Contents lists available at ScienceDirect

Comparative Biochemistry and Physiology, Part D

journal homepage: www.elsevier.com/locate/cbpd

Characterization and analysis of a transcriptome from the boreal spider crab *Hyas araneus*

Lars Harms [a,*], Stephan Frickenhaus [b], Melanie Schiffer [a], Felix C. Mark [a], Daniela Storch [a], Hans-Otto Pörtner [a], Christoph Held [c], Magnus Lucassen [a]

[a] *Integrative Ecophysiology, Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany*
[b] *Scientific Computing, Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany*
[c] *Functional Ecology, Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany*

ARTICLE INFO

Article history:
Received 22 May 2013
Received in revised form 5 September 2013
Accepted 30 September 2013
Available online xxxx

Keywords:
Spider crab (*Hyas araneus*)
Crustacea
Transcriptome
Illumina sequencing
454 sequencing
Markov Clustering Algorithm
Anti-viral immunity
Reverse transcriptase

ABSTRACT

Research investigating the genetic basis of physiological responses has significantly broadened our understanding of the mechanisms underlying organismic response to environmental change. However, genomic data are currently available for few taxa only, thus excluding physiological model species from this approach. In this study we report the transcriptome of the model organism *Hyas araneus* from Spitsbergen (Arctic). We generated 20,479 transcripts, using the 454 GS FLX sequencing technology in combination with an Illumina HiSeq sequencing approach. Annotation by Blastx revealed 7159 blast hits in the NCBI non-redundant protein database. The comparison between the spider crab *H. araneus* transcriptome and EST libraries of the European lobster *Homarus americanus* and the porcelain crab *Petrolisthes cinctipes* yielded 3229/2581 sequences with a significant hit, respectively. The clustering by the Markov Clustering Algorithm (MCL) revealed a common core of 1710 clusters present in all three species and 5903 unique clusters for *H. araneus*. The combined sequencing approaches generated transcripts that will greatly expand the limited genomic data available for crustaceans. We introduce the MCL clustering for transcriptome comparisons as a simple approach to estimate similarities between transcriptomic libraries of different size and quality and to analyze homologies within the selected group of species. In particular, we identified a large variety of reverse transcriptase (RT) sequences not only in the *H. araneus* transcriptome and other decapod crustaceans, but also sea urchin, supporting the hypothesis of a heritable, anti-viral immunity and the proposed viral fragment integration by host-derived RTs in marine invertebrates.

1. Introduction

The great spider crab, *Hyas araneus*, is a benthic decapod crab that lives on sublitoral rocky or sandy substrates to a depth of 50 m (Hayward and Ryland, 1990). Within the North-East Atlantic region it is distributed along a latitudinal gradient from the English Channel up to the Arctic regions of Spitsbergen, where it represents one of the most prominent brachyuran crabs (Zittier et al., 2012). The size of its distribution range and the corresponding cline in environmental conditions make *H. araneus* an ideal species to study the effects of environmental changes as well as functional differentiation between populations. For example, decreased larval developmental rates in Arctic compared to temperate populations suggest adaptation to the polar cold (Walther et al., 2010). Elevated seawater PCO_2 (as projected by ocean acidification scenarios) caused an increase in metabolic rate during larval development pointing to higher metabolic costs in larvae (Schiffer et al., 2012). Adult *H. araneus* displayed increased heat sensitivity under elevated CO_2 levels with potential consequences for biogeographical boundaries (Walther et al., 2009). However, the genetic basis of these responses to environmental changes has so far only been investigated for a limited number of candidate genes. For example, hyastatin, a peptide involved in haemolymph antimicrobial defense, has been isolated, and the importance of the cys-containing region for the antimicrobial activity and a possible multifunctional character has been demonstrated (Sperstad et al., 2009). The reason for the small number of studies is likely the lack of genomic information in databases like the National Center for Biotechnology Information (NCBI). Currently, only 26 nucleotide sequences of *H. araneus* are published in NCBI.

In recent years, Next Generation Sequencing (NGS) has made it possible to approach this problem by sequencing and assembly of entire genomes of ecologically relevant species (for review see Wheat, 2010). However, for non-model organisms, sequencing a transcriptome rather than the genome to obtain the genetic data is advantageous for many reasons. The generation of sequence data is quick, it is relatively cost-effective and can thus provide the genetic basis for studies with fewer resources. Further, transcriptome sequencing can provide both

Abbreviations: MCL, Markov Clustering Algorithm; CEGMA, Core Eukaryotic Genes Mapping Approach.

* Corresponding author.
E-mail address: LHarms@gmx.net (L. Harms).

1744-117X/$ – see front matter © 2013 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.cbd.2013.09.004

expression and coding data, using RNA-seq (Martin and Wang, 2011). Using different tissues and differentially treated animals it is possible to capture variations in coding sequences, stress induced sequences as well as differences in the expression level. Respective approaches have already been applied to a number of marine invertebrates to achieve insights into expression information (Giant Ezo scallop (Hou et al., 2011); common octopus (Zhang et al., 2012); 2 Mollusca, 2 Arthropoda, 2 Annelida, 2 Memertea, 2 Porifera (Riesgo et al., 2012); pearl oyster (Shi et al., 2013)) thereby expanding the existing genetic resources massively.

Thus, the objective of the present study was to fundamentally characterize the transcriptome of *H. araneus*. For analyzing specific homologies within decapod transcriptomes and for identifying common and specific gene clusters of the selected group of species we introduced the Markov Clustering Algorithm (MCL) clustering approach.

To develop an extensive transcriptome of *H. araneus* we combined the 454 and Illumina sequencing technologies on normalized and common cDNA libraries constructed from pooled samples of multiple tissues from animals treated with different environmental conditions (see Materials and methods). We assembled the sequences to reconstruct transcripts potentially representing the *H. araneus* transcriptome. Because no reference genome is available for *H. araneus* we assembled the transcriptome *de novo*. There are several *de novo* tools available, but none represent the perfect solution (Kumar and Blaxter, 2010). To obtain a comprehensive and high-quality de novo assembly of the *H. araneus* transcriptome, we tested different assembling tools and compared the resulting assemblies. In the second part we analyzed the functionally annotated transcriptome for particular features and compared the identified sequences with available sequence information of other decapod crustaceans using the MCL-clustering to reveal homologies within the selected group of species.

The approach illustrates a potential methodological framework and may promote further transcriptome studies in non-model organisms. The transcriptome obtained for *H. araneus* will become essential for future analyses and annotations and also provide useful information for future functional genomic studies in crustaceans.

2. Materials and methods

2.1. Sample preparation and RNA extraction

Adult specimens of the boreal spider crab *H. araneus* were collected in the Kongsfjord (N 78° 58.635′; E 11° 29.454′) at the west coast of Spitsbergen (Norway). Animals were acclimated for 10 weeks in flow through aquaria systems to 6 different treatments of 3 seawater PCO_2 values (390, 1120 and 1960 μatm) combined with two temperatures (5 and 10 °C), respectively. Tissue samples comprising of all 6 gill-arches, tegument, heart, hepatopancreas, testis and pincer muscle were collected from four to six animals per treatment and directly frozen in liquid nitrogen. Samples were stored at −80 °C until used for RNA extraction. Total tissue RNA was extracted by using the RNeasy Mini Kit according to the "Purification of Total RNA from Animal Tissue" protocol (QIAGEN, Hilden, Germany). RNA quantities were determined by a NanoDrop 2000c spectrometer (PeqLab, Erlangen, Germany), and RNA was analyzed for quality by microfluidic electrophoresis in an Agilent 2100 Bioanalyzer (Agilent Technologies).

