COMPARATIVE TRANSCRIPTOMICS OF AMPHINOMIDA (ANNELIDA)

An Undergraduate Research Scholars Thesis

by

ARIANNA P. BARTLETT

Submitted to the Undergraduate Research Scholars program at
Texas A&M University
in partial fulfillment of the requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by Research Advisor: Dr. Jessica Labonté
Dr. Elizabeth Borda

May 2018

Major: Marine Biology

TABLE OF CONTENTS

ABSTRACT

Comparative Transcriptomics of Amphinomida (Annelida)

Arianna P. Bartlett
Department of Marine Biology
Texas A&M University

Research Advisors: Drs. Jessica Labonté and Elizabeth Borda
Department of Marine Biology
Texas A&M University

Annelida is a diverse phylum that includes leeches, earthworms, polychaetes, and several model species like *Platynereis dumerilii* (clam worm), *Helobdella robusta* (leech), and *Capitella capitata* (polychaete worm), that are important in the fields of evolutionary developmental biology, neurobiology, ecology, evolution and phylogenomics. Our research seeks to identify and characterize gene annotations from transcriptome data collected from previously unevaluated amphinomid clades, Euphrosinidae (*Euphrosine capensis*), and Archinominae (*Chloeia pinnata*), and to supplement knowledge of Amphinominae (*Paramphinome jeffreysii* and *Hermodice carunculata*). Amphinomida remains understudied in terms of whole transcriptome analyses relative to more well studied annelids such as *Capitella teleta* and *Helobdella robusta*. To make transcriptomic comparisons, orthologous proteins within the amphinomids and a Lophotrochozoan database were identified and annotated for downstream analysis of phylogenomic relationships and exploration of biological pathways. Expanding transcriptomic analyses from previously unevaluated clades in Amphinomida may provide key evolutionary insights into the biological, physiological and morphological diversity of Annelida.

ACKNOWLEDGEMENTS

I would like to thank my research advisors, Dr. Jessica Labonté and Dr. Elizabeth Borda for their guidance and support throughout the course of this research. I also want to thank Dr. Ken Halanych at Auburn University for providing the genomic data for this study. Thank you to the predecessor of this project, Giovanni Madrigal, for always being optimistic and available for questions. As always, I am thankful for my family and partner for supporting me during this process.

CHAPTER I

INTRODUCTION

Background

The phylum Annelida consists of segmented worms that are globally distributed and inhabit a diversity of ecological habitats within aquatic systems (i.e. lakes, streams, coral and rocky reefs, deep sea chemosynthetic environments), and damp terrestrial environments (Rouse and Pleijel 2001). Annelida also exhibits a variety of morphological forms, perform diverse ecological functions (Rouse and Pleijel 2001), and is considered a key taxon in understanding the evolution of segmentation and the nervous system, and the developmental biology in the last common ancestor of Bilateria (Ferrier 2012; Weigert et al. 2014). Amphinomid fireworms and relatives of the Amphinomida exhibit interesting characteristics, such as the ability to regenerate lost segments (Ahrens et al. 2014; Weidhase et al. 2016) or using their brittle calcareous chaetae as a defense mechanism, causing mild to severe skin irritations on the skin of predators (Nakamura et al. 2008). Over 200 species of amphinomids have been described in approximately 25 genera, divided into two families, Euphrosinidae and Amphinomidae, with the latter further subdivided into subfamilies Archinominae and Amphinominae (Borda et al., 2012, 2015) (see also Figure 1).

To date, amphinomid research has mainly focused on taxonomy, higher-level phylogenetic relationships, population genetics, systematics (Borda et al. 2012, 2013, 2015; Ahrens et al. 2013; Schulze et al. 2017; Sun et al. 2017; Bonyadi-Naeini et al. 2017), and placement within the annelid tree of life (Struck et al. 2011; Weigert et al. 2014; Andrade et al. 2015). The evaluation of whole genome and/or transcriptomic data remains limited in

amphinomids and relatives, with attention only given to Amphinominae (Amphinomidae). Few transcriptomic analyses exist for Amphinomida, including *Hermodice carunculata* (Mehr et al. 2015; Verdes et al. 2018), *Eurythoe complanata*, and *Paramphinome jeffreysii* (Verdes et al. 2018). Studies which include amphinomid representatives as part of broader phylogenomic studies have only included Amphinomidae (Struck et al. 2011; Weigert et al. 2014; Andrade et al. 2015). Thus, we have limited knowledge regarding the genomic content of key biological function and pathways found in other recognized amphinomid lineages, Euphrosinidae and Archinominae (Borda et al. 2015).

