Discovery of novel hemocyanin-like genes in Metazoans

Elisa M. Costa-Paiva[1,2*], Carlos G. Schrago[1*], Christopher J. Coates[3] and Kenneth M. Halanych[2]

1. Departamento de Genética, Laboratório de Biologia Evolutiva Teórica e Aplicada, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brazil
2. Department of Biological Sciences, Molette Biology Laboratory for Environmental and Climate Change Studies, Auburn University, Auburn, AL, 36849, USA
3. Department of Biosciences, College of Science, Swansea University, Singleton Park, Swansea, Wales SA2 8PP, UK

*Author for Correspondence: elisapolychaeta@gmail.com; carlos.schrago@gmail.com

Running head: Hemocyanin-like genes in Metazoans

Abstract

Among animals, two major groups of oxygen-binding proteins are found: proteins that use iron to bind oxygen (hemoglobins and hemerythrins) and two non-homologous hemocyanins that use copper. Although arthropod and mollusc hemocyanins (herein designated HcA and HcM, respectively) bind oxygen in the same manner, they are distinct in their molecular structures. In order to better understand the range of natural variation in Hcs, we searched for Hcs in a diverse array of metazoan transcriptomes using bioinformatics tools to examine Hc evolutionary history and consequently revive the discussion about whether all metazoan Hcs shared a common origin with frequent losses, or, originated separately after the divergence of Lophotrochozoa and Ecdysozoa. We confirm that the distribution of Hc-like genes is more widespread than previously reported, including five putative novel HcM genes in two annelid species from Chaetopteridae. For HcA, 16 putative novel genes were retained, and the presence of HcA in 11 annelid species represent novel observations. Interestingly, Annelida is the lineage that presents the greatest repertoire of oxygen transport proteins reported to date, possessing all the main superfamily proteins, which could be explained partially by the immense variability of life styles and habitats. Work presented here contradicts the canonical view that hemocyanins are restricted to molluscs and arthropods, suggesting the occurrence of copper-based blood pigments in metazoans has been underestimated. Our results also support the idea of the presence of oxygen carrier Hcs being widespread across metazoans with an evolutionary history characterized by frequent losses.

Introduction

Four families of oxygen-binding/transporting proteins have been characterised into two major groups: iron-based including hemoglobins and hemerythrins, and copper-based including two non-homologous families of hemocyanins (Hcs) (Terwilliger et al. 1976; Burmester 2002; Coates and Decker, 2017). These proteins are inextricably linked to metazoan evolution as they are constrained by oxygen requirements of tissues, and therefore selection has presumably favored

proteins that can reversibly bind and transport oxygen (Terwilliger, 1998; Schmidt-Rhaesa 2007). Although these molecules can reversibly bind oxygen, their binding affinities and evolutionary origins differ and it is generally regarded that the diversity of blood pigments in animals has been underestimated (Martín-Durán et al, 2013; Koch et al, 2016, Costa-Paiva et al, 2017 A-B).

Hcs are extracellular, oligomeric proteins that bind molecular oxygen reversibly between two copper ions (CuA and CuB; Markl and Decker, 1992) in a side-on bridging arrangement. These megamolecules, which can be similar in size to viruses or ribosomes (up to 13 MDa), are found predominantly in the hemolymph of gastropod molluscs and numerous arthropods (Burmester, 2002; Coates and Decker, 2017). They are also characterized by their affinity for oxygen, which can vary from low to moderate, and are regulated by a variety of allosteric factors (e.g. urea, calcium) related to specific ecophysiological adaptations (Decker et al, 2007). Arthropod and mollusc Hcs, despite being given the same name due to the conserved dicupric active sites, are distinct in their sequence and structural compositions (Terwilliger, 1998; van Holde et al, 2001).

Hcs present in Arthropoda (HcA) are organized into hexamers, where each subunit has three separate domains: (I) 5 or 6 α-helices, (II) 4 α-helices grouped together wherein lies the dicopper active site, and (III) 7 antiparallel strands forming a β-barrel (Magnus et al., 1994). HcAs, are members of a protein superfamily that include (a) insect and crustacean phenoloxidases whose functions include sclerotization of the cuticle, wound healing, and humoral immune defense (Whitten and Coates, 2017); (b) hexamerins (HEX), metal-free proteins present in insects that do not bind oxygen, but are considered storage proteins associated with molting cycles or nutritional conditions (Burmester, 1999a); (c) copper-free pseudo-hemocyanins or cryptocyanins (pHc or CRY) are similar to hemocyanin but appear to act as storage proteins in the haemolymph (Burmester, 1999b); and (d) hexamerin receptors (reHEX), which are present in dipterans and are related to their own ligands (Burmester and Scheller, 1996). Although these proteins form a functionally diverse superfamily, primary structures have highly conserved core elements that allow their evolutionary history to be traced (Burmester, 2001).

79 Mollusc Hcs (HcM) consist of paralogous functional units derived from successive gene
80 duplication events (Markl, 2013). They are large proteins (up to 13 MDa) with numerous
81 polypeptide channels consisting of 10-12 or more peptides, and 7 or 8 functional units connected to
82 each other by bridging peptides (Van Holde et al., 2001). Each functional unit consists of two
83 domains: the tyrosinase or α-helical core domain where the active site is located, and the β-
84 sandwich domain (Cuff et al., 1998). Functionally, the α-domain in HcM corresponds to domain II
85 in HcA and the β-domain corresponds to domain III, respectively (Decker et al, 2007). There is high
86 similarity between HcA and HcM near the ligands of the CuB site, comprising a segment of about
87 50 residues or 10% of the polypeptide chain length (Drexel et al, 1987; Preaux et al, 1988). HcA
88 and HcM are structurally heterogeneous and represent two different classes of proteins, however the
89 similar segment around CuB indicates that both Hcs have evolved independently from a common
90 ancestral mononuclear copper protein (Markl and Decker, 1992).
91 Despite extensive studies on hemocyanins (Hcs) over the years (Burmester 2002, 2015;
92 Coates and Nairn, 2014), knowledge remains limited for animals except in molluscs and arthropods
93 where Hcs are broadly distributed (Burmester, 2001; Lieb et al, 2006; Markl, 2013; Kato et al,
94 2017). Isolated records of Hc in single species of Porifera (*Amphimedon queenslandica),*
95 Hemichordata (*Saccoglossus kowalevski*), and Ctenophora (*Mnemiopsis leidyi*) (Martín-Durán et al.
96 2013) have been reported. The presence of Hc-like proteins in the tunicate *Ciona intestinalis* with
97 putative phenoloxidase-like activity suggested that respiratory Hcs evolved from a phenoloxidase
98 (Immesberger and Burmester, 2004), in addition to the well-characterized inducible phenoloxidase
99 activity of HcAs and HcMs (reviewed by Coates and Nairn, 2014). Thus, given the absence of
100 functional data for non-bilaterian animals (i.e., ctenophore and sponge) and recent evidence
101 concerning greater Hc distribution in animals, we revive the discussion about whether all metazoan
102 Hcs shared a common origin with frequent losses (Martín-Durán et al. 2013), or, if Hcs originated
103 separately after the Lophotrochozoa and the Ecdysozoa diverged (van Holde et al, 2001). In
104 addition to metazoan records, similarities of HcAs with sequences in fungi (*Aspergilus niger*) have

also been reported, suggesting that the origin of Hcs (or type-III copper proteins more generally) occurred in Opisthokonta, followed by multiple independent loss events (Martín-Durán et al., 2013). However, evidence is scant to establish what the function of this ancestral protein might have been (Burmester, 2015), as well as large gaps in sampling across Metazoa. In order to understand the range of natural variation in Hcs, we searched for Hcs in a diverse array of metazoan transcriptomes using a phylogenetic approach to examine Hc evolutionary history in the context of animal phylogeny.

Methods

Sample collection

Information on species employed herein is provided in Table 1 and Appendix 1. Transcriptomes of these species were collected using a variety of techniques, which include intertidal sampling, dredge and box cores. Samples collected were preserved in RNALater or frozen at -80^{o}C.

Data collection & sequence assembly

Methods for RNA extraction, cDNA preparation and high-throughput sequencing generally followed Kocot et al. (2011) and Whelan et al. (2015). The total RNA was extracted from either whole animals (for small specimens) or from the body wall and coelomic region (for larger specimens). RNA purifications were performed after extraction using TRIzol (Invitrogen) or the RNeasy kit (Qiagen) with on-column DNase digestion, respectively. Reverse transcription used single stranded RNA template and the SMART cDNA Library Construction Kit (Clontech) with double stranded cDNA synthesis employing The Advantage 2 PCR system (Clontech). cDNA libraries were barcoded and sequenced using Illumina technology by The Genomic Services Lab at the Hudson Alpha Institute (Huntsville, Alabama, USA). Since sequencing was performed from 2012-2015, Paired End (PE) runs were of 100bp or 125bp lengths, utilizing either v3 or v4

chemistry on Illumina HiSeq 2000 or 2500 platforms (San Diego, California). CD Hit was employed to look for redundant sequences. In order to facilitate sequence assembly, paired-end transcriptome data were digitally normalized to an average k-mer coverage of 30 using normalize-by-median.py (Brown et al, 2012; McDonald and Brown, 2013) and assembled using Trinity r2013-02-25 with default settings (Grabherr et al, 2011).

Data mining and gene identification

Methods employed here were similar to those described in Costa-Paiva et al. (2017b). Two approaches were utilized to mine transcriptomic data from 179 metazoan species and two choanoflagellate species for putative Hc genes *in silico* (Appendix 1).

The first approach employed BLASTX (Altschul et al, 1990) at an e-value cutoff of 10^{-6} in order to compare each assembled transcriptome contig ('queries') to a protein database composed of 22 Hcs sequences from the National Center for Biotechnology (NCBI) database (Appendix 2) of at least 500 amino acid residues. The BLASTX approach assured that any transcriptome contig with a significant 'hit' to an Hc would be further evaluated in the pipeline. Then, initial contigs recovered from these BLASTX searches were utilized in a second set of BLASTX searches against the NCBI protein database (minimum e-value of 10^{-10}) and only top hits longer than 600 nucleotides retained and were considered as putative Hc genes.

A second approach processed the transcriptomic data from the same species (Appendix 1) through the Trinotate annotation pipeline (Grabherr et al, 2011), which utilizes a BLAST-based approach to provide, among others, GO annotation (The Gene Ontology Consortium, 2004). Transcripts annotated as Hcs, using the 10^{-5} e-value cutoff obtained by using BLASTX, were also considered putative Hc-like gene orthologs.

