Haig J. 1960. The Porcellanidae (Crustacea Anomura) of the Eastern Pacific. Allan Hancock Pacific Expedition, 24, pp. viii+-440.

Havird JC, Santos SR. 2016. Here we are, but where do we go? a systematic review of crustacean transcriptomic studies from 2014–2015. Integr Comp Biol 56:1055–66.

Johnson JG, Paul MR, Kniffin CD, Anderson PE, Burnett LE, Burnett KG. 2015. High CO2 alters the hypoxia response of the Pacific whiteleg shrimp (*Litopenaeus vannamei*) transcriptome including known and novel hemocyanin isoforms. Physiol Genomics 47:548–58.

Krueger F, Kreck B, Franke A, Andrews SR. 2012. DNA methylome analysis using short bisulfite sequencing data. Nat Meth 9:145–51.

Liu Y, Zhou J, White KP. 2014. RNA-seq differential expression studies: More sequence or more replication? Bioinformatics 30:301–4.

Magoč T, Salzberg SL. 2011. FLASH: Fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27:2957–63.

Miner BE, De Meester L, Pfrender ME, Lampert W, Hairston NG. 2012. Linking genes to communities and ecosystems: Daphnia as an ecogenomic model. Proc R Soc B: Biol Sci 279:1873–82.

Muller I, Stocco P, Marques M. 2010. Differentially expressed genes in gills of farmed shrimp Litopenaeus vannamei infected with White Spot syndrome Virus. NCBI.

Nilsen F, Malde K, Kongshaug H. 2010. Functional genomics in sea lice. NCBI.

Paganini AW, Miller NA, Stillman JH. 2014. Temperature and acidification variability reduce physiological performance in the intertidal zone porcelain crab *Petrolisthes cinctipes*. J Exp Biol 217:3974–80.

Poley JD, Igboeli OO, Fast MD. 2015. Towards a consensus: Multiple experiments provide evidence for constitutive expression differences among sexes and populations of sea lice (*Lepeophtheirus salmonis*) related to emamectin benzoate resistance. Aquaculture 448:445–50.

Robalino J, Almeida JS, McKillen D, Colglazier J, Trent HF, Chen YA, Peck ME, Browdy CL, Chapman RW, Warr GW et al. 2007. Insights into the immune transcriptome of the shrimp *Litopenaeus vannamei*: tissue-specific expression profiles and transcriptomic responses to immune challenge. Physiol Gen 29:44–56.

Roeding F, Borner J, Kube M, Klages S, Reinhardt R, Burmester T. 2009. A 454 sequencing approach for large scale phylogenomic analysis of the common emperor scorpion (Pandinus imperator). Mol Phylog Evol 53:826–34.

Ronges D, Walsh JP, Sinclair BJ, Stillman JH. 2012. Changes in extreme cold tolerance, membrane composition and cardiac transcriptome during the first day of thermal acclimation

in the porcelain crab *Petrolisthes cinctipes*. J Exp Biol 215:1824–36.

Stahl BA, Gross JB, Speiser DI, Oakley TH, Patel NH, Gould DB, Protas ME. 2015. A transcriptomic analysis of cave, surface, and hybrid isopod crustaceans of the species *Asellus aquaticus*. PLoS ONE 10:1–14.

Stepanyan R, Day K, Urban J, Hardin DL, Shetty RS, Derby CD, Ache BW, McClintock TS. 2006. Gene expression and specificity in the mature zone of the lobster olfactory organ. Physiol Gen 25:224–33.

Stillman JH. 2002. Causes and consequences of thermal tolerance limits in rocky intertidal porcelain crabs, genus Petrolisthes. Int Comp Biol 42:790–6.

Stillman JH. 2004. A comparative analysis of plasticity of thermal limits in porcelain crabs across latitudinal and intertidal zone clines. Intern Cong Ser 1275:267–74.

Stillman JH, Armstrong E. 2015. Genomics are transforming our understanding of responses to climate change. BioScience 65:237–46.

Stillman JH, Colbourne JK, Lee CE, Patel NH, Phillips MR, Towle DW, Eads BD, Gelembuik GW, Henry RP, Johnson EA et al. 2008. Recent advances in crustacean genomics. Int Comp Biol 48:852–68.

Stillman JH, Reeb C. a. 2001. Molecular phylogeny of Eastern Pacific porcelain crabs, genera Petrolisthes and Pachycheles, based on the mtDNA 16S rDNA sequence: phylogeographic and systematic implications. Mol Phylog Evol 19:236–45.

Stillman JH, Somero GN. 2000. A comparative analysis of the upper thermal tolerance limits of eastern Pacific porcelain crabs, genus Petrolisthes: influences of latitude, vertical zonation, acclimation, and phylogeny. Physiol Biochem Zool?: PBZ 73:200–8.

Stillman JH, Tagmount A. 2009. Seasonal and latitudinal acclimatization of cardiac transcriptome responses to thermal stress in porcelain crabs, *Petrolisthes cinctipes*. Mol Ecol 18:4206–26.

Stillman JH, Teranishi KS, Tagmount A, Lindquist EA, Brokstein PB. 2006. Construction and characterization of EST libraries from the porcelain crab, *Petrolisthes cinctipes*. Int Comp Biol 46:919–30.

Stillman J, Somero G. 1996. Adaptation to temperature stress and aerial exposure in congeneric species of intertidal porcelain crabs (genus Petrolisthes): correlation of physiology, biochemistry and morphology with vertical distribution. J Exp Biol 199:1845–55.

Tagmount A, Wang M, Lindquist E, Tanaka Y, Teranishi KS, Sunagawa S, Wong M, Stillman JH. 2010. The porcelain crab transcriptome and PCAD, the porcelain crab microarray and sequence database. PloS One 5:e9327.

Tarrant AM, Baumgartner MF, Hansen BH, Altin D, Nordtug T, Olsen AJ. 2014. Transcriptional profiling of reproductive development, lipid storage and molting throughout the last juvenile stage of the marine copepod Calanus finmarchicus. Front Zool 11:91.

Teranishi KS, Stillman JH. 2007. A cDNA microarray analysis of the response to heat stress in hepatopancreas tissue of the porcelain crab *Petrolisthes cinctipes*. Comp Biochem Physiol - Part D: Gen Prot 2:53–62.

Vega EDI, O'Leary NA, Robalino J, Peck ME, Bartlett TC, Richards M, Hikima S, Browdy CL, Warr GW, Chapman RW, Gross PS. 2008. *Litopenaeus vannamei* Gene Discovery Project. NCBI.

Verslycke T, Tarrant A, Stegeman J, McDowell J. 2009. Express sequence tags associated with asymptomatic and shell-diseased lobsters. NCBI.

Wang X, Wang S, Li C, Chen K, Qin JG, Chen L, Li E. 2015. Molecular Pathway and Gene Responses of the Pacific White Shrimp *Litopenaeus vannamei* to Acute Low Salinity Stress. J Shellfi Res 34:1037–48.

Xu Z, Li T, Li E, Chen K, Ding Z, Qin JG, Chen L, Ye J. 2016. Comparative transcriptome analysis reveals molecular strategies of oriental river prawn *Macrobrachium nipponense* in response to acute and chronic nitrite stress. Fi Shellfi Immunol 48:254–65.

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