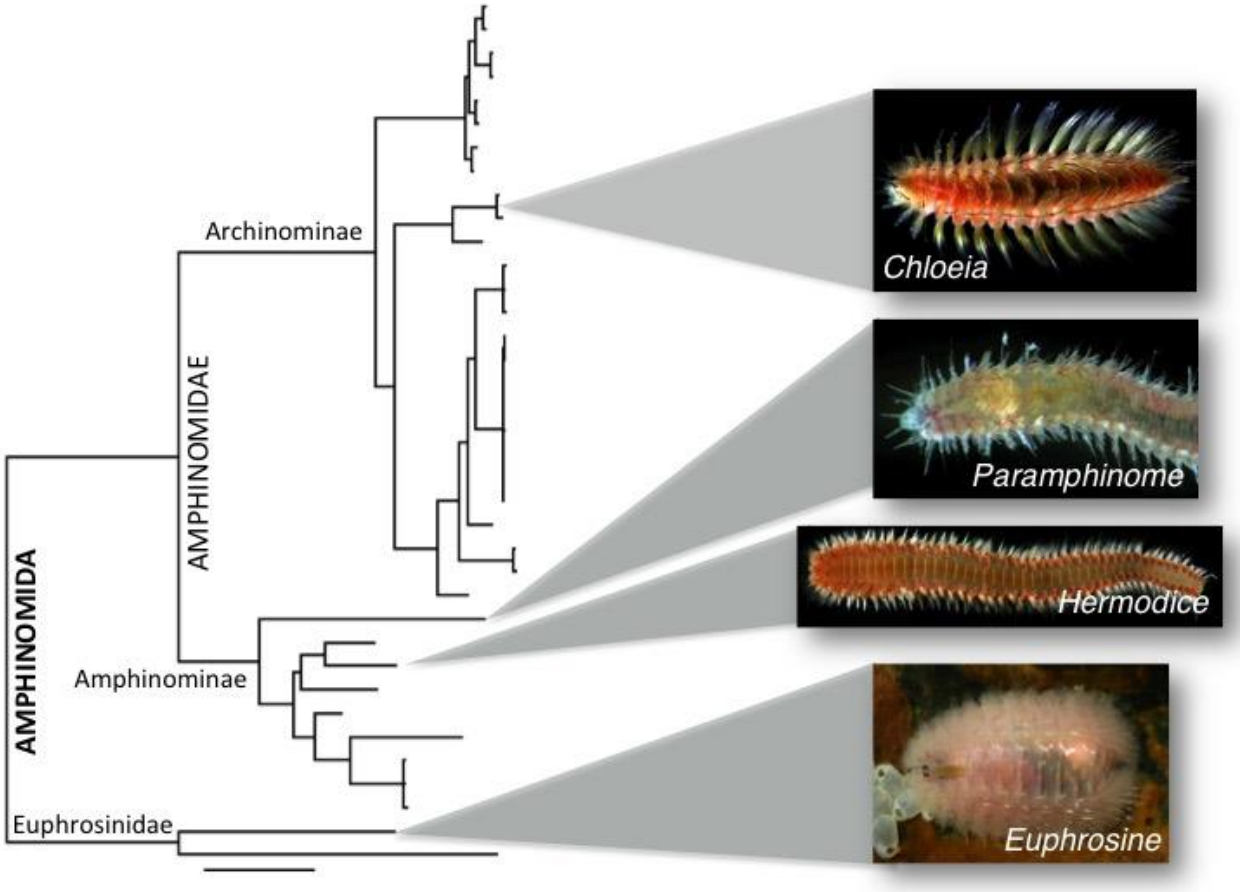


Figure 1. Phylogenetic tree of Amphinomida showing evolutionary relationships between the four Amphinomids. *Euphrosinidae* is sister to *Amphinomidae*, which includes two clades, Amphinominae (*Paramphinome jeffreysii* and *Hermodice carunculata*) and Archinominae (*Chloeia pinnata*). Image adapted from Borda et al. 2015.