Contigs identified as putative Hc genes using the two approaches described above were subsequently translated into amino acids using TransDecoder with default settings (Haas et al, 2013). All translations were additionally subject to a Pfam domain evaluation using the EMBL-EBI

database with an e-value cutoff of 10^{-5}. Translations which returned Hc domains N, M or C or Hc beta associated to tyrosinases in Pfam and that were longer than 200 amino acids residues were retained for subsequent analyses. Transcripts passing the criteria described above were considered Hc genes (Table 1; Appendix 1).

Sequence alignment

Because Hcs have been treated as two distinct proteins (Terwilliger, 1998), two protein datasets were formed based on the Pfam domain evaluation results: (a) an HcM dataset and (b) an HcA dataset.

a) HcM dataset sequence alignment

The HcM dataset included eight mollusc sequences previously used as 'queries' (Appendix 2) and five new sequences from translated transcripts. As HcM consists of a series of functional units contain α and β domains, we opted to used partial sequences consisting in two functional units (2[α domain + β domain]) for each sequence, with the exception of *Mesochaetopterus* sequences which presented just one single functional unit each. The dataset was aligned with MAFFT using default "FFT-NS-2" algorithm (Katoh and Standley, 2013), followed by visual inspection and manual curation in order to remove spuriously aligned sequences based on similarity to the protein alignment as a whole. Subsequently, ends of aligned sequences were manually trimmed in Geneious 9.1.3 (Kearse et al, 2013) to exclude residues 5' of the putative start of a Tyr domain and 3' residues following the amino acid subsequent to the end of the second Hc domain. The resulting alignment was used for all subsequent analyses (Supplemental File 1 available online).

b) HcA dataset sequence alignment

The HcA dataset was formed using 40 arthropod Hc superfamily sequences (Burmester, 2001; Aguilera et al, 2013; Martín-Durán et al, 2013; Appendix 3) and a remaining 16 sequences

183 from translated transcripts presenting at least two of the three HcA domains I, II, and III. Dataset
184 were aligned with MAFFT using default “FFT-NS-2" algorithm (Katoh and Standley, 2013),
185 followed by visual inspection and manual curation in order to remove spuriously aligned sequences
186 based on similarity to the protein alignment as a whole. Subsequently, aligned sequences were
187 trimmed using trimAl (Capella-Gutiérrez et al, 2009) with a 90% gap threshold were also
188 performed in order to eliminate poorly aligned regions. The resulting alignment was used for all
189 subsequent analyses (Supplemental File 2 available online).
190
191 **Phylogenetic analysis**
192 For each dataset, ProtTest3.4 was applied to carry out statistical selection of best-fit models
193 of protein evolution for each dataset separately using the Akaike and Bayesian Information Criteria
194 (AIC and BIC, respectively) methods (Darriba et al, 2011).
195 Bayesian phylogenetic inference was performed with MrBayes 3.2.1 (Ronquist and
196 Huelsenbeck, 2003) for each database separately employing two independent MCMC runs. In each
197 run, four Metropolis-coupled chains were sampled every 500^{th} cycle for 10^7 generations. In order to
198 confirm if chains achieved stationary and determine an appropriate burn-in, we evaluated trace plots
199 of all MrBayes parameter output in Tracer v1.6 (Rambaut et al, 2014). The first 25% of samples
200 were discarded as burn-in and a majority rule consensus tree generated using MrBayes. Bayesian
201 posterior probabilities were used for assessing statistical support of each bipartition.
202
203 **Contamination screening**
204 We under took procedures to reduce the chance of false positives in our analysis, meaning
205 genes that are not homologous to Hcs or genes obtained from contamination. In most samples
206 included in our work, the total RNA was extracted from the body wall and coelomic region, which
207 excluded any possible contamination from food residues. Moreover, all the species included in our

analyses in which Hcs were found were prepared and sequenced separate from any mollusc or arthropod species, which makes cross-contamination highly unlikely.

Furthermore, for *in silico* analyses, we opted for a more conservative approach and employed very stringent e-values cutoff ($< 10^{-5}$). After translation, proteins from contigs identified as putative Hc genes were subsequently subject to a Pfam Domain evaluation as described above and a BLASTP (Altschul et al, 1990) search against NCBI protein database (minimum e-value of 10^{-10}). The top hits were Hc sequences from either mollusc (for new HcM sequences) or arthropod (for new HcA sequences) and this could be easily explained since sequences of Hcs from other metazoan species are still rare in NCBI.

Therefore, as we opted for a very conservative approach, we decided to not include two sequences identified as HcAs in the dataset, one from an annelid *Cossura longocirrata* and one from a cycliophor *Symbion americanus*. Although these sequences had an Hc Pfam domain and presented the six conservative histidine residues, they also showed high identity values (> 80%) to crustacean Hcs when submitted to BLASTP and MEGABLAST searches.

Results

a) HcM

Our bioinformatic pipeline (Figure 1) recovered a total of 18 unique protein sequences of mollusc Hc-like genes from 181 transcriptomes representing 15 metazoan phyla and two choanoflagellate species (Appendix 1). Following translation, Pfam domain evaluation, and removal of sequences with fewer than 200 amino acid residues, five putative novel HcM genes were retained from all taxa examined here, representing two annelid species from Chaetopteridae (Table 1, Supplemental File 1 available online). For both choanoflagellate species and all other metazoan phyla, we did not find any HcM gene (Appendix 1). Alignment of translated transcripts possessed 956 residue positions and partial sequences consisted of two functional units (2[α domain + β domain]) for each sequence, with the exception of *Mesochaetopterus* sequences which had one

functional unit each. New sequences were combined with eight publically available HcMs previously used as 'queries' (Appendix 2) to produce a final dataset of 13 HcM sequences (Figure 2; Supplemental File 1available online; Costa-Paiva, 2018a).

The Bayesian inference analysis (Figure 3) revealed two highly supported clades (Figure 3, p.p. = 0.99) even though other internal nodes were less resolved, which are often observed in gene genealogies (DeSalle, 2015). Aside from previous mollusc records (Markl, 2013), we found novel HcM genes in two annelid species (Table 1) both within Chaetopteridae. All five sequences included canonical functional units composed of α- and β-domains. The topology of the HcM gene tree did not mirror recent phylogenies of Mollusca and the relationship between Mollusca and Annelida (Kocot et al, 2011; Halanych, 2016; Kocot et al, 2017). We found a strongly supported sub-group (Figure 3, pink clade B, p.p. = 0.99) with two *Phyllochaetopterus* Hcs and HcM sequences from caudofoveats, cephalopods, and bivalves. However, the remaining *Phyllochaetopterus* sequence formed a clade with other *Mesochaetopteus* sequences (Figure 3, blue clade A, p.p. = 0.76), which is in a clade with other polyplacophorans, gastropods, and cephalopods.

Our data were based on available transcriptomes, and the absence of a certain gene when searching in transcriptomes does not necessarily mean that it is absent from the genome. Thus, we carefully addressed that we did not find any HcM genes in any other metazoan and choanoflagellates as expected (Appendix 1).

b) HcA

Our bioinformatic analyses for HcA (Figure 1) recovered a total of 137 unique protein sequences of arthropod Hc-like genes from 181 transcriptomes from 15 metazoan phyla and two choanoflagellate species (Table 1). After nucleic acid translation, Pfam domain evaluation, and removal of sequences with fewer than 200 amino acid residues, 16 putative novel genes representing at least two of the three HcA domains were retained, representing 16 individual species distributed across four phyla (Table 1). Concerning these 16 new putative HcA genes, just one

possessed all three canonical arthropod domains (I, II, and III), the annelid *Streblosoma hartmanae* (Table I). The remaining 15 sequences contained two domains (I and II or II and III). Domain II, the location of the dicopper active site with six histidine residues, was found in each species, therefore we included these records as putative HcA genes (Figure 4). In order to understand the relationship of these putative HcA genes with other members of the Hc superfamily, the alignment of translated transcripts possessing 359 residue positions were performed for a dataset containing these 16 putative HcA genes and 40 publically available sequences from arthropod Hc superfamily representatives (Burmester, 2001; Aguilera et al., 2013; Martín-Durán et al., 2013) (Supplemental File 2 available online; Costa-Paiva, 2018b).

The Bayesian inference analysis of HcA superfamily revealed five supported clades: A) a green clade (Figure 5, p.p. = 1) formed by hexapod hexamerin sequences; B) a blue clade (Figure 5, p.p. = 1) formed exclusively by crustaceans HcAs, cryptocyanins and pseudo-hemocyanins; C) an orange clade (Figure 5, p.p. = 1) formed by myriapod and chelicerate HcAs; D) a pink clade (Figure 5, p.p. = 1) formed by hexapod and crustacean prophenoloxidase sequences; and E) a yellow clade (Figure 5, p.p. = 1) formed by non-arthropod HcA sequences, including the 16 novel sequences from annelids, hemichordates, sponges, and ctenophores. The circled clade inside the yellow clade was formed exclusively by ctenophores HcAs sequences.

Discussion

Herein, we confirm that the distribution of Hc genes is more widespread than previously reported. Our results describe actively transcribed HcMs genes in two chaetopterid annelid species, and HcAs in 16 species distributed across four metazoan phyla (Table 1, Figure 6). Of the four phyla, HcAs in Ctenophora, Porifera, and Hemichordata were reported previously (Aguilera et al. 2013; Martín-Durán et al. 2013). Importantly, the presence of HcAs in Annelida represent new records. Our work is contrary to the traditional view that Hcs are restricted to Molluscs and Arthropods (Burmester, 2001; Markl, 2013), corroborating the recent findings of Martin-Duran et al

286 (2013) that the presence of Hcs, as well as other oxygen carrier molecules such as hemerythrins
287 (Costa-Paiva et al., 2017 B) and globins (Koch et al., 2016), are underestimated in animals.
288 Although our results cannot empirically prove that these newly discovered genes effectively
289 transport oxygen, we present evidence here that this function is entirely possible (e.g., orthology to
290 Hcs, PFAM structure, and GO ontology). Additionally, these genes were identified based on
291 sequence similarity to previously well-characterized Hcs under the assumption that sequence
292 similarity is generally indicative of function (Gabaldón and Huynen, 2004).
293 Previous studies (e.g., Martín-Durán et al., 2013) have demonstrated that the α domain
294 (tyrosinase domain) has a wide distribution across metazoan lineages, with the exception of
295 arthropods, which can be explained by the expansion and diversification of the Hc domain II in this
296 group of animals. However, the tyrosinase domain itself can play several roles in addition to
297 respiratory function, such as melanin biosynthesis (Sugumaran, 2002). Although respiratory
298 function requires the presence of both domains α and β, as we found in molluscs, the presence of
299 few functional units in chaetopterids could indicate another function besides oxygen transport for
300 these molecules in agreement with recent studies on the functional versatility of Hcs. Examples
301 include antimicrobial peptide production, host-symbiont dynamics, and Hc-derived phenoloxidase
302 activity (Zhuang et al., 2015; Coates and Nairn, 2014; Kremer et al., 2014).
303 In relation to HcAs, our data corroborate previous findings of HcAs in hemichordates,
304 sponges, and ctenophores (Martín-Durán et al, 2013). Moreover, not all of our newly discovered
305 genes possessed the three domains, but all of them were considered to be an HcA because they
306 possessed domain II, where the active site is located (Decker et al, 2007). The presence of domain II
307 spread across metazoan lineages and the presence of the same domain in amebozoans and in the
308 filamentous fungus *Aspergilus niger* (Martín-Durán et al, 2013) are likely to be an unikont
309 synapomorphy as suggested before. Despite previous suggestions that the domain I (*N*-domain) of
310 HcA can be used as a specific molecular signature of the Pan-arthropoda (Martín-Durán et al,

311 2013), our findings concerning the presence of domain I in annelids and ctenophores contradict this
312 idea (Figure 6).