2.2. Sequencing and assembly

To generate the transcriptome of the non-model organism *H. araneus*, two different sequencing approaches were used. First, a 454 pyrosequencing approach based on normalized cDNA libraries was applied, serving as a basis for the assembly. Using samples from multiple tissues and differentially treated animals (rearing temperature and PCO_2 level) as well as using a normalized cDNA libraries allow for a comprehensive transcriptome, capturing variations in coding sequences, stress induced sequences as well as low expressed genes. Two separate cDNA libraries were sequenced by 454: a library exclusively based on gill samples and a library based on samples of a mixture of tissues. For the *H. araneus* gill library, the same amount of RNA was collected from each gill of 4 animals per treatment and pooled in one sample. The same was done for all other tissues to prepare the material for a mixed tissue library. Both mixtures were used for the library constructions by the Max Planck Institute for Molecular Genetics (Berlin, Germany). Total RNA of the two pools (gill and mixed tissue) was used for cDNA synthesis using the SMART protocol (Mint-Universal cDNA synthesis kit, Evrogen, Moscow, Russia). The cDNA was subsequently normalized using duplex-specific nuclease and re-amplified thereafter following the instructions of the "Trimmer Kit" (Evrogen, Moscow, Russia). Sequencing libraries were prepared from cDNA using the "GS FLX Titanium General Library Preparation Kit" (Roche, Basel, Switzerland). Before sequencing, the libraries were amplified by polymerase chain reaction (PCR) using the 'GS FLX Titanium LV emPCR Kit' (Roche, Basel, Switzerland) (De Gregoris et al., 2011). Sequencing was performed by the Max Planck Institute for Molecular Genetics (Berlin, Germany) on a 454 Genome Sequencer FLX using the Titanium chemistry (Roche). Initial quality control and filtering of adapters and barcodes was performed at the Max Planck Institute for Molecular Genetics (Berlin, Germany). Both cleaned libraries were combined for the subsequent *de novo* assembly. To optimize the quality of the *de novo* transcriptome assembly, we compared two different assembler programs: GS De Novo Assembler version 2.6 (Newbler, Roche) and MIRA 3.0 (Chevreux and Wetter, 1999). We tested each program with the following main assembly parameters: minimum percentage identities of 95%, and minimum overlap length of 40 bp for MIRA, and 40 bp for the GS De Novo Assembler. The "-cdna" mode was used for the GS De Novo Assembler. The final de novo assembly by GS De Novo Assembler was chosen based on basic assembly metrics and performance in terms of completeness and contiguity.

Secondly, an Illumina sequencing approach was used to enhance the 454 based transcriptome. Six different cDNA libraries based on samples of the six different treatments were sequenced. For each treatment, total RNA from all gills of 4 animals was pooled and used for the library construction by GATC Biotech (Konstanz, Germany). Libraries for each treatment were constructed according to the 'SMART protocol for Illumina sequencing' (Clontech, Mountain View, CA, USA). Illumina single-end sequencing was performed on a HiSeq 2000 Sequencer by GATC Biotech. Initial quality control and filtering of adapters was performed by GATC Biotech. In addition, obtained raw reads were quality controlled by FastQC (Babraham Institute, Cambridge, UK) and cleaned using the FastX-Toolkit (Hannon Lab — Cold Spring Harbor Laboratory, NY, USA). Quality control was performed using the following parameters: minimum quality score of 20, minimum percentage of bases within the quality score of 90 and a minimum length of 25 bases. To enhance the set of GS De Novo Assembler-assembled contigs, obtained Illumina-data from the six libraries were combined and reads were assembled *de novo* with ABySS version 1.3.2 (Simpson et al., 2009) with k = 26, minimum overlap length of 30 bp and minimum sequence identity of 0.9. Considerable overlaps with the GS De Novo Assembler-assembled 454-data were detected with blastn (word size 8), removing Abyss-contigs above E-value 10^{-10} and length below 500 bp. The transcriptome of *H. araneus* was deposited in the 'European Nucleotide Archive' (ENA) at the 'European Molecular Biological Laboratory-European Bioinformatics Institute' (EMBL-EBI) (Accession range: HAAI01000001–HAAI01019199).

2.3. Functional annotation

Functional annotation of the *H. araneus* transcriptome was accomplished using the Blast2GO software v.2.6.0 (Conesa et al., 2005; Gotz et al., 2008). Homology searches were performed using Blastx against the NCBI non-redundant protein database. Blast searches were performed with an E-value cut-off of $1E^{-3}$. For the Gene Ontology (GO)

Please cite this article as: Harms, L., et al., Characterization and analysis of a transcriptome from the boreal spider crab *Hyas araneus*, Comp. Biochem. Physiol., D (2013), http://dx.doi.org/10.1016/j.cbd.2013.09.004

classification of the blasthits, the default parameters were used (E-value $<1E^{-6}$, annotation cut-off >55 and a GO weight >5). The annotated most specific GO terms were traced back to the second level parent term using the R Bioconductor package "GO.db" (Carson et al., 2010).

2.4. Comparative analysis

The assembled transcript sequences were compared with EST sequence libraries from *Homarus americanus* and *Petrolisthes cinctipes* obtained from Genbank. After filtering for length >500, a set of 25,185, 75,208 and 13,706 sequences (from *H. americanus*, *P. cinctipes*, and *H. araneus*, respectively) were clustered following the Markov Cluster Algorithm (MCL) (Enright et al., 2002) based on tBlastx tables (all against all) with an E-value $<10^{-9}$ and negative $\log_{10}$ E-value as similarity. The MCL-inflation parameter was $I = 2$. The obtained set of 35,440 MCL-clusters was divided in species-specific or overlapping groups. Library-specific sequence counts within these cluster sets were computed. All analyses were performed in R (R Core Team, 2012). The R-script producing the counts in Venn-diagrams for clusters and sequences is available on our web-server http://www.awi.de/en/go/bioinformatics. The common core of clustered transcript sequences comprises 3245 *H. araneus* sequences of which a total of 2194 were found annotated in Blast2GO. This set has been taken for a GO enrichment analysis against the full set of annotated *H. araneus* transcript sequences (Fisher's exact test). The dataset of CEGMA 2.4 (Parra et al., 2007) was used to screen transcript sequences longer than 200 bp for universal eukaryotic functions using trpsblastn (Altschul et al., 1997) applied as in Windisch et al. (2012). Top CEGMA hits were analyzed on the basis of MCL clustering results for the core set of *H. araneus* sequences and the *H. araneus* specific MCL clusters.