Objectives and Hypotheses

The objectives of the study were to expand the evaluation of transcriptomic data to previously unsampled amphinomid lineages, Euphrosinidae and Archinominae, and to expand knowledge of amphinomid fireworms (Figure 1). Another objective was to provide the functional annotations of orthologous genes shared between the four amphinomids and a set of Lophotrochozoans.

CHAPTER II

MATERIALS AND METHODS

Identification of Coding Regions

The Halanych Lab at Auburn University was responsible for RNA extraction and sequencing, and shared the assembled contigs of *P. jeffreysii*, *H. carunculata*, *C. pinnata*, and *E. capensis*. Trinity and transdecoder were used to assemble the contigs and identify candidate open reading frames (ORFs), respectively (Haas et al. 2013). ORFs were compared to GenBank non-redundant database to find protein homologs (BLASTp, e-value 10^{-5}) (Altschul et al. 1990). To infer the quality of the assemblies, Benchmarking Universal Single-Copy Orthologs (BUSCO) was used to provide quantitative measures of the transcriptome completeness using *Caenorhabditis* (Nematoda) as the closest reference lineage (e-value 10^{-3}) (Simão et al. 2015).

Annotation and Gene Ontology

Using Blast2Go, a program that assigns gene ontology, protein function, and functional analysis (Götz et al. 2008), each ORF and corresponding BLAST results (top 10 hits) were uploaded for each transcriptome. Using the Blast2Go pipeline, the mapping and annotation functions retrieved gene ontology terms (GO terms) and assigned them to sequences using generic parameters. InterProScan (IPS), a tool that identifies candidate gene ontology (GO) using multiple online databases (Panther, Pfam, PIR, BlastProDom, etc.) was used to assign GO terms and IPS IDs. IPS results were merged with previously generated blast annotations to validate and add GO terms.

Identifying Core Orthologs in Amphinomids

Reciprocal BLASTp (e-value 10^{-10}) searches were performed serially between the proteins of each transcriptome dataset to identify the core proteins conserved among all four organisms. Due to the presence of duplicate sequences and similar sequences with varying lengths, deletions of duplicates were performed using Geneious (Kearse et al. 2012) and clustering of similar sequences based on global sequence identity (sequence identity cut-off=0.95, alignment bandwidth=20) was performed with cd-hit (Huang et al. 2010).

Identifying Core Orthologs Shared with Lophotrochozoa

To determine broader phylogenomic relationships between the core-orthologs of *E. capensis, C. pinnata, H. carunculata,* and *P. jeffreysii* to distantly related taxa, a database was created using Weigert et al. (2014) core-ortholog set "Lophotrochozoa_hmmr3" which contains 2,339 orthologous proteins, and a total of 14,626 sequences, from seven Lophotrochozoan species: *Helobdella robusta, Capitella teleta, Lottia gigantea* (owl limpet), *Schistosoma mansoni* (human blood fluke), *Daphnia pulex* (water flea), *Apis mellifera* (western honey bee), and *Caenorhabditis elegans* (nematode). Using BLASTp, comparisons were performed of the amphinomid orthologous proteins against the Lophotrochozoan orthologous set to identify shared proteins (e-value 10^{-10}). To identify homologous proteins, sequences with a sequence similarity less than 30% were removed. A reference dataset of the amphinomid and Lophotrochozoan orthologous set was generated and annotated with the top BLAST hit for the amphinomid proteins and marked for presence/absence between the four species.

To make comparisons of annotations in Blast2Go, the orthologous sets of each amphinomid were isolated from the larger transcriptome dataset. Statistics related to sequence similarity and top blast hit species were generated. A generic GO-Slim was made to visualize the

biological processes using GO terms for each amphinomid. An alignment was generated for one gene, 21889, to describe protein overlap and redundancy using Multiple Sequence Alignment by Log-Expectation (Edgar 2004).