313 Interestingly, Annelida is the lineage that presents the greatest repertoire of oxygen transport
314 proteins reported to date, possessing all the main superfamily proteins: Hrs, Hbs, HcAs, and HcMs
315 (Rouse and Pleijel, 2001; Costa-Paiva et al, 2017a). This fact could be explained by their ancient
316 origin and the early radiation of this group, and also by the great diversification of life forms and
317 habitats which leads to a great variety of oxygen absorption and transport strategies inside the body
318 (Schumway, 1979; Rouse & Pleijel, 2001). Furthermore, the high diversity of oxygen-binding
319 proteins can be found even in the same lineage within annelids, as observed in chaetopterids. Based
320 on our data, a single individual of *Mesochaetopterus alipes* can actively transcribe Hrs (Costa-Paiva
321 et al, 2017a) and mollusc-like Hcs. Other annelids also presented more than one family of oxygen-
322 binding protein, for example some species of Terebellidae, Opheliidae and Sipuncula (Bailly et al,
323 2008; Liu et al, 2013). Such organisms may simultaneously express more than one protein or may
324 have different protein expression in different parts of the body (Bailly et al, 2008).

325 Although this unexpected large repertoire of oxygen-binding proteins in an annelid could be
326 partially explained by the need to transport oxygen molecules, secondary functional specializations
327 could also have been an important trigger for the diversification of these molecules throughout their
328 evolutionary history in this group. There are records of Hrs molecules involved in other functions
329 besides oxygen loading, such as in the storage of iron atoms, detoxification of heavy metals, and
330 metabolic pathways related to the innate immunity of some species of annelids (e.g., *Theromyzon*
331 *tessulatum*, *Hirudo medicinalis*, and *Neanthes diversicolor*; Baert et al, 1992; Demuynck et al,
332 1993; Vergote et al, 2004; reviewed by Coates and Decker, 2017). Moreover, Hcs are known for
333 contributing in many ways to immune defenses, such as the inhibition of viral replication, precursor
334 of antimicrobial peptides, and the conformational switch to a phenoloxidase–like enzyme (Coates
335 and Nairn, 2014; Coates and Talbot, 2018).

Our results support the idea of the presence of Hcs being widespread across metazoans (Figure 6) with an evolutionary history characterized by frequent losses (Aguilera et al, 2013; Martín-Durán et al, 2013). Such losses are observed within several lineages of molluscs and arthropods. For example, the gastropod family Planorbidae, which lack HcM in their hemolymph and utilize extracellular hemoglobin for oxygen transport (Ochiai et al, 1989; Arndt and Santoro, 1998), probably due to hemoglobin's higher affinity for oxygen than the ancestral Hc (Lieb et al, 2001). The same can be found in crustacean lineages, as branchiopods, ostracods, copepods, cirripeds, and decapods that lost HcA and presented hemoglobin for handling oxygen (Terwilliger and Ryan, 2001).

The revised distribution in expression of HcM and HcA genes across metazoans could be explained by the differences in physio-chemical properties of the oxygen binding domains and the life histories of disparate animal lineages. Obtaining functional data on these newly discovered Hc genes is needed to evaluate the significance of their widespread occurrence in metazoans, and oxygen-binding/transport proteins in general.

Acknowledgements

We thank Fernando Avila Queiroz for valuable help with the figures. EMCP was supported by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazil). Use of SkyNet computational resources at Auburn University is acknowledged. This work was funded by National Science Foundation grant DEB-1036537 to K. M. Halanych and Scott R. Santos and OCE-1155188 to K. M. Halanych. This is Molette Biology Laboratory contribution ## and Auburn University Marine Biology Program contribution ###. We are grateful to the three anonymous reviewers and editorial board for their candor.

References

Aguilera F., C. McDougall, and B. M. Degnan. 2013. Origin, evolution and classification of type-3 copper proteins: lineage-specific gene expansions and losses across the Metazoa. *BMC Evol. Biol.* **13**(1): 96.

Altschul S. F., W. Gish, W. Miller, E. W. Myers, D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.***215**: 403–410.

Arndt M. H., and M. M. Santoro. 1998. Structure of the extracellular hemoglobin of *Biomphalaria glabrata*. *Comp. Biochem. Physiol. B: Biochem. Mol. Biol.***119**(4): 667-675.

Baert J.L., M. Britel, P. Sautière, and J. Malécha. 1992. Ovohemerythrin, a major 14-kDa yolk protein distinct from vitellogenin in leech. *Eur. J. Biochem.*209: 563-569.

Bailly X., S. Vanin, C. Chabasse, K. Mizuguchi, and S. N. Vinogradov. 2008. A phylogenomic profile of hemerythrins, the nonheme diiron binding respiratory proteins. *BMC Evol. Biol.*8: 244.

Brown C. T., A. Howe, Q. Zhang, A. B. Pyrkosz, and T. H. Brom. 2012. A reference-free algorithm for computational normalization of shotgun sequencing data. arXiv:1203.4802 [q-bio.GN], available: https://arxiv.org/pdf/1203.4802.pdf. [2018, August 05].

Burmester T. 1999. Evolution and function of the insect hexamerins. *Eur. J. Entomol.***96**: 213-226. **(A)**

Burmester T. 1999. Identification, molecular cloning, and phylogenetic analysis of a non-respiratory pseudo-hemocyanin of *Homarus americanus*. *J. Biol. Chem.* **274**(19): 13217-13222. **(B)**

Burmester T. 2001. Molecular evolution of the arthropod hemocyanin superfamily. *Mol. Biol. Evol.* **18**(2): 184-195.

Burmester T. 2002. Origin and evolution of arthropod hemocyanins and related proteins. *J. Comp. Physiol. B.***172**(2): 95–107.

Burmester T. 2015. Evolution of respiratory proteins across the Pancrustacea. *Integr. Comp. Biol.***55**(5): 765-770.

Burmester T., and K. Schellen. 1996. Common origin of arthropod tyrosinase, arthropod hemocyanin, insect hexamerin, and dipteran arylphorin receptor. *J. Mol. Evol.***42**(6): 713-728.

Capella-Gutiérrez S., J. M. Silla-Martínez, and T. Gabaldón. 2009. TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, ***25***(15): 1972-1973.

Coates C. J. and H. Decker. 2017. Immunological properties of oxygen-transport proteins: hemoglobin, hemocyanin and hemerythrin. *Cell. Mol. Life Sci.**74***(2): 293-317.

Coates C. J. and J. Nairn. 2014. Diverse immune functions of hemocyanins. *Dev. Comp. Immunol.***45**(1): 43-55.

Coates C. J. and J. Talbot. 2018. Hemocyanin-derived phenoloxidase reaction products display anti-infective properties. *Dev. Comp. Immunol.*86: 47-51.

Costa-Paiva E. M. 2018. Mollusc hemocyanin dataset [online]. Figshare file DOI: 10.6084/m9.figshare.5301730 [2018, May 08]. (A)

Costa-Paiva E. M. 2018. Arthropod hemocyanin superfamily dataset [online]. Figshare file DOI: 10.6084/m9.figshare.5301691) [2018, May 08]. (B)

Costa-Paiva E. M., N. V. Whelan, D. S. Waits, S. R. Santos, C. G. Schrago, and K. M. Halanych. 2017. Discovery and evolution of novel hemerythrin genes in annelid worms. *BMC Evol. Biol.* **17**: 85-96. (A)

Costa-Paiva E. M., C. G. Schrago, and K. M. Halanych. 2017. Broad phylogenetic occurrence of the oxygen-binding hemerythrins in bilaterians. *Genome Biol. Evol.*,9(10): 2580-2591. (B)

Cuff M., K. Miller, K. van Holde, and W. Hendrickson. 1998. Crystal structure of a functional unit from Octopus hemocyanin. *J. Mol. Biol.* **278**: 855–870.

Darriba D., G. L. Taboada, R. Doallo, and D. Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*. **27**(8): 1164-1165.

Decker H., N. Hellmann, E. Jaenicke, B. Lieb, U. Meissner, and J. Markl. 2007. Recent progress in hemocyanin research. *Integr. Comp. Biol.* **47**(4): 631–644.

Demuynck S., K. W. Li, R. van der Schors, and N. Dhainaut-Courtois. 1993. Amino acid sequence of the small cadmium-binding protein (MP-II) from *Nereis diversicolor* (Annelida, Polychaeta) – evidence for a myohemerythrin structure. *Eur. J. Biochem.*217: 151–156.

DeSalle R. 2015. Can single protein and protein family phylogenies be resolved better? *J. Phylogenetics Evol. Biol.* **3**: e116.

Drexel R., S. Siegmund, H. J. Schneider, B. Linzen, C. Gielens, G. Préaux, R. Lontie, J. Kellermann, and F. Lottspeich. 1987. Complete amino-acid sequence of a functional unit from a molluscan hemocyanin (*Helix pomatia*). *Biol. Chem.* **368**(1): 617-636.

Gene Ontology Consortium. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**(1): D258-D261.

Grabherr M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnol.***29**: 644–652.

Haas B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P.D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*.**8**(8): 1494-1512.

Halanych K. M. 2016. How our view of animal phylogeny was reshaped by molecular approaches: lessons learned. *Org. Divers. Evol.***16**(2): 319-328.

Immesberger A., and T. Burmester. 2004. Putative phenoloxidases in the tunicate *Ciona intestinalis* and the origin of the arthropod hemocyanin superfamily. *J. Comp. Physiol. B*.**174**(2): 169-180.