2.5. Comparative analysis of reverse transcriptase sequences

For an extended analysis resolving similarity features with reverse transcriptase (RT) sequences, tBlastx analyses of published transcript sequences from *Drosophila melanogaster* ($N = 27{,}539$) and *Strongylocentrotus purpuratus* ($N = 23{,}057$) were incorporated in MCL with a more stringent lower E-value cut-off $1E^{-25}$. For this, *H. araneus* sequences were pooled with the *H. americanus* and *P. cinctipes* sequence libraries as crustaceans. Sequence IDs of all non-*H. araneus* libraries used in the tBlastx runs are listed in the supplement text file contained in Xseq-IDs.zip.

3. Results and discussion

3.1. 454-sequencing and assembly

The two 454 pyro-sequencing runs based on normalized cDNA libraries constructed with total RNA from 6 different tissues (gills, tegument, heart, hepatopancreas, testis and pincer muscle) yielded a total of 1,111,880 reads with 335 Mbp and an average length of 550 bp (Table 1). The 454 reads originating from the two sequenced cDNA libraries were assembled with GS De Novo Assembler 2.6 (Newbler, Roche). After internal trimming, a total of 824,230 reads (260 Mbp) with an average length of 300 bp were assembled into 16,614 isotigs. The size of the reads extends to a maximum of 871 bp with a peak between 300 and 480 bp (Fig. 1A). The obtained isotigs had a maximum length of 6697 bp, an average length of 668 bp and a N50 isotig size of 751 bp (Table 1). Isotigs with a length shorter than 100 bp were excluded from the analysis. The size distribution of the isotigs ranges from 100 to 6697 bp with a peak between 400 and 600 bp (Fig. 1B). The estimated average fold coverage of the isotigs was 6 and ranged from 1 to more than 2000 (Fig. 2).

Table 1
454 sequence and assembly statistics. Gill and mixed tissue sequencings are combined for statistics. Only isotigs with a length greater than 100 bp are considered in the assembly statistics.

Raw sequencing reads	
Number of reads (gill tissue)	551,904
Number of reads (mixed tissue)	559,976
Number of reads (total)	1,111,880
Total size (bp)	335,440,200
Average size (bp)	550
Aligned reads	
Number of reads	824,230
Total size (bp)	259,700,556
Average size (bp)	300
Assembly statistics	
Number of isotigs	16,614
Total size (bp)	11,105,636
Average size (bp)	668
Maximum length, bp	6697

Q2

3.2. Enhancement of the transcriptome by Illumina sequencing

An Illumina sequencing approach was carried out with total RNA from gill tissue generated from animals long-term exposed to six different climate conditions. The obtained sequences were used to enhance the transcriptome generated by the 454 *de novo* assembly. The Illumina sequencing runs yielded a total of 98,508,658 reads with 9457 Mbp and an average length of 96 bp (Table 2). The Illumina reads originating from the six treatments were assembled into contigs by use of the ABySS Assembler (Simpson et al., 2009). A total of 55,354,912 reads with an

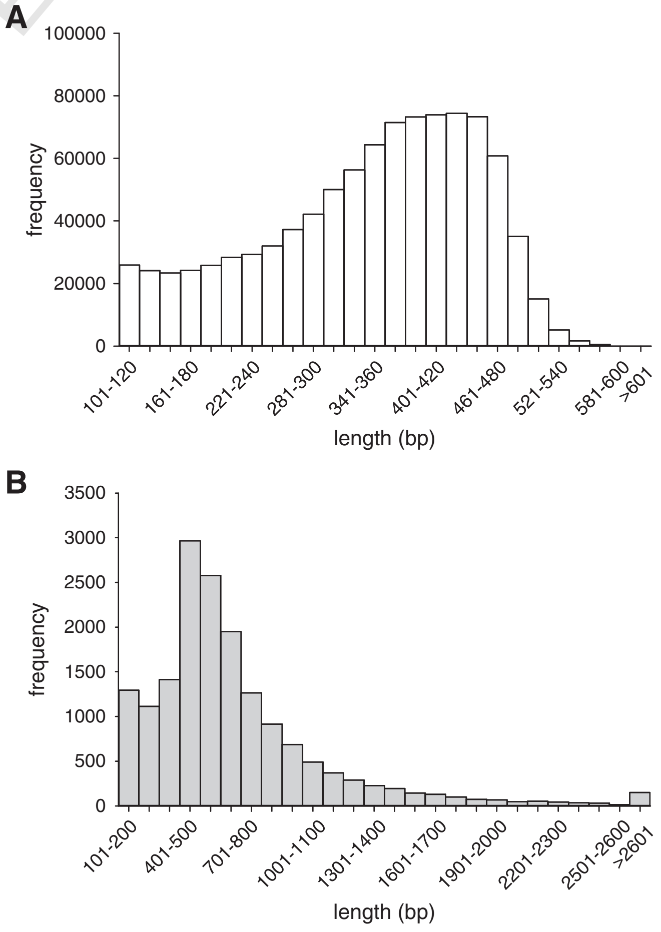

Fig. 1. Size distribution of reads and isotigs from 454 pyrosequencing. Reads (A) and isotigs (B) longer than 100 bp are considered.

Please cite this article as: Harms, L., et al., Characterization and analysis of a transcriptome from the boreal spider crab *Hyas araneus*, Comp. Biochem. Physiol., D (2013), http://dx.doi.org/10.1016/j.cbd.2013.09.004

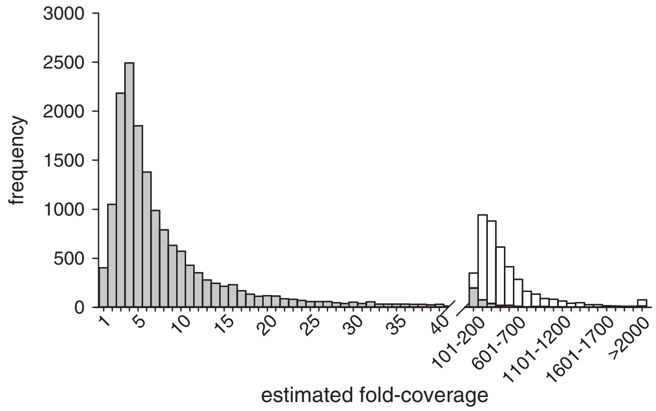

Fig. 2. Estimated fold-coverage of assembled isotigs/contigs and frequency of sequences with an according fold-coverage. Isotigs from 454 assembly (gray) and contigs from Illumina assembly (white).

average length of 61 bp were assembled into 175,612 contigs with a length greater than 100 bp. The contigs of the Illumina assembly had a maximum length of 3094 bp and an average length of 195 bp and a N50 isotig size of 213 bp (Table 2). A relatively short average contig length of only 195 bp can be explained by the short reads of the Illumina sequencing (61 bp) and the lack of a reference transcriptome/genome in non-model organisms. Similar results were reported for the non-model organism *Radix balthica* (snail) testing four different assemblers (Feldmeyer et al., 2011). We focused on transcripts of potentially greater functional relevance and excluded contigs with a length shorter than 500 bp from the ABySS Illumina assembly. All Illumina-based contigs showing an overlap with the Newbler 454 contigs were excluded to avoid redundancy. In total, 3865 contigs were used for further analysis and added to the existing 454 assembly to complement the transcriptome to a total of 20,479 transcript sequences. The estimated average fold coverage of the reduced set of contigs was 471 and ranged from 100 to over 2000, whereas – due to the larger sequence volume – the contigs led to a distinctly greater coverage than the isotigs of the 454 assembly (Fig. 2).