CHAPTER III

RESULTS

Whole Transcriptome Assembly and Annotations

Transcriptomes were sequenced for four species of amphinomids, *Paramphinome jeffreysii* and *Hermodice carunculata* (Amphinominae), *Chloeia pinnata* (Archinominae), and *Euphrosine capensis* (Euphrosinidae). After assembly, there were a total of 165,337; 110,813; 130,037; and 72,220 contigs for *P. jeffreysii*, *H. carunculata*, *C. pinnata*, and *E. capensis*, respectively (Table 1). *P. jeffreysii* had the largest number of contigs and highest mean length followed by *C. pinnata*, *H. carunculata*, and lastly *E. capensis* (Table 1). A BUSCO analysis showed that the higher amount of contigs correlated with "completeness" of transcriptomes, with an estimated completeness of 36.2% in *P. jeffreysii*, 20.7% in *C. pinnata*, 9.1% in *H. carunculata*, and 8.2% in *E. capensis* (Table 1).

Table 1. Number of contigs and mean length among the amphinomid transcriptome data. BUSCO transcriptome completeness analysis using *Caenorhabditis* (Nematoda) as the reference lineage. Complete single-copy (Complete-S) and compete duplicated (Complete-D) BUSCOs are complete based on BUSCO scores and length alignment using a reference lineage. Fragmented BUSCOs have met the required score, but the range of length alignments to the BUSCO profile are not met. Missing BUSCOs are contigs without a significant match or scored too low to the reference lineage.

Amphinomid	Contigs	Mean Length	Complete-S (%)	Complete-D (%)	Fragmented (%)	Missing (%)
P. jeffreysii	165337	631	14.3	21.9	7.6	56.2
H. carunculata	110813	388	5.0	4.1	4.5	86.4
C. pinnata	130037	534	9.7	11.0	6.5	72.8
E. capensis	72220	376	5.0	3.2	4.1	87.7

Using the pipeline in Blast2Go, all the amphinomids had blast results and/or IPS hits for all proteins, except for *H. carunculata*, where around 11% of the proteins had no results (Figure

2). Although there are proteins which have BLAST, IPS, and mapping results, only the proteins which were successfully annotated are assigned GO terms. For example, a protein can have BLASTp, IPS, and mapping results but no annotation because candidate GOs retrieved from the mapping stage do not meet the score set by the Blast2Go annotation algorithm and associated generic parameters. The percentages of annotations compared to the total proteins are 53.6%, 36.1%, 51.3%, and 45.4% for *P. jeffreysii*, *H. carunculata*, *C. pinnata*, and *E. capensis*, respectively (Figure 2).