Kato S., T. Matsui, C. Gatsogiannis, and Y. Tanaka. 2018. Molluscan hemocyanin: structure, evolution, and physiology. *Biophys. Rev.* **10**(2):191-202.

Katoh K., and D. M. Standley. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**(4): 772–780.

Kearse M., R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics;***28**(12): 1647–1649.

Koch J., J. Lüdemann, R. Spies, M. Last, C. T. Amemiya, and T. Burmester. 2016. Unusual diversity of myoglobins genes in the lungfish. *Mol. Biol. Evol.***33**(12): 3033-3041.

Kocot K. M., Cannon J.T., Todt C., Citarella M.R., Kohn A.B., Meyer A., Santos S.R., Schander C., Moroz L.L., Lieb B., et al. 2011. Phylogenomics reveals deep molluscan relationships. *Nature*; **477**: 452–456.

Kocot K. M., T. H. Struck, J. Merkel, D. S. Waits, C. Todt, P. M. Brannock, D. A. Weese, J. T. Cannon, L. L. Moroz, B. Lieb, et al. 2017. Phylogenomics of Lophotrochozoa with consideration of systematic error. *Syst. Biol.***66**(2): 256-282.

Kremer N., J. Schwartzman, R. Augustin, L. Zhou, E. G. Ruby, S. Hourdez, and M. J. McFall-Ngai. 2014. The dual nature of haemocyanin in the establishment and persistence of the squid–vibrio symbiosis. *Proc. R. Soc. Lond. B: Biol. Sci.**281***(1785): 20140504.

Lieb B., B. Altenhein, J. Markl, A. Vincent, E. van Olden, K. E. van Holde, and K. I. Miller. 2001. Structures of two molluscan hemocyanin genes: significance for gene evolution. *Proc. Natl. Acad. Sci.***98**(8): 4546-4551.

Liu Y., C. Li, X. Su, M. Wang, Y. Li, Y. Li, and T. Li. 2013. Cloning and characterization of hemerythrin gene from Sipuncula Phascolosoma esculenta. *Genes Genom.*35(1), 95-100.

Magnus K., B. Hazes, H. Ton-That, C. Bonaventura, J. Bonaventura, and W. Hol. 1994. Crystallographic analysis of oxygenated and deoxygenated states of arthropod hemocyanin shows unusual differences. *Proteins;***19**: 302–309.

Markl J. 2013. Evolution of molluscan hemocyanin structures. *Biochim. Biophys. Acta (BBA)-Proteins and Proteomics*;**1834**(9): 1840-1852.

Markl J., and H. Decker. 1992. Molecular structure of the arthropod hemocyanins. Pp. 325-376 in *Advances in Comparative & Environmental Physiology Vol 13 – Blood and Tissue Oxygen Carriers,* Mangum ChP, ed. Springer, Berlin, Heidelberg.

Martín-Durán J.M., A. De Mendoza, A. Sebé-Pedrós, I. Ruiz-Trillo, and A. Hejnol. 2013. A broad genomic survey reveals multiple origins and frequent losses in the evolution of respiratory hemerythrins and hemocyanins. *Genome Biol. Evol.***5**: 1435–1442.

McDonald E. and C. T. Brown. 2013. Khmer: working with big data in bioinformatics. *CoRR*, abs/1303.2223.

Ochiai T., Y. Enoki, and I. Usuki. 1989. Physicochemical properties of the extracellular hemoglobin from the planorbid snail, *Indoplanorbis exustus*. *Comp. Biochem. Physiol. B: Comp. Biochem.***93**(4): 935-940.

Préaux G., C. Gielens, R. Witters, and R. Lontie. 1988. The structure of molluscan haemocyanins and their homology with tyrosinases. *B. Soc. Chim. Belg.***97**(11-12): 1037-1044.

Rambaut A., M. A. Suchard, D. Xie, and A. J. Drummond. 2014. Tracer v1.6. Available from http://beast.bio.ed.ac.uk/Tracer. 2014. [2017, May 15]

Ronquist F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*;**19**(12): 1572-1574.

Rouse G. and F. Pleijel. 2001. *Polychaetes*. Oxford university press. 354 pp.

Schmidt-Rhaesa A. 2007. *The Evolution of Organs Systems*. Oxford University Press. 385 pp.

Sugumaran M. 2002. Comparative biochemistry of eumelanogenesis and the protective roles of phenoloxidase and melanin in insects. *Pigment Cell & Melanoma Research;***15**(1): 2-9.

Terwilliger N. B. 1998. Functional adaptations of oxygen-transport proteins. *J. Exp. Biol.***201**: 1085-1098.

Terwilliger N. B., and M. Ryan. 2001. Ontogeny of crustacean respiratory proteins. *Am. Zool.***41**(5): 1057-1067.

Terwilliger R. C., N. B. Terwilliger, and E. Schabtach. 1976. Comparison of chlorocruorin and annelid hemoglobin quaternary structures. *Comp. Biochem. Physiol. A.***55**(1): 51-55.

van Holde K. E., K. I. Miller, and H. Decker. 2001. Hemocyanins and Invertebrate Evolution. *J. Biol. Chem.***276**(19): 15563–66.

Vergote D., P. E. Sautière, F. Vandenbulcke, D. Vieau, G. Mitta, E. R. Macagno, and M. Salzet. 2004. Up-regulation of neurohemerythrin expression in the central nervous system of the medicinal leech, *Hirudo medicinalis*, following septic injury. *J. Biol. Chem.*279 (42): 43828–37.

Weigert A., and C. Bleidorn. 2016. Current status of annelid phylogeny. *Org. Divers. Evol.***16**(2): 345-362.

Whelan N. V., K. M. Kocot, L. L. Moroz, and K. M. Halanych. 2015. Error, signal, and the placement of Ctenophora sister to all other animals. *Proc. Natl. Acad. Sci. USA.***112**(18): 5773-5778.

Whitten M., and C. J. Coates. 2017. Re-evaluation of insect melanogenesis research: Views from the dark side. *Pigment Cell & Melanoma Research*;***30***(4): 386-401.

Zhuang J., C. J. Coates, H. Zhu, P. Zhu, Z. Wu, and L. Xie. 2015. Identification of candidate antimicrobial peptides derived from abalone hemocyanin. *Dev. Comp. Immunol.***49**(1), 96-102.

506 **Table 1.** List of taxa analyzed with novel genes and total number of contigs after
507 assembly. Number and type of putative Hc genes and accession numbers are also
508 provided.

Taxon	Total contigs number	Number and type of putative Hc genes	Accession number
METAZOA			
Porifera			
Kirkpatrickia variolosa (Kirkpatrick, 1907)	100,231	1 partial HcA Domains II + III	MF998096
Latrunculia apicalis Ridley & Dendy, 1886	76,210	1 partial HcA Domains II + III	MF998097
Ctenophora			
Coeloplana astericola Mortensen, 1927	222,614	1 partial HcA Domains I + II	MF998091
Mnemiopsis leidyi A. Agassiz, 1865	385,798	1 partial HcA Domains II + III	MF998101
Pleurobrachia bachei A. Agassiz, 1860	38,856	2 partial HcA Domains II + III	MF998107 MF998108
Hemichordata			
Harrimaniidae gen sp. (from Iceland)	230,054	1 partial HcA Domains II + III	MF998095
Annelida			
Dorydrilus michaelseni Piguet, 1913	136,096	1 partial HcA Domains II + III	MF998093
Eupolymnia nebulosa (Montagu, 1819)	139,021	1 partial HcA Domains II + III	MF998094
Lumbrineris perkinsi Carrera-Parra, 2001	144,648	1 partial HcA Domains II + III	MF998098
Mesochaetopterus alipes Monro, 1928	83,209	2 HcM	MF998099 MF998100
Paramphinome jeffreysii (McIntosh, 1868)	165,337	1 partial HcA Domains II + III	MF998102
Phyllochaetopterus prolifica Potts, 1914	193,836	3 HcM	MF998103 MF998104 MF998105
Pista macrolobata Hessle, 1917	126,764	1 partial HcA Domains II + III	MF998106
Streblosoma hartmanae Kritzler, 1971	108,080	1 HcA Domains I + II + III	MF998109
Stylodrilus heringianus Claparede, 1862	239,935	1 partial HcA Domains II + III	MF998110
Terebellides stroemii Sars, 1835	169,760	1 partial HcA Domains II + III	MF998112
Thelepus crispus Johnson, 1901	67,478	1 partial HcA Domains II + III	MF998113

Figure Legends

Figure 1 – Bioinformatics pipeline. Rounded rectangles represent input / output files, ovals represent software or scripts, and the hexagon represents a step which involving manual evaluation. Eight mollusc Hc sequences previously used as query sequences from Genbank (Appendix 2) were also included in the HcM dataset and 42 arthropod HcA superfamily protein sequences (Appendix 3) were also included in the HcA dataset for posterior analyses.

Figure 2 – HcM dataset partial alignment evidencing the active-site of six conserved histidine residues (arrows).

Figure 3 – Unrooted bayesian tree for HcM using MrBayes 3.2.1. A) Blue clade represents the remaining annelid sequences with polyplacophoran, gastropods, and cephalopod sequences. The circled blue clade represents the *Phyllochaetopterus* sequence with other *Mesochaetopteus* novel sequences. B) Pink clade represents two *Phyllochaetopterus* Hcs (circled clade) and HcM sequences from caudofoveats, cephalopods, and bivalves. The number after the name of each sequence indicates the GenBank accession numbers for each previously identified HcM gene and it is indicated in Appendix 2.

Figure 4 – HcA dataset partial alignment evidencing the active-site of six conserved histidine residues (arrows).

Figure 5 – Unrooted bayesian tree for HcA using MrBayes 3.2.1. A) green clade is formed by hexapod hexamerin sequences; B) blue clade by crustacean HcAs, cryptocyanins and pseudo-hemocyanins; C) orange clade included myriapod and chelicerate HcAs; D) pink clade prophenoloxidase sequences; and E) yellow clade is formed by non-arthropod HcA sequences, which include the 16 novel sequences from annelids, hemichordates, sponges, and ctenophores. The circled yellow clade represented a supported clade formed by ctenophores HcAs sequences. The number after the name of each sequence indicates the GenBank accession numbers for each previously identified HcA superfamily gene and it is indicated in Appendix 3.

Figure 6 - Hypothesized relationships among metazoan phyla derived from recent phylogenomic studies. Red rectangles represent HcM records and blue rectangles represent HcA records. The domain structure of the Hc protein found is demonstrated alongside each respective taxon.