Recently, a comparative description of ten invertebrate transcriptomes was based solely on Illumina *de novo* sequencing and assembly (Riesgo et al., 2012), emphasizing the importance of sequencing invertebrate non-model species as a powerful basis for phylogenetic and functional genomic studies. In average, about 40% of all reads could be assembled, resulting in about 67,000 to 210,000 contigs across the ten species (Riesgo et al., 2012). Based on our Illumina approach, we were able to assemble slightly more reads (56%) into a comparable number of contigs within the present sequencing project.

Table 2
Illumina sequence and assembly statistics. Sequencings of treatments are combined for statistics. Only contigs with a length greater than 100 bp are considered in the Illumina sequence statistics.

Raw sequencing reads	
Number of reads (treatment 1)	12,670,506
Number of reads (treatment 2)	18,137,025
Number of reads (treatment 3)	12,363,308
Number of reads (treatment 4)	21,113,665
Number of reads (treatment 5)	21,110,842
Number of reads (treatment 6)	13,113,312
Number of reads (total)	98,508,658
Total size (bp)	9,456,831,168
Average size (bp)	96
Aligned reads	
Number of reads	55,354,912
Total size (bp)	3,397,642,905
Average size (bp)	61
Assembly statistics	
Number of contigs	175,612
Total size (bp)	34,271,175
Average size (bp)	195
Maximum length (bp)	3094

3.3. Functional analysis

In order to annotate the consensus sequences, a Blastx search against the NCBI non-redundant (nr) protein database was performed using the Blast2GO suite (Conesa et al., 2005; Gotz et al., 2008). For the 20,479 transcripts the search revealed 7159 (35%) significant blast hits (1E^{-3} cutoff threshold), which corresponded to 5962 unique accession numbers. The large number of transcripts without a significant blast hit (65%) is probably caused by a high proportion of novel genes and the lack of fully annotated transcriptomes in closely related crustaceans. The distribution of annotated and non-annotated transcripts is only slightly influenced by the length, which can be explained by the fact that a moderately restrictive E-value cut-off was used to obtain a comprehensive set of blasthits. A similar ratio of annotated and non-annotated isotigs/contigs was observed in the transcriptome (Fig. 3). In contrast, the quality of the annotation strongly depended on the transcript lengths. Table 3 lists the 20 consensus sequences with the highest E-value and the highest score. All transcripts with a strong match in the blast search belonged to sequences with a great length (>1500 bp). Unsurprisingly, there are several heat-shock proteins included in the top 20 list, as heat-shock proteins are often conserved across phyla (Lindquist and Craig, 1988). Furthermore, a potential bias due to the large number of studies with a focus on specific gene groups must certainly be taken into account.

Gene Ontology (GO) terms of the *H. araneus* transcriptome were analyzed using Blast2GO (Consortium, 2008). Blast2GO provides information on the 'Molecular Function', the 'Cellular Component' and the 'Biological Process' for each sequence. In total, 27,074 GO terms could be allocated for 4156 (58.1%) sequences. The annotated GO terms are grouped in 7226 (26.69%) on 'Molecular Function', in 6414 (23.69%) on 'Cellular Component' and in 13,434 (49.62%) on 'Biological Process' (Fig. 4). For each sequence, the specific annotated GO term was mapped to the second level parent term to obtain a broader overview of the functionally grouped transcripts for the three GO ontologies (Fig. 4). The hierarchical order of the GO allows to consider gene sets involved in a specific process at a specific detail level of interest. For the 'Biological Processes', the most frequent categories were 'cellular process' (28.7%), 'biological regulation' (23.3%), 'cellular component organization or biogenesis' (13.5%) and 'developmental process' (13.1%), followed by 'response to stimulus' (6.5%), 'establishment of localization' (6.4%), and

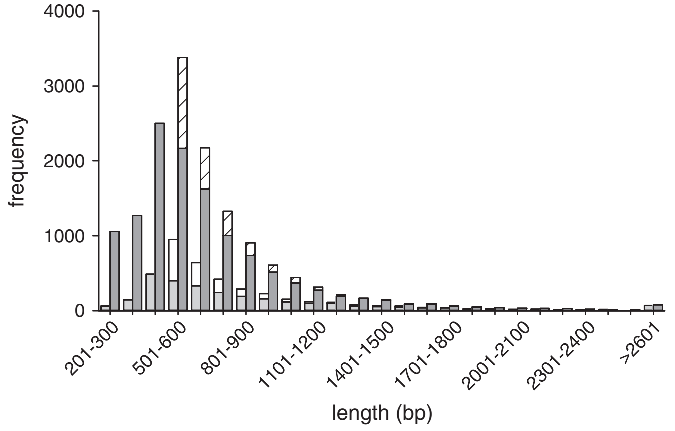

Fig. 3. Size distribution of annotated and non-annotated isotigs (454 sequencing)/contigs (Illumina sequencing). Annotated isotigs (gray), annotated contigs (white), non-annotated isotigs (dark gray) and non-annotated contigs (white with diagonal lines).

Table 3
Top 20 high quality annotations of the *Hyas araneus* transcriptome. Sequences with highest score in Blastx search.

Putative sequence description	Length	Score	ACC number	Species	Type
Myosin heavy chain type a	6490	2510	BAK61429.1	*Marsupenaeus japonicus*	Full length
Elongation factor 2	3273	1556	ACS36538.1	*Homarus americanus*	Full length
Na^+/K^+ ATPase alpha subunit	4496	1427	AAG47843.1	*Callinectes sapidus*	Partial
Myosin heavy chain type b	5507	1413	BAK61430.1	*Marsupenaeus japonicus*	Full length
Low-density lipoprotein receptor protein like	2477	1161	XP_002430267.1	*Pediculus humanus corporis*	Partial
UDP-n-acetylglucosamine, n-acetylglucosaminyltransferase	2180	1153	XP_003249419.1	*Apis mellifera*	Partial
Heat shock protein 70	2427	1046	CAL68989.1	*Cyanagraea praedator*	Full length
Heat shock protein 70	2204	1046	ACE79213.1	*Scylla paramamosain*	Full length
hypothetical protein	3414	1045	EFX68045.1	*Daphnia pulex*	Partial
DNA topoisomerase 2 like	1808	988	XP_002428978.1	*Pediculus humanus corporis*	Partial
Ubiquitin-activating enzyme like	3834	966	EFX89910.1	*Daphnia pulex*	Partial
Pre-mRNA-processing-splicing factor like	1896	950	EFX85628.1	*Daphnia pulex*	Partial
Peroxinectin	2721	944	ABB55269.2	*Fenneropenaeus chinensis*	Full length
Translation initiation factor like	2460	910	EFX65461.1	*Daphnia pulex*	Partial
Elongation factor	1651	863	ADK25705.1	*Cancer borealis*	Full length
Glucose regulated protein 78 (GRP78)	1875	840	ABM92447.1	*Fenneropenaeus chinensis*	Partial
Catalase	3032	834	ACX46120.1	*Scylla paramamosain*	Partial
Polyadenylate-binding protein 1 like isoform	2852	828	XP_003398393.1	*Bombus terrestris*	Full length
ATP-synthase subunit mitochondrial	1710	827	ADC55251.1	*Litopenaeus vannamei*	Full length
Tubulin beta-2c chain	1406	823	Q94571.1	*Homarus americanus*	Full length