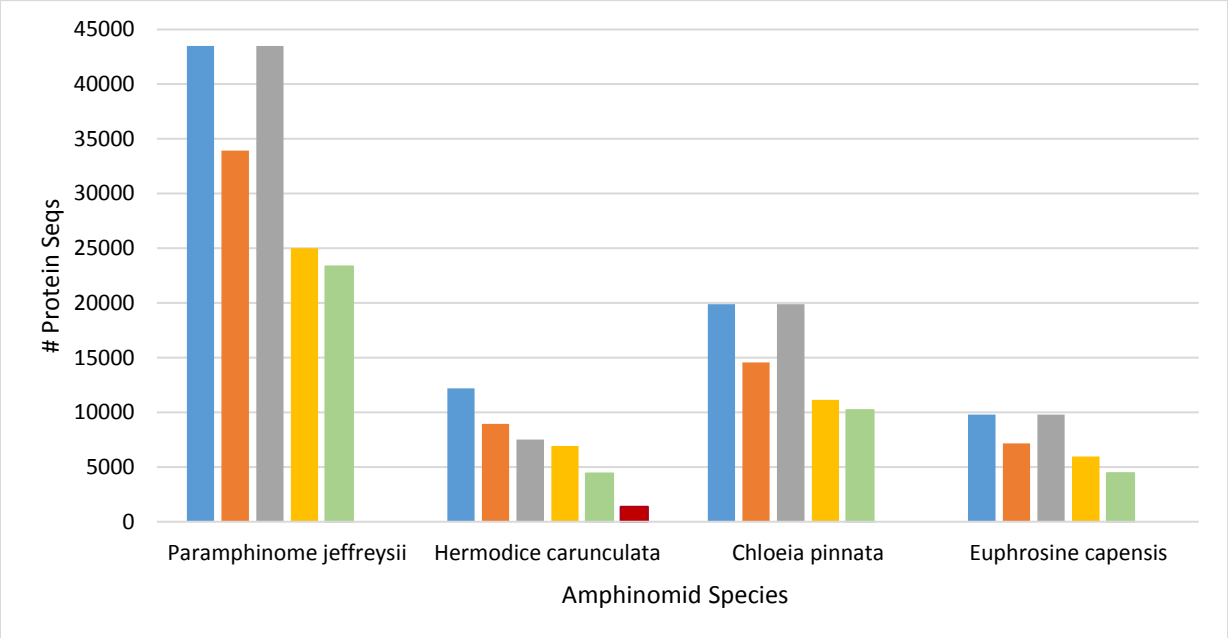


Figure 2. Gene ontology (GO) and annotation results of putative proteins using the Blast2Go pipeline. For comparison, total protein sequences (with identical sequences removed) are shown in blue. BLASTp (orange) and InterProScan (grey) were two databases are used to assign putative gene ontology. Blast2Go has functions including mapping (yellow) and annotation (green) for the recovery and assignment of GO terms. Proteins without hits to any database are shown in red.

IPS GO terms were added to existing blast annotation, Figure 3 shows the number of GO terms before and after merging, where for *P. jeffreysii*, *H. carunculata*, and *C. pinnata* the number of GOs decreased, but increased for *E. capensis* (Figure 3). Confirmed IPS GOs were

highest in *P. jeffreysii*, which had the highest initial IPS hits, followed by *C. pinnata* which also had the second highest initial IPS hits. *E. capensis* had a higher amount of initial IPS hits than *H. carunculata* (Figure 2), but the lowest amount of confirmed IPS after merging, with *H. carunculata* having the third highest confirmed ISP GOs (Figure 3). *P. jeffreysii* had the larger number of IPS GOs which were too general, followed by *C. pinnata*, *E. capensis*, and *H. carunculata*.

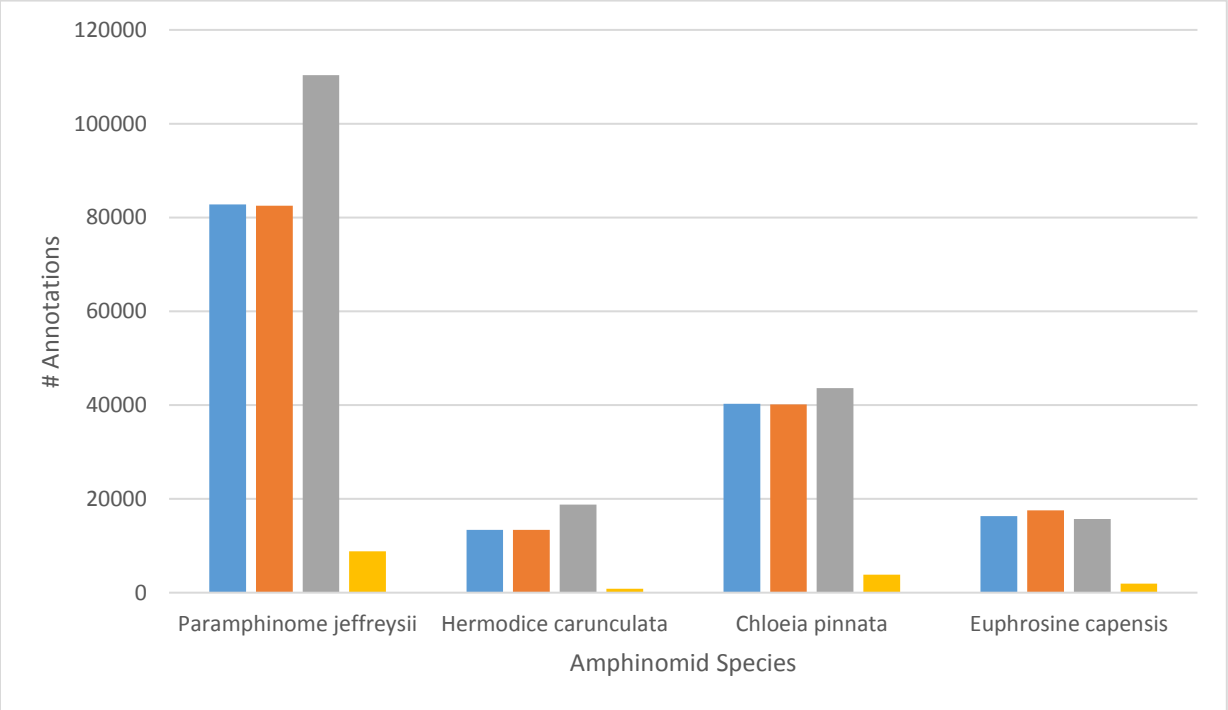


Figure 3. Annotations generated using InterProScan (IPS) were merged to add and validate existing GO terms generated by BLASTp. The figure shows the GOs before merging (blue), GOs after merging (orange), confirmed IPS GOs (grey), and IPS GOs which were too general (yellow).