Appendix

Appendix 1 - List of all taxa analyzed and total number of contigs after assembly. For underlined taxa, number and type of putative Hc genes and accession numbers are also provided.

Appendix 2 - Queries used to search the assembled translated transcriptomes. All HcM sequences were also included in the dataset previous to the alignment.

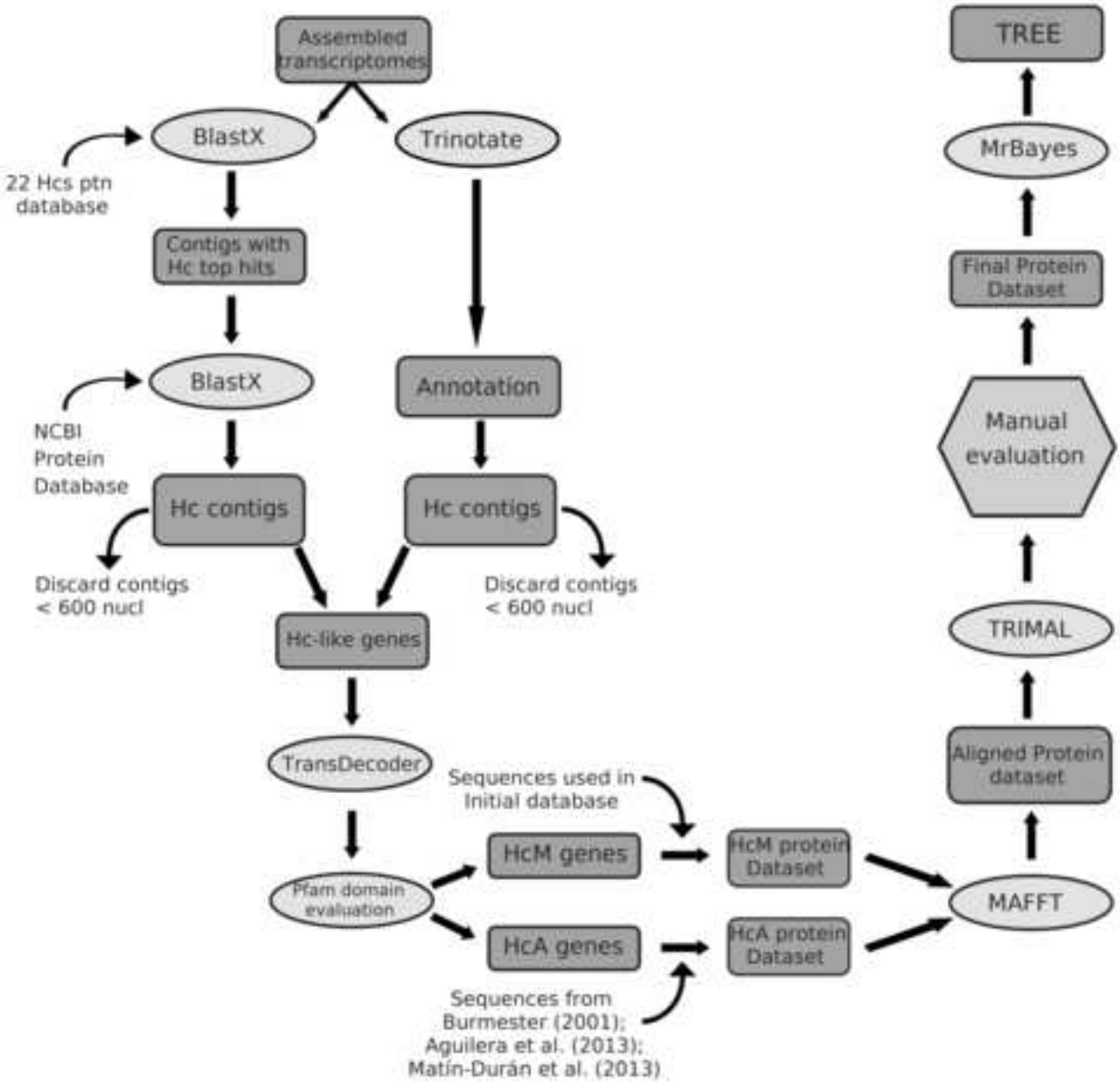
Appendix 3 - HcA superfamily protein sequences used in Burmester (2001), Aguilera et al. (2013), and Martín-Durán et al. (2013) with genes accession numbers for each species.

Supplemental Files

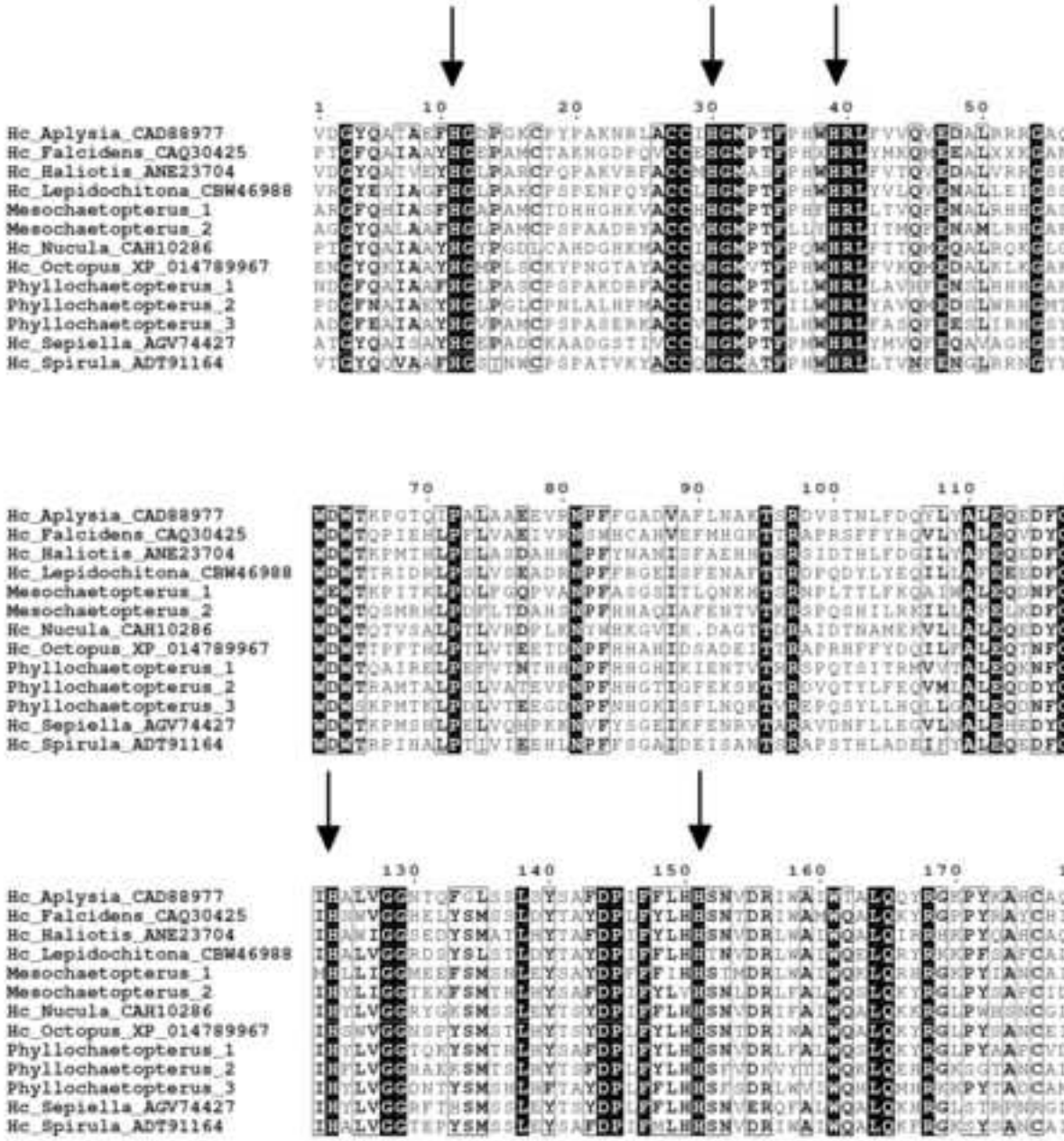
Supplemental File 1 - The amino acid alignment for HcM used in analyses.

Supplemental File 2 - The amino acid alignment for HcA used in analyses.