'metabolic process' (3.6%). Other 'Biological Process' categories such as 'localization', 'multicellular organismal process' are present, but at a lower percentage. In the 'Molecular Function' category, most of the terms are grouped into the 'binding' (54.3%) and 'catalytic activity' (33.3%) categories, followed by 'transporter activity' (3.8%) and 'structural molecule activity' (3.5%). Terms such as 'enzyme regulator activity, molecular transducer activity', 'nucleic acid binding transcription factor activity' and 'protein binding transcription factor activity' are also present, but constitute a smaller proportion. The 'Cellular Component' category indicates that over 95% ('cell part') of annotated sequences are of cellular origin. Other categories such as 'extracellular region part', 'cell junction', 'synapse and macromolecular complex' are only present in small numbers.

In comparison to other studies, the distribution of genes based on the GO terms and the three categories is consistent. In a study carried out on the porcelain crab *P. cinctipes*, 'binding and catalytic activity' were the most represented terms in the 'Molecular Function' category (Tagmount et al., 2010). In addition, 'cellular process' was the major term in 'Biological Process'. The study used different GO category levels, thus the distributions are difficult to compare. However, a similar classification was obtained for the scallop *Patinopecten yessoensis* and the octopus *Octopus vulgaris* (Hou et al., 2011; Zhang et al., 2012). Only

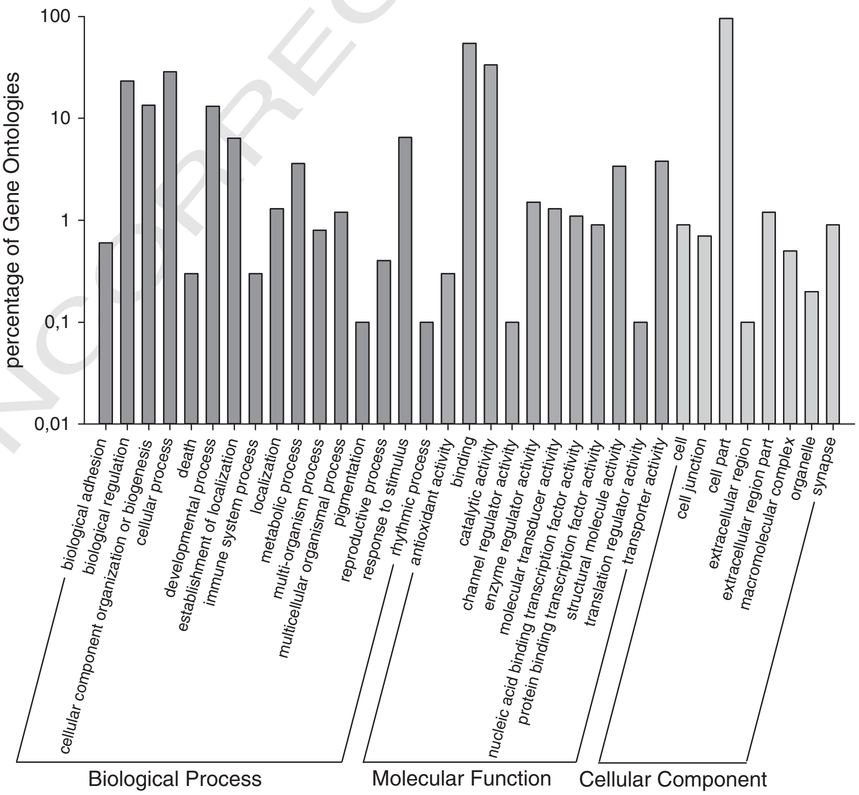

Fig. 4. Gene ontologies. Transcript counts for Gene Ontology (GO) classification of the *Hyas araneus* transcriptome for biological process (dark gray), molecular function (white) and cellular component (gray) categories.

Please cite this article as: Harms, L., et al., Characterization and analysis of a transcriptome from the boreal spider crab *Hyas araneus*, Comp. Biochem. Physiol., D (2013), http://dx.doi.org/10.1016/j.cbd.2013.09.004

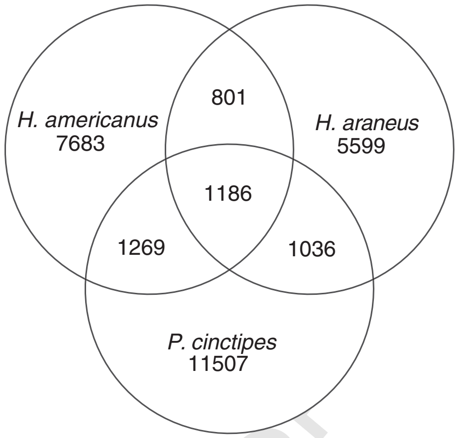

Fig. 5. Distribution of the MCL clusters built from tBlastx hits. Comparative sequence analysis of the *Petrolisthes cinciptes* and *Homarus americanus* EST libraries with the *Hyas araneus* transcriptome.

the ‘metabolic process’ category seems to be underestimated in the ‘Biological Process’ category of *H. araneus*, as a distinctively larger proportion of ‘metabolic process’ GO terms (12–30%) was observed in the former studies (Tagmount et al., 2010; Hou et al., 2011; Zhang et al., 2012). One explanation could be that a large fraction of the sequencing volume was based on gill tissue due to the focus of the sequencing project, but it could also be differences in quality and degree of sequence clustering in the assemblies. Furthermore, it must be taken into account that a possible bias exists due to the large proportion of vertebrate sequences in the common databases. However, the results of our gene ontology analysis suggest a diverse and representative gene set of the *H. araneus* transcriptome. In addition, when using the transcriptome to identify unknown proteins in a parallel proteomic study, the quality of the transcriptome was confirmed by its capacity to identify 58% of the proteins (Harms et al. unpublished).

3.4. Comparison with H. americanus and P. cinctipes databases

We used the *H. araneus* transcriptome, the porcelain crab *P. cinctipes* (Stillman et al., 2006) and the European lobster *H. americanus* (Towle and Smith, 2006) EST libraries for a comparative analysis to identify similarities and differences between decapod crustaceans. A tBlastx approach with an E-value cut-off of 1E^{-5} was performed for all sequences from each species as query against all others. The results for all sequences with a length ≥500 bp are shown in Table 4 in terms of counts of best hits. For *H. araneus*, 1154 cDNAs have a hit in *P. cinctipes* and 1851 in *H. americanus*. The blast hits suggest a low similarity of *H. araneus* transcripts with those from the other species, with a slightly greater sequence similarity between *H. araneus* and *H. americanus*. However, considering the unequal sizes of these EST databases and the *H. araneus* transcriptome combined with a possibly different redundancy suggest that the data sets are hardly comparable by counting blast hits alone. To take the differences into account, a Markov Cluster Algorithm (MCL) clustering was applied to cluster transcripts into putative homologies. By clustering putatively related sequences into groups, the bias introduced through potentially different degrees of redundancy of transcript sequences is greatly reduced.