Figure

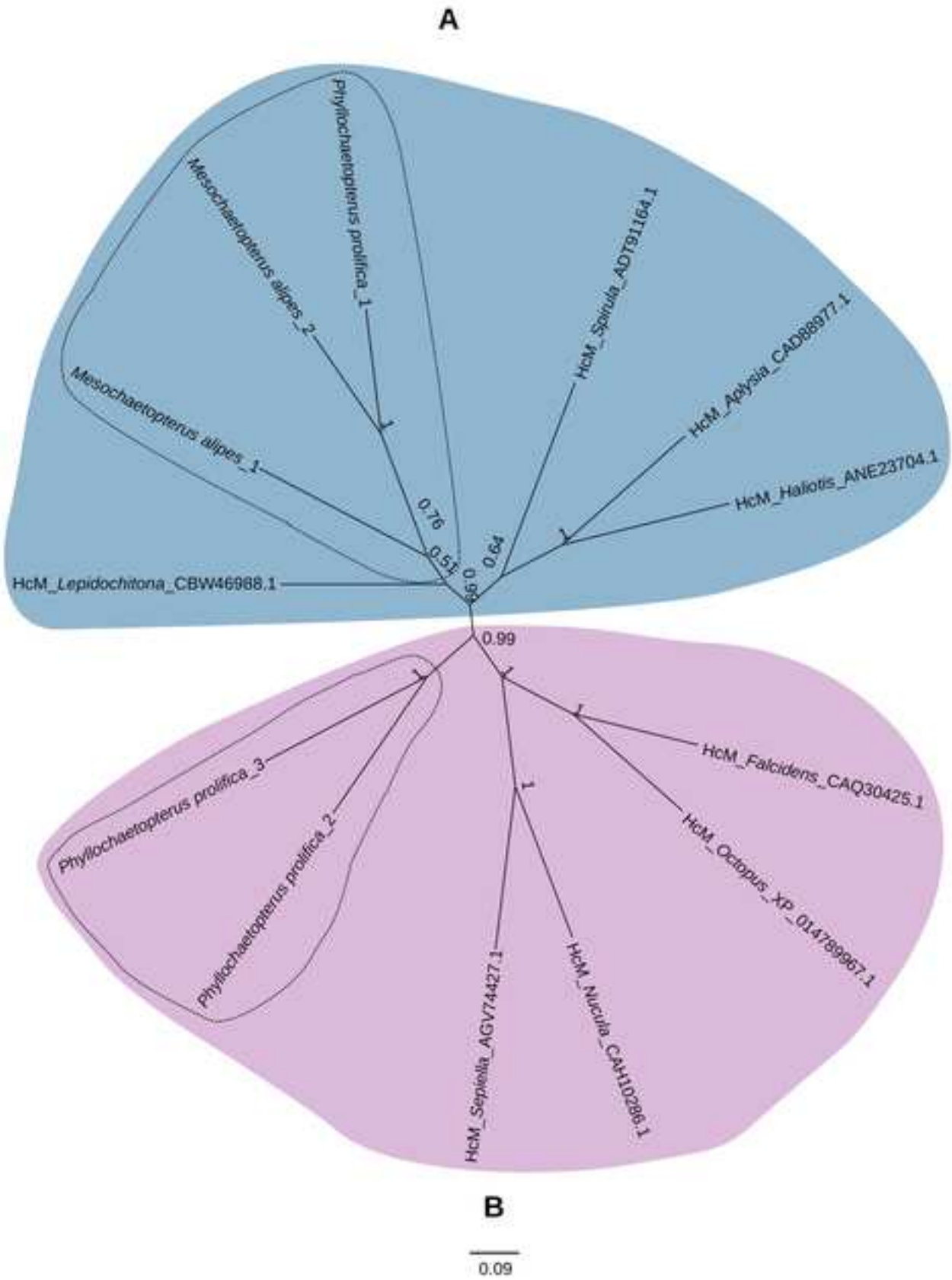

Assembled transcriptomes
BlastX
Trinotate
22 Hcs ptn database
Contigs with Hc top hits
BlastX
Annotation
NCBI Protein Database
Hc contigs
Hc contigs
Discard contigs < 600 nucl
Discard contigs < 600 nucl
Hc-like genes
TransDecoder
Sequences used in Initial database
HcM genes
HcM protein Dataset
Pfam domain evaluation
HcA genes
HcA protein Dataset
Sequences from Burmester (2001); Aguilera et al. (2013); Matín-Durán et al. (2013)
MAFFT
Aligned Protein dataset
TRIMAL
Manual evaluation
Final Protein Dataset
MrBayes
TREE

Figure


Hc_Aplysia_CAD88977
Hc_Falcidens_CAQ30425
Hc_Haliotis_ANE23704
Hc_Lepidochitona_CBW46988
Mesochaetopterus_1
Mesochaetopterus_2
Hc_Nucula_CAH10286
Hc_Octopus_XP_014789967
Phyllochaetopterus_1
Phyllochaetopterus_2
Phyllochaetopterus_3
Hc_Sepiella_AGV74427
Hc_Spirula_ADT91164

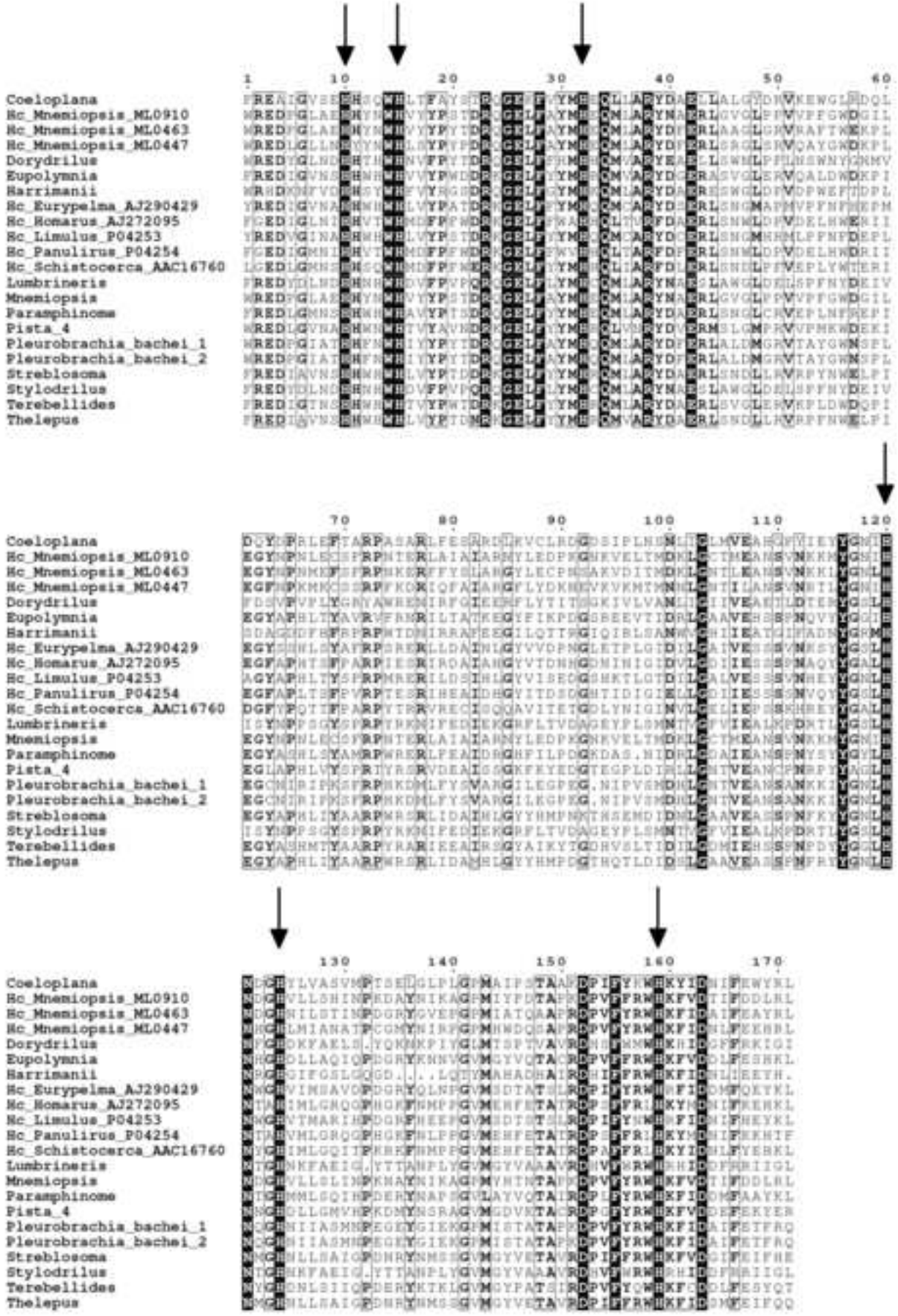
Figure

Click here to download Figure Figure3.jpg



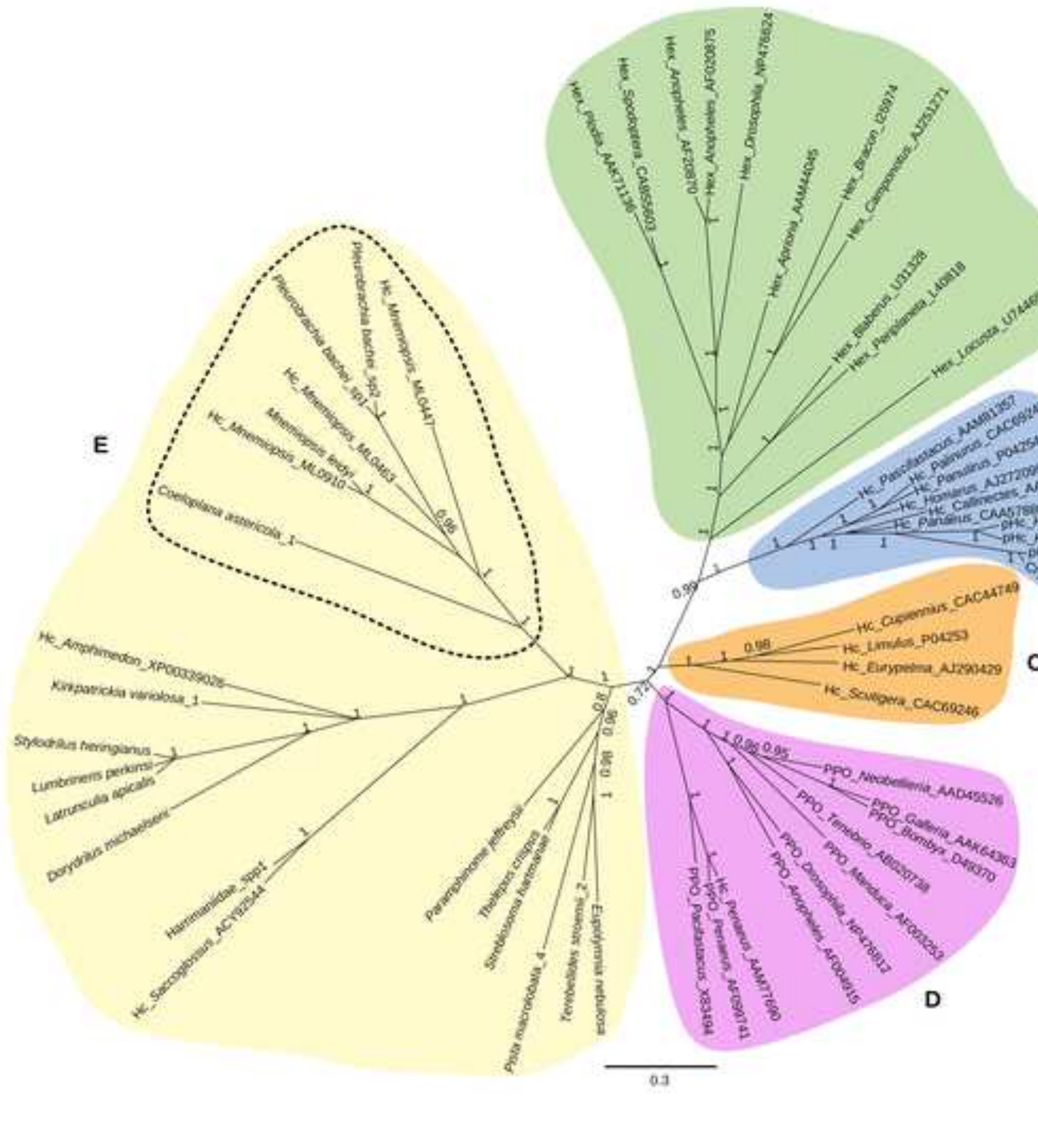
Figure

Click here to download Figure Figure4.jpg



Figure

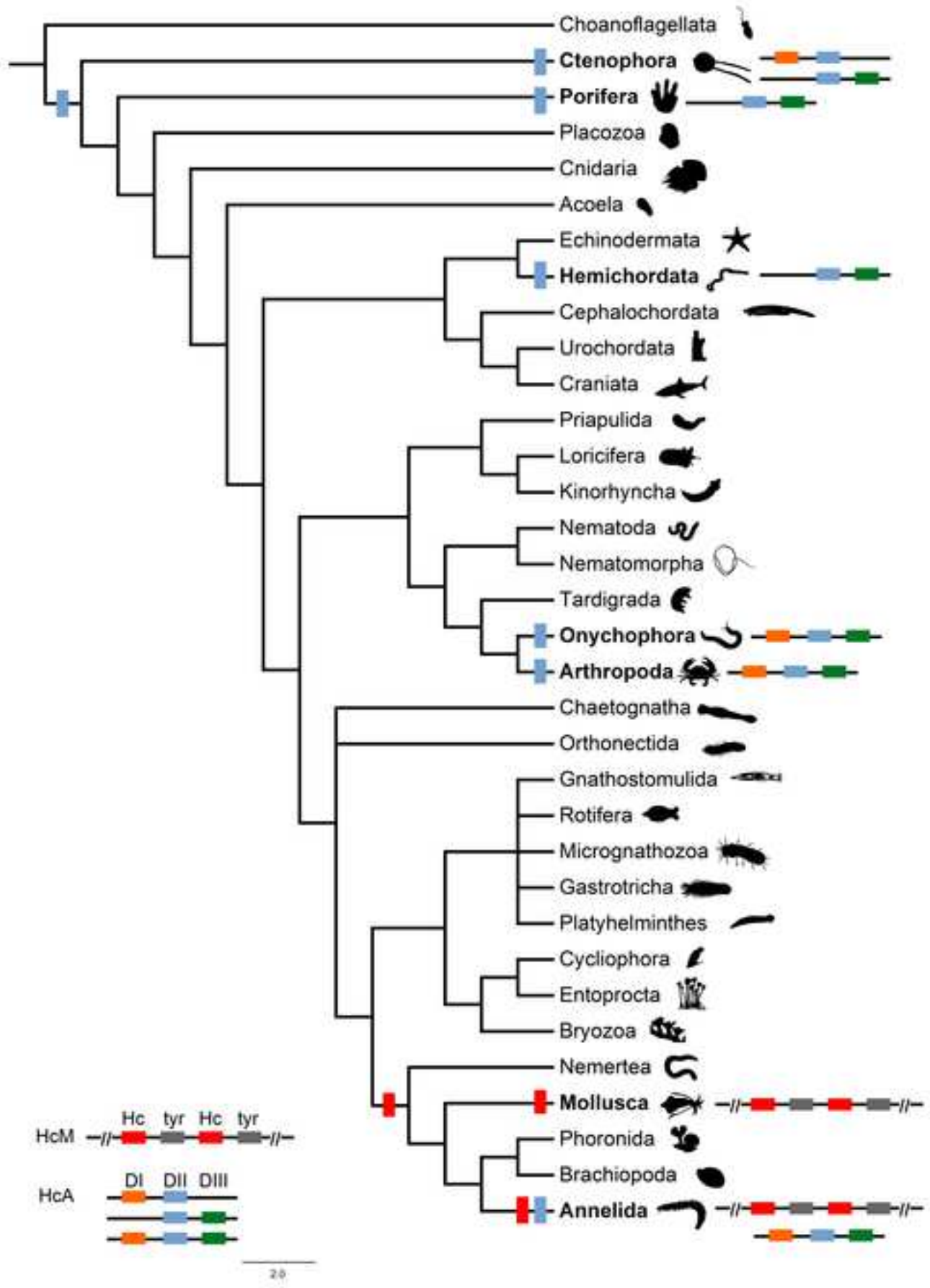
Cl



E
D
C
0.3
Hc_Amphimedon_XP003389026
Kirkpatrickia variolosa_1
Stylodrilus heringianus
Lumbrineris perkinsi
Lutrunculia apicalis
Dorydrilus michaelseni
Harrimaniidae_spp1
Hc_Saccoglossus_ACY92544
Coeloplana astericola_1
Hc_Cupiennius_CAC44749
Hc_Limulus_P04253
Hc_Eurypelma_AJ290429
Hc_Scutigera_CAC69246
PPO_Neobellieria_AAD45526
PPO_Galleria_AAK64363
PPO_Bombyx_D49370
PPO_Tenebrio_AB020738
PPO_Manduca_AF003253
PPO_Drosophila_NP476812
PPO_Anopheles_AF004915
Hc_Penaeus_AAM77690
PPO_Penaeus_AF099741
PPO_Pacifastacus_X83494
Hex_Locusta_U74469
Hex_Blaberus_U31328
Hex_Periplaneta_L40818
Hex_Apriona_AAM44045
Hex_Bracon_I25974
Hex_Drosophila_NP476624
Hex_Spodoptera_CAB55603

Figure

Click here to download Figure Figure6.jpg



Appendix 1. List of all taxa analyzed and total number of contigs after assembly. For undelined taxa, number and type of putative Hc genes and accession numbers are also provided.

Taxon	Total contigs number	Number and type of putative Hc genes	Accession number
CHOANOFLAGELATA			
Acanthoeca spectabilis W.Ellis, 1930	198,922		
Salpingoeca pyxidium Kent	202,399		
METAZOA			
Porifera			
Hyalonema populiferum Schulze, 1899	58,839		
Kirkpatrickia variolosa (Kirkpatrick, 1907)	100,231	1 partial HcA Domains II + III	MF998096
Latrunculia apicalis Ridley & Dendy, 1886	76,210	1 partial HcA Domains II + III	MF998097
Rossella fibulata Schulze & Kirkpatrick, 1910	40,103		
Sympagella nux Schmidt, 1870	85,237		
Ctenophora			
Beroe abyssicola Mortensen, 1927	83,798		
Coeloplana astericola Mortensen, 1927	222,614	1 partial HcA Domains I + II	MF998091
Dryodora glandiformis (Mertens, 1833)	101,598		
Euplokamis dunlapae Mills, 1987	321,550		
Mnemiopsis leidyi A. Agassiz, 1865	385,798	1 partial HcA Domains II + III	MF998101
Pleurobrachia bachei A. Agassiz, 1860	38,856	2 partial HcA Domains II + III	MF998107 MF998108
Vallicula multiformis Rankin, 1956	339,814		
Cnidaria			
Gersemia antarctica (Kukenthal, 1902)	20,023		
Periphylla periphylla (Peron & Lesueur, 1810)	212,658		
Staurozoa gen. sp.	45,023		
Echinodermata			
Apostichopus californicus (Stimpson, 1857)	134,640		
Astrotomma agassizii Lyman, 1875	156,062		
Labidiaster annulatus Sladen, 1889	108,871		

Labidiaster sp.	168,720		
Leptosynapta clarki Heding, 1928	242,126		
Hemichordata			
Balanoglossus aurantiaca Girard, 1853	143,815		
Cephalodiscus gracilis Harmer, 1905	57,139		
Cephalodiscus hodgsoni Ridewood, 1907	200,052		
Cephalodiscus nigrescens Lankester, 1905	11,565		
Harrimaniidae gen sp. (from Iceland)	230,054	1 partial HcA Domains II + III	MF998095
Harrimaniidae gen sp. (from Norway)	274,434		
Ptychodera bahamensis Spengel, 1893	115,310		
Rhabdopleura sp.	4,790		
Saccoglossus mereschkowskii Wagner, 1885	145,937		
Schizocardium brasiliense Spengel, 1893	101,493		
Stereobalanus canadensis Spengel, 1893	12,741		
Torquaratoridae gen. sp.	102,971		
Annelida			
Abarenicola pacifica Healy & Wells, 1959	94,376		
Aeolosoma sp.	190,647		
Aglaophamus verrilli (McIntosh, 1885)	118,343		
Alciopa sp.	233,051		
Amphisamytha galapagensis Zottoli, 1983	14,313		
Ancistrosyllis groenlandica McIntosh, 1878	94,924		
Andiorrhinus sp.	139,858		
Ankyrodrilus legaeus Holt, 1965	54,246		
Antarctodrilus proboscidea (Brinkhurst & Fulton, 1979)	49,656		
Aphelochaeta sp.	165,566		
Aphrodita japonica Marenzeller, 1879	120,025		
Arabella sp.	217,183		
Areco reco Righi, Ayres & Bittencourt, 1978	170,510		
Arenicola loveni Kinberg, 1866	27,028		
Arhynchite pugettensis Fisher, 1949	20,724		
Arichlidon gathofi Watson Russell, 2000	140,980		
Aricidea quadrilobata Webster & Benedict, 1887	81,139		

Armandia sp.	137,440			
Aspidosiphon laevis Quatrefages, 1865	168,072			
Auchenoplax crinita Ehlers, 1887	144,974			
Aulodrilus japonicus Yamaguchi, 1953	109,361			
Autolytus tuberculatus (Schmarda, 1861)	137,934			
Axiothella rubrocincta (Johnson, 1901)	107,215			
Bathydrilus rohdei (Jamieson, 1977)	226,538			
Bdellodrilus illuminatus (Moore, 1894)	67,562			
Bhawania goodei Webster, 1884	70,615			
Boccardia proboscidea Hartman, 1940	117,570			
Cambarincola holti Hoffman, 1963	46,015			
Capilloventer sp.	221,627			
Chaetogaster diaphanus (Gruithuisen, 1828)	128,034			
Chaetopterus variopedatus (Renier, 1804)	147,132			
Chaetacanthus magnificus (Grube, 1876)	95,443			
Chaetozone sp.	143,597			
Chloeia pinnata Moore, 1911	130,037			
Chone sp.	106,577			
Cirratulus spectabilis (Kinberg, 1866)	120,244			
Clymenella torquata (Leidy, 1855)	111,567			
Cossura longocirrata Webster & Benedict, 1887	75,079			
Crucigera zygophora (Johnson, 1901)	116,092			
Dichogaster saliens (Beddard 1893)	98,665			