Counts of such overlapping clusters are a more objective statistic to compare datasets because they are less susceptible to bias when similar degrees of redundancy among data sources cannot be ensured. In the MCL clustering, 801 clusters of *H. araneus* show an overlap with *H. americanus* and 1036 clusters share sequence similarity with *P. cinctipes* (Fig. 5). In total, 1186 clusters contain sequences from all three species' libraries. This group of sequences may represent an assemblage of genes with putative core functions within decapod crustaceans. The large numbers of species-specific clusters are in agreement with the high genomic divergence as expectable from the fact that decapods comprise a set of highly diverse taxa (Martin et al., 2009). In total, for 5599 unique clusters found in *H. araneus* alone no corresponding analog could be identified in the annotation databases for the two other decapods. In comparison to the initial tBlastx analysis, the results show distinct differences. For example, the MCL cluster numbers indicate a slightly stronger similarity of *P. cinctipes*, not *H. americanus*, to *H. araneus*. All three species belong to the order of decapod crustaceans, yet differ in their classification to different infraorders (*P. cinctipes*: Anomura; *H. araneus*: Brachyura; *H. americanus*: Astacidea). Even if the phylogenetic taxonomy is far from being completely understood, the closer relation of *H. araneus* and *P. cinctipes* demonstrated by the MCL clustering is supported by numerous morphological and molecular analyses. Phylogenetic studies proposed an Anomura and Brachyura clade and a more distant Astacidea clade (Scholtz and Richter, 1995; Ahyong and O'Meally, 2004; Tsang et al., 2008). However, the approach is considered as an initial effective method and more comprehensive analyses including multiple species are needed to demonstrate in how far the results of the performed library clustering are taking sequencing biases implicitly into consideration, and how interpretation in functional terms can be achieved.

In terms of sequence counts, 15,111 *H. americanus* ESTs (60%), 43,005 *P. cinctipes* ESTs (57%) and 7459 *H. araneus* transcripts (54%) turned out to be species-specific. In the common core of 1186 MCL-clusters, corresponding sequence counts were 5009 (*H. americanus*), 17,773 (*P. cinctipes*) and 3245 (*H. araneus*), respectively. It is observed that the mean cluster sizes of *P. cinctipes*-specific transcript sequences are significantly larger than those of *H. araneus*. This highlights that the MCL-clustering leads to cluster sizes roughly proportional to the size of the libraries, indicating that differences in, for example, redundancy or sequencing depths are considered by clustering in a plausible manner. Overlap estimates between transcriptomic libraries can be effectively computed by clustering to reduce the effects of extensive amounts of transcript variants or large genome expansions.

The derived clustering structure can be analyzed in more detail by relating the transcript sequences to a defined set of assumed universal homologies. For this, we used Core Eukaryotic Genes Mapping Approach (CEGMA) profiles to screen transcripts for universal eukaryotic functions using trpsblastn and an E-value cutoff of 1E^{-9} (Altschul et al., 1997; Windisch et al., 2012). A total of 961 hits of *H. araneus* transcripts within the CEGMA dataset including multiple hits to 377 unique CEGMA profiles were found, corresponding to a 82% CEGMA-hit coverage of the *H. araneus* library. From these, 278 were found in the MCL-derived core set of transcripts, and 57 CEGMA-profiles were located in the *H. araneus*-specific MCL clusters. This finding can be explained by an inappropriate clustering and/or limited library depths resulting in

Table 4
Comparative tBlastx analysis. Intercomparison between *Hyas araneus* transcript sequence dataset and EST libraries of two other crustacean species. Given are the total numbers of tested sequences for each species, the number of sequences with a blast hit in the comparative species/common core and the percentage of the respective total number of sequences.

Species	Total number of sequences	No. of blasthits in *H. araneus*	No. of blasthits in *P. cinctipes*	No. of blasthits in *H. americanus*	No. of blasthits in the common core
H. araneus	13,709		1154 (8.4%)	1851 (13.5%)	3245 (23.7%)
P. cinctipes	75,298	7468 (9.9%)		7034 (9.3%)	17,773 (23.6%)
H. americanus	25,185	1939 (7.7%)	3126 (12.4%)		5009 (19.9%)

Please cite this article as: Harms, L., et al., Characterization and analysis of a transcriptome from the boreal spider crab *Hyas araneus*, Comp. Biochem. Physiol., D (2013), http://dx.doi.org/10.1016/j.cbd.2013.09.004

insufficient assemblies of the non-*H. araneus* libraries. In terms of amounts of clusters, in the core set of transcript 321 clusters contained hits to CEGMA-profiles from *H. araneus* transcripts, with 115 clusters containing more than one hit. In these clusters, 28 had hits to more than one CEGMA-profile, with a maximum of 6 different CEGMA-profiles (multiplicity 6). This points to limitations of the clustering approach when combined with homology information derived from partly incomplete transcriptomic sequences from non-model organisms with model organism genome databases. This is further supported by the finding that cluster-size and multiplicity of CEGMA-hits weakly correlate ($p = 0.64$, Spearman rank). It should be noted that the non-*H. araneus* libraries also cover ~60% of the CEGMA-profiles within the core set of clusters, confirming that the overlapping clustering contains the majority of preserved core functions.

The GO enrichment analysis of the core set of annotated sequences of *H. araneus* revealed a variety of over-represented terms from the ontologies of 'Molecular Functions', 'Biological Processes' and 'Cellular Compounds', respectively, but only 3 under-represented terms from the GO category 'Cellular Components' (see supporting information Table A). A clear picture of categories associated with the common core that could be interpreted as a representative functional clustering (e.g. with housekeeping genes dominating) within the decapod crustaceans does not become obvious. However, we observed a majority of closely interrelated terms under the GO term 'nucleotide metabolic process' within the 'Biological Process' category. This finding deserves further critical analyses with respect to the influence of assembly quality as well as of transcriptome complexity, e.g. presence of splice variants, in general.

3.5. Special characteristic in the H. araneus transcriptome — hypothesis for heritable, anti-viral immunity

One striking observation in the *H. araneus* transcriptome was the large amount of sequences identified as reverse transcriptase (RT). A total of 56 transcripts with a significant blast hit (E-value of $\leq 1E^{-25}$; score ≥ 150) were identified as RT or RT-like sequences and thereby constitute about 0.8% of all annotated transcripts of the *H. araneus* transcriptome. RTs are used to generate cDNA and are typically found in retroviruses to integrate their RNA genomes into the host genome, resulting in a replication along with the host cell. However, sequences for RTs from retro-transposons, retro-viruses, or viral-like elements have been previously observed in the genome of insects (Terzian et al., 2001; Eickbush and Jamburuthugoda, 2008). Furthermore, the occurrence of viral and viral-like sequences in the DNA of insects and crustaceans was reported (Crochu et al., 2004; Tang and Lightner, 2006). In the genome of the black tiger prawn *Penaeus monodon* for example, non-infectious sequences of the *Penaeus stylirostris* densovirus have been found (Tang and Lightner, 2006). Based on these findings a hypothesis for a heritable, anti-viral immunity was proposed for crustaceans and insects (Flegel, 2009). According to the author, an integration of viral genome fragments into the host genome by host-derived RT and integrases (IN) could result in the generation of antisense mRNA sequences that are capable to suppress the replication of the virus. These antisense mRNAs provide protection by the RNA interference pathway, which has been already validated in shrimp (Robalino et al., 2005). Due to the variety and number of RTs of the *H. araneus* transcriptome an acute infection of the sampled animals seems unlikely, and the finding could indicate a viral recognition process similar to the pathogen-associated molecular pattern recognition system of the known innate immune system defense mechanisms of crustaceans (for review see Vasquez et al., 2009). To support the proposed viral recognition mechanism for crustaceans or at least decapod crustaceans, and concomitantly reduce the possibility that the RTs are an assemblage artifact of the *H. araneus* transcriptome as well as a contamination of viral RNA, we scanned the core-set of the MCL-clustering (see above) comprising sequences that are present in all three crustacean species' libraries for RT sequences. We found 45 RT-transcripts (significant blasthit: E-value of $\leq 1E^{-25}$; score ≥ 150) of *H. araneus* in the core set (1.4%), a nearly two-fold enrichment of RTs compared to the proportion of RTs of the total transcriptome (0.8%), which suggests that the proposed viral recognition mechanism is a common feature in crustaceans. For *H. americanus* and *P. cinctipes*, 49 and 70 RT-sequences could be found, respectively, which correspond to 0.9% and 0.4% of the sequences of the core set.

Consequently, the presence of several RT-sequences in the core-set of all three species makes the presence of an assemblage artifact unlikely and reduces the possibility for a contamination, but supports the presence of a viral recognition mechanism proposed for crustaceans. To further test the reliability of these findings we used the previously identified RT-sequences from *H. araneus* to co-locate the sequences in the common fruit fly *D. melanogaster* and the purple sea urchin *S. purpuratus* sequence libraries. *D. melanogaster* as a model organism with a fully sequenced genome provides an excellent basis for this hypothesis. It is further known that *D. melanogaster* comprises RT-sequences as well as virus like fragments in the genome (Kim et al., 1994; Nefedova et al., 2011). The *S. purpuratus* genome was chosen as marine outlier. The overlap of *H. araneus* RT-sequences with *D. melanogaster* revealed no RT-sequences of *D. melanogaster*, while for *S. purpuratus* 34 RT-sequences could be identified. The presence of overlapping RT-sequences with the *S. purpuratus* transcriptome indicates that the hypothesis proposed for crustaceans and insects possibly can be expanded to other invertebrates. The absence of homologous RT-sequences in the *D. melanogaster* library suggests thereby that the RT-sequences found in the *H. araneus* transcriptome, in the EST libraries of the other crustaceans as well as in the sea urchin library seem to be specific for marine species possibly indicating an adaptation to marine habitats and a different viral composition. Several sequences, identified as integrases and transposases of the *H. araneus* transcriptome further support the possible integration of viral fragments in the genome and thus the proposed heritable, anti-viral immunity.

The present study could demonstrate the occurrence of a variety of RT-sequences in different decapod crustaceans and thus support the hypothesis of an integration of viral genome fragments into the host genome by host-derived RT. Besides in insects and crustaceans our data indicate a possible presence of a similar mechanism in other invertebrates (sea urchin). Furthermore, the findings suggest that the identified RT-sequences are marine specific. Although, the presence of the transcribed sequences alone is insufficient to verify the hypothesis and further investigations of the genome for viral inserts are indispensible. However, the several findings presented here already support the hypothesis and should promote further studies.

4. Conclusion

In this study we characterized the transcriptome of the Arctic spider crab *H. araneus*. The use of normalized cDNA libraries with samples from different tissues, collected after animal exposure to a variety of different abiotic conditions, and a high-throughput GS FLX sequencing in combination with additional Illumina sequencing, resulted in high-quality reads. The reads were assembled to 20,479 transcripts, 35% of them were functionally annotated. Thus, the *H. araneus* transcriptomic data provides a solid basement for future expression profiling and genomic studies in this physiological model.

The transcripts will significantly enhance the still small amount of available sequence data for crustaceans. This is even more important in light of the expected high genomic diversity within the decapods, requiring additional genome projects besides the *Daphnia* genome. The proposed overlap estimates in terms of clusters of similar transcript sequences by MCL, adopted here on transcriptomic data for the first time, allowed to effectively compare non-model organism transcriptomic libraries. Since we were able to determine special features and homologies (e.g. RTs) even in preliminary transcriptomes of crustaceans and

Please cite this article as: Harms, L., et al., Characterization and analysis of a transcriptome from the boreal spider crab *Hyas araneus*, Comp. Biochem. Physiol., D (2013), http://dx.doi.org/10.1016/j.cbd.2013.09.004

other marine invertebrates (sea urchin), its general applicability as methodological framework has to be validated by similar questions of further organism groups.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.cbd.2013.09.004.

Acknowledgments

Sincere thanks go to the scientific divers of the Alfred Wegener Institute and especially Max Schwanitz for animal collection. We would further like to thank the Max Planck Institute of Molecular Genetics and Dr. Richard Reinhardt for cDNA library construction and 454 sequencing.

References

Ahyong, S.T., O'Meally, D., 2004. Phylogeny of the Decapoda Reptantia: resolution using three molecular loci and morphology. Raffles Bull. Zool. 52, 673–693.

Q10 Altschul, S.F., Madden, T.L., Schäfler, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-Blast: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-2402.

Carson, M., Falcon, S., Pages, H., Li, N., 2010. GO.db: a set of annotation maps describing the entire gene ontology. R Package version 2.7.1.

Chevreux, B., Wetter, T.S.S., 1999. Genome sequence assembly using trace signals and additional sequence information. Comput. Sci. Biol. 99, 45–56.

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

Consortium, G.O., 2008. The gene ontology project in 2008. Nucleic Acids Res. 36, D440–D444.

R Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Crochu, S., Cook, S., Attoui, H., Charrel, R.N., De Chesse, R., Belhouchet, M., Lemasson, J.J., de Micco, P., de Lamballerie, X., 2004. Sequences of flavivirus-related RNA viruses persist in DNA form integrated in the genome of *Aedes* spp. mosquitoes. J. Gen. Virol. 85, 1971–1980.

De Gregoris, T.B., Rupp, O., Klages, S., Knaust, F., Bekel, T., Kube, M., Burgess, J.G., Arnone, M.I., Goesmann, A., Reinhardt, R., Clare, A.S., 2011. Deep sequencing of naupliar-, cyprid- and adult-specific normalised Expressed Sequence Tag (EST) libraries of the acorn barnacle *Balanus amphitrite*. Biofouling 27, 367–374.