Diopatra cuprea (Bosc, 1802)	138,779		
Dodecaceria pulchra Day, 1955	229,501		
Dorydrilus michaelseni Piguet, 1913	136,096	1 partial HcA Domains II + III	MF998093
Eteone sp.	41,912		
Enchytraeus crypticus Westheide & Graefe, 1992	161,842		
Eulalia myriacyclum (Schmarda, 1861)	110,762		
Eunice norvegica (Linnaeus, 1767)	122,784		
Euphrosine capensis Kinberg, 1857	72,220		
Eupolymnia nebulosa (Montagu, 1819)	139,021	1 partial HcA Domains II + III	MF998094
Galathowenia oculata (Zachs, 1923)	179,612		
Galeolaria caespitosa Lamarck, 1818	143,655		
Gatesona chaetophora (Bouché, 1972)	104,334		
Glossodrilus sp.	122,993		
Glycera americana Leidy, 1855	126,229		
Glycera dibranchiata Ehlers, 1868	101,455		
Glycinde armigera Moore, 1911	79,528		
Glyptonotobdella antarctica (Sawyer & White, 1969)	64,208		
Goniada brunnea Treadwell, 1906	89,398		
Harmothoe oculinarum (Storm, 1879)	94,991		
Hemipodia simplex (Grube, 1857)	55,653		
Hermenia verruculosa Grube, 1856	111,026		
Hermodice carunculata (Pallas, 1766)	110,813		
Heteromastus filiformis (Claparède, 1864)	148,196		
Histriobdella homari Beneden, 1858	143,130		
Idanthursus sp.	201,049		
Laetmonice producta Grube, 1876	73,530		

Leanira sp.	115,908		
Lumbrineris crassicephala Hartman, 1965	196,426		
Lumbrineris perkinsi Carrera-Parra, 2001	144,648	1 partial HcA Domains II + III	MF998098
Lysilla sp.	104,324		
Magelona berkeleyi Jones, 1971	50,123		
Marphysa sanguinea (Montagu, 1813)	110,924		
Melinna maculata Webster, 1879	135,712		
Mesochaetopterus alipes Monro, 1928	83,209	2 HcM	MF998099 MF998100
Microphthalmus similis Bobretzky, 1870	169,427		
Myxicola infundibulum (Montagu, 1808)	217,996		
Naineris laevigata (Grube, 1855)	218,272		
Neosabellaria cementarium (Moore, 1906)	82,479		
Nephasoma flagriferum (Selenka, 1885)	170,216		
Nephtys incisa Malmgren, 1865	188,338		
Nicolea macrobranchia (Schmarda, 1861)	53,572		
Nicomache venticola Blake & Hilbig, 1990	124,708		
Notomastus tenuis Moore, 1909	129,745		
Odontosyllis gibba Claparède, 1863	131,487		
Oenone fulgida (Savigny in Lamarck, 1818)	144,726		
Ophelina acuminata Örsted, 1843	81,846		
Ophiodromus pugettensis (Johnson, 1901)	92,341		
Ophryotrocha globopalpata Blake & Hilbig, 1990	129,450		
Owenia fusiformis Delle Chiaje, 1844	106,476		
Palola sp.	211,279		
Paralvinella palmiformis Desbruyères & Laubier, 1986	85,363		
Paramphinome jeffreysii (McIntosh, 1868)	165,337	1 partial HcA Domains II + III	MF998102
Pectinaria gouldii (Verrill, 1874)	92,091		
Phascolosoma agassizii Keferstein, 1866	87,403		
Pherecardia striata (Kinberg, 1857)	216,722		
Phyllochaetopterus prolifica Potts, 1914	193,836	3 HcM	MF998103 MF998104 MF998105

Pista macrolobata Hessle, 1917	126,764	1 partial HcA Domains II + III	MF998106
Prionospio dubia Day, 1961	119,949		
Sabaco elongatus (Verrill, 1873)	84,082		
Schizobranchia insignis Bush, 1905	102,002		
Sclerolinum brattstromi Webb, 1964	149,694		
Scolelepis squamata (Müller, 1806)	147,343		
Serpula vermicularis Linnaeus, 1767	151,097		
Siboglinum ekmani Jägersten, 1956	270,658		
Siboglinum fiordicum Webb, 1963	75,226		
Sphaerodorum papillifer Moore, 1909	52,411		
Spirobranchus kraussii (Baird, 1865)	167,761		
Sternaspis scutata (Ranzani, 1817)	10,634		
Sternaspis sp.	10,878		
Streblosoma hartmanae Kritzler, 1971	108,080	1 HcA Domains I + II + III	MF998109
Stylodrilus heringianus Claparede, 1862	239,935	1 partial HcA Domains II + III	MF998110
Stygocapitella subterranea Knöllner, 1934	74,556		
Syllis cf. *hyalina* Grube, 1863	106,283		
Terebellides stroemii Sars, 1835	169,760	1 partial HcA Domains II + III	MF998112
Tharyx kirkegaardi Blake, 1991	114,157		
Thelepus crispus Johnson, 1901	67,478	1 partial HcA Domains II + III	MF998113
Themiste pyroides (Chamberlin, 1919)	88,157		
Travisia brevis Moore, 1923	69,827		
Trypanosyllis sp.	167,501		
Brachiopoda			

Glottidia pyramidata (Stimpson, 1860)	131,562			
Hemithiris psittacea (Gmelin, 1791)	103,581			
Laqueus californicus (Koch, 1848)	133,086			
Macandrevia cranium (O. F. Müller, 1776)	9,695			
Novocrania anomala (O. F. Müller, 1776)	117,369			
Phoronida				
Phoronis psammophila Cori, 1889	193,702			
Phoronopsis harmeri Pixell, 1912	283,821			
Nemertea				
Malacobdella grossa (Müller, 1779)	79,313			
Paranemertes peregrina Coe, 1901	99,203			
Parborlasia corrugatus (McIntosh, 1876)	911,662			
Tubulanus polymorphus Renier, 1804	109,120			
Bryozoa				
Pectinatella magnifica (Leidy, 1851)	191,465			
Cycliophora				
Symbion americanus Obst, Funch & Kristensen, 2006	135,725			
Entoprocta				
Barentsia gracilis M. Sars, 1835	146,310			
Loxosoma pectinaricola Franzen, 1962	144,339			
Platyhelminthes				
Acipensericola petersoni Bullard, Snyder, Jensen & Overstreet, 2008	152,140			
Cardicola currani Bullard & Overstreet, 2004	86,962			
Cardicola palmeri Bullard & Overstreet, 2004	52,837			
Elaphrobates euzeti Bullard & Overstreet, 2003	118,013			
Elopicola sp.	64,384			
Hapalorhynchus sp.	42,863			
Myliobaticola richardheardi Bullard & Jensen, 2008	15,147			
Myliobaticola sp.	73,883			
Psettarium anthicum Bullard & Overstreet, 2006	39,616			
Sanguinicola sp.	145,041			

Selachohemecus olsoni Short, 1954	135,169			
Orthonectida				
Orthonectida gen. sp.	231,032			
Priapulida				
Priapulus sp.	50,034			

Appendix 2: Queries used to search the assembled translated transcriptomes. All HcM sequences were also included in the dataset previous to the alignment.

Taxon	Protein	GenBank accession number
Arthropoda		
Archispirosreptus gigas	Hc subunit type I	CCC55877.1
Cherax quadricarinatus	Hc	AFP23115.1
Cupiennius salei	Hc subunit 1	CAC44749.1
Cyamus scammoni	Hc	ABB59715.1
Limulus polyphemus	Hc II	NP_001301072.1
Macrobrachium nipponense	Hc	AHJ90473.1
Nebalia kensleyi	Hc	ACV33306.1
Penaeus monodon	Hc	AEB77775.1
Periplaneta americana	Hc subunit 1 precursor	CAR85701.1
Scutigera coleoptrata	Hc subunit A	CAC69246.1
Zootermopsis nevadensis	Hc	KDR21641.1
Mollusca		
Aplysia californica	Hc	CAD88977.1
Falcidens crossotus	Hc fgh, partial	CAQ30425.1
Haliotis rubra	Hc type 1	ANE23704.1
Lepidochitona cinerea	Hc, partial	CBW46988.1
Nucula nucleus	Hc isoform 1	CAH10286.1
Octopus bimaculoides	Hc units G and H-like, partial	XP_014789967
Sepiella maindroni	Hc	AGV74427.1
Spirula spirula	Hc, partial	ADT91164.1

Taxon	Protein	GenBank accession number
Porifera		
Amphimedon queenslandica	Phenoloxidase subunit 2-like	XP_003390261.1
Ctenophora		
Mnemiopsis leidyi	Hc	Contig ML0910 (Supplementary Data in Martín-Durán et al., 2013)
Hemichordata		
Saccoglossus kowalevski	Hc-like, partial	ACY92544

Appendix 3. HcA superfamily protein sequences used in Burmester (2001), Aguilera et al. (2013), and Martín-Durán et al. (2013) with genes accession numbers for each species.

Protein	Species	Accession number
Prophenoloxidase		
	Penaeus monodon	AAD45201
	Pacifastacus leniusculus	X83494
	Tenebrio molitor	AB020738
	Bombyx mori	BBA08368
	Manduca sexta	AAC05796
	Drosophila melanogaster	NP476812
	Neobellieria bullata	AAD45526
	Galleria mellonella	AAK64363
	Anopheles gambiae	AF004915
Hemocyanin		
Arthropoda		
	Eurypelma californicum	AJ290429
	Limulus polyphemus	P04253
	Callinectes sapidus	AAF64305
	Cupiennius salei	CAC44749
	Penaeus semisulcatus	AAM77690
	Pacifastacus leniusculus	AAM81357
	Panaeus vannamei	CAA57880
	Panulirus interruptus	P04254
	Homarus americanus	AJ272095

	Palinurus vulgaris	CAC69243
	Scutigera coleptrata	CAC69246
Non-Arthropoda		
	Amphimedon quenslandica	XP003390261
	Mnemiopsis leidyi	Contig ML0910 (Supplementary Data in Martín-Durán et al., 2013)
	Mnemiopsis leidyi	Contig ML0463 (Supplementary Data in Martín-Durán et al., 2013)
	Mnemiopsis leidyi	Contig ML0447 (Supplementary Data in Martín-Durán et al., 2013)
	Saccoglossus kowalevskii	ACY92544
Cryptocyanin		
	Cancer magister	AF091261
Pseudo-hemocyanin		
	Homarus americanus	CAB38042
	Homarus americanus	CAB38043
	Metacarcinus magister	AAD09762
Hexamerin		
	Locusta migratoria	U74469
	Drosophila melanogaster	NP476624
	Periplaneta americana	AAB09629
	Blaberus discoidalis	AAA74579
	Spodoptera litura	CAB55603
	Camponotus festinatus	AJ251271
	Apriona germari	AAM44045
	Plodia interpunctella	AAK71136

	Bracon hebetor	I25974
	Anopheles gambiae	AF020870
	Anopheles merus	AF020875