Eickbush, T.H., Jamburuthugoda, V.K., 2008. The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res. 134, 221–234.

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

Feldmeyer, B., Wheat, C.W., Krezdorn, N., Rotter, B., Pfenninger, M., 2011. Short read Illumina data for the de novo assembly of a non-model snail species transcriptome (*Radix balthica*, Basommatophora, Pulmonata), and a comparison of assembler performance. BMC Genomics 12, 317.

Flegel, T.W., 2009. Hypothesis for heritable, anti-viral immunity in crustaceans and insects. Biol. Direct 4, 32.

Gotz, S., Garcia-Gomez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talon, M., Dopazo, J., Conesa, A., 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36, 3420–3435.

Hayward, P., Ryland, J., 1990. The Marine Fauna of the British Isles and North-West Europe: Introduction and Protozoans to Arthropods. Clarendon Press, Oxford.

Hou, R., Bao, Z., Wang, S., Su, H., Li, Y., Du, H., Hu, J., Wang, S., Hu, X., 2011. Transcriptome sequencing and de novo analysis for yesso scallop (*Patinopecten yessoensis*) using 454 GS FLX. PLoS ONE 6, e21560.

Kim, A., Terzian, C., Santamaria, P., Pélisson, A., Prud'Homme, N., 1994. Retroviruses in invertebrates: the *gypsy* retrotransposon is apparently an infectious retrovirus of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. U. S. A. 91, 1285–1289.

Kumar, S., Blaxter, M.L., 2010. Comparing de novo assemblers for 454 transcriptome data. BMC Genomics 11, 571.

Lindquist, S., Craig, E.A., 1988. The heat-shock proteins. Annu. Rev. Genet. 22, 631–677.

Martin, J.A., Wang, Z., 2011. Next-generation transcriptome assembly. Nat. Rev. Genet. 12, 671–682.

Martin, J.W., Crandall, K.A., Felder, D.L., 2009. Decapod Crustacean Phylogenetics. CRC PressI Llc.

Nefedova, L.N., Mannanova, M.M., Kim, A.I., 2011. Integration specificity of LTR-retrotransposons and retroviruses in the *Drosophila melanogaster* genome. Virus Genes 42, 297–306.

Parra, G., Bradnam, K., Korf, I., 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23, 1061–1067.

Riesgo, A., Andrade, S.C., Sharma, P.P., Novo, M., Perez-Porro, A.R., Vahtera, V., Gonzalez, V.L., Kawauchi, G.Y., Giribet, G., 2012. Comparative description of ten transcriptomes of newly sequenced invertebrates and efficiency estimation of genomic sampling in non-model taxa. Front. Zool. 9, 33.

Robalino, J., Bartlett, T., Shepard, E., Prior, S., Jaramillo, G., Scura, E., Chapman, R.W., Gross, P.S., Browdy, C.L., Warr, G.W., 2005. Double-stranded RNA induces sequence-specific antiviral silencing in addition to nonspecific immunity in a marine shrimp: convergence of RNA interference and innate immunity in the invertebrate antiviral response? J. Virol. 79, 13561–13571.

Schiffer, M., Harms, L., Pörtner, H.O., Lucassen, M., Mark, F.C., Storch, D., 2012. Tolerance of *Hyas araneus* zoea I larvae to elevated seawater PCO_2 despite elevated metabolic costs. Mar. Biol. Q11

Scholtz, G., Richter, S., 1995. Phylogenetic systematics of the Reptantian Decapoda (Crustacea, Malacostraca). Zool. J. Linnean Soc. 113, 289–328.

Shi, Y., Yu, C., Gu, Z., Zhan, X., Wang, Y., Wang, A., 2013. Characterization of the pearl oyster (*Pinctada martensii*) mantle transcriptome unravels biomineralization genes. Mar. Biotechnol. 15, 175–187.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., Birol, I., 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19, 1117–1123.

Sperstad, S.V., Haug, T., Vasskog, T., Stensvag, K., 2009. Hyastatin, a glycine-rich multi-domain antimicrobial peptide isolated from the spider crab (*Hyas araneus*) hemocytes. Mol. Immunol. 46, 2604–2612.

Stillman, J.H., Teranishi, K.S., Tagmount, A., Lindquist, E.A., Brokstein, P.B., 2006. Construction and characterization of EST libraries from the porcelain crab, *Petrolisthes cinctipes*. Integr. Comp. Biol. 46, 919–930.

Tagmount, A., Wang, M., Lindquist, E., Tanaka, Y., Teranishi, K.S., Sunagawa, S., Wong, M., Stillman, J.H., 2010. The porcelain crab transcriptome and PCAD, the porcelain crab microarray and sequence database. PLoS One 5, e9327.

Tang, K.F.J., Lightner, D.V., 2006. Infectious hypodermal and hematopoietic necrosis virus (IHHNV)-related sequences in the genome of the black tiger prawn *Penaeus monodon* from Africa and Australia. Virus Res. 118, 185–191.

Terzian, C., Pélisson, A., Bucheton, A., 2001. Evolution and phylogeny of insect endogenous retroviruses. BMC Evol. Biol. 1.

Towle, D.W., Smith, C.M., 2006. Gene discovery in *Carcinus maenas* and *Homarus americanus* via expressed sequence tags. Integr. Comp. Biol. 46, 912–918.

Tsang, L.M., Ma, K.Y., Ahyong, S.T., Chan, T.-Y., Chu, K.H., 2008. Phylogeny of Decapoda using two nuclear protein-coding genes: origin and evolution of the Reptantia. Anglais 48, 359–368.

Vasquez, L., Alpuche, J., Maldonado, G., Agundis, C., Pereyra-Morales, A., Zenteno, E., 2009. Review: immunity mechanisms in crustaceans. Innate Immun. 15, 179–188.

Walther, K., Sartoris, F.-J., Bock, C., Pörtner, H.O., 2009. Impact of anthropogenic ocean acidification on thermal tolerance of the spider crab *Hyas araneus*. Biogeosciences 6.

Walther, K., Anger, K., Pörtner, H.O., 2010. Effects of ocean acidification and warming on the larval development of the spider crab *Hyas araneus* from different latitudes (54° vs. 79°N). Mar. Ecol. Prog. Ser. 417, 159–170.

Wheat, C.W., 2010. Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. Genetica 138, 433–451.

Windisch, H.S., Lucassen, M., Frickenhaus, S., 2012. Evolutionary force in confamiliar marine vertebrates of different temperature realms: adaptive trends in zoarcid fish transcriptomes. BMC Genomics 13, 549.

Zhang, X., Mao, Y., Huang, Z., Qu, M., Chen, J., Ding, S., Hong, J., Sun, T., 2012. Transcriptome analysis of the *Octopus vulgaris* central nervous system. PLoS ONE 7, e40320.

Zittier, Z.M.C., Hirse, T., Pörtner, H.-O., 2012. The synergistic effects of increasing temperature and CO_2 levels on activity capacity and acid–base balance in the spider crab, *Hyas araneus*. Mar. Biol. xx, xxx.