Orthologous Gene Comparisons

Transdecoder identified the largest amount of putative proteins in P. *jeffreysii*, followed by *C. pinnata*, *H. carunculata*, and *E. capensis* (Table 2). After finding the shared orthologs within the amphinomids, the protein hits within each species were not uniform, the clustering

resulted in the most uniformity between *H. carunculata* and *E. capensis* within a 224-protein difference, but *C. pinnata* still had 3,633 proteins, and *P. jeffreysii* had 6,000 proteins (Table 2).

Table 2. The total putative proteins identified using Transdecoder before (Total Putative Proteins) and after (Identical Sequences Removed) removing identical sequences. The protein amounts of core orthologous proteins of *P. jeffreysii*, *H. carunculata*, *C. pinnata*, and *E. capensis*, and subsequent amounts of proteins after clustering similar sequences.

Amphinomid	Total Putative Proteins	Identical Sequences Removed	Core Orthologs	Clustered (Seq Identity Cutoff=0.95)
P. jeffreysii	*56278*	*43501*	10527	6000
H. carunculata	*13669*	*12186*	4509	2714
C. pinnata	*23822*	*19897*	5815	3633
E. capensis	*10478*	*9791*	3790	2490

These proteins were then compared against the Lophotrochozoan orthologous set which contained seven different Lophotrochozoans. The initial hits were narrowed down to 59,769; 18,833; 34,830; and 16,256 for *P. jeffreysii*, *H. carunculata*, *C. pinnata*, and *E. capensis*, respectively, after the removal of sequences with less than 30% sequence similarity, which resulted in a 30% reduction in the total amount of protein hits between the four amphinomids (Table 3). These amounts are larger than the 14,636 sequences in the Lophotrochozoan orthologous set, due to single amphinomid sequence within a species hit to multiple genes (i.e. *P. jeffreysii* sequence #84430 hit to Lophotrochozoan gene 23208 and 22706, see "Hits Including Redundancy to Lophotrochzoan Database*" in Table 3). Despite these redundant amounts of proteins, the actual number of proteins for each species is smaller, as shown in "Individual Protein Hits" (Table 3). Another redundancy creating phenomenon was when more than one protein hit to a single Lophotrochozoan gene, for example, Appendix 3 shows multiple different protein sequences from each amphinomid species for gene 21889 (39s ribosomal mitochondrial-like).

Table 3. A summation of the number of data generated by comparing the orthologous proteins in the amphinomids to the Lophotrochozoan orthologous set using BLASTp (14,636 sequences representing 2,339 genes among 7 taxa). The total hits were narrowed by excluding protein sequences with less than 30% sequence similarity.

Amphinomid	Total Hits	Homologous Hits (≥30% Seq Similarity)	Hits Including Redundancy to Lophotrochzoan Database	Individual Protein Hits	Hits to 2,339 Lophotrochozoan Gene List
P. jeffreysii	96419	59769	12138	2935	992
H. carunculata	22671	18833	3579	1200	825
C. pinnata	48329	34830	2986	1726	816
E. capensis	19967	16256	6620	1070	896

*Values include the phenomenon where individual protein sequences within one amphinomid species hit to multiple genes in the Lophotrochozoan database which includes multiple taxa.
**Shows the actual number of individual protein sequences per amphinomid species.

Total hits for each amphinomid to the Lophotrochozoan database were highest in *P. jeffreysii* with 992 proteins, followed by 896 proteins in *E. capensis*, though it had the smallest number of contigs, proteins, and the lowest transcriptome completeness (Table 1, Table 3). There were 1,026 overall amphinomid protein hits to the Lophotrochozoan gene list, of these, 764 proteins were shared between the amphinomids and Lophotrochozoans overall (Figure 4). A sample of the gene list, containing the first 10 genes with associated gene descriptions generated by the amphinomid top BLAST hits are provided, see Appendix 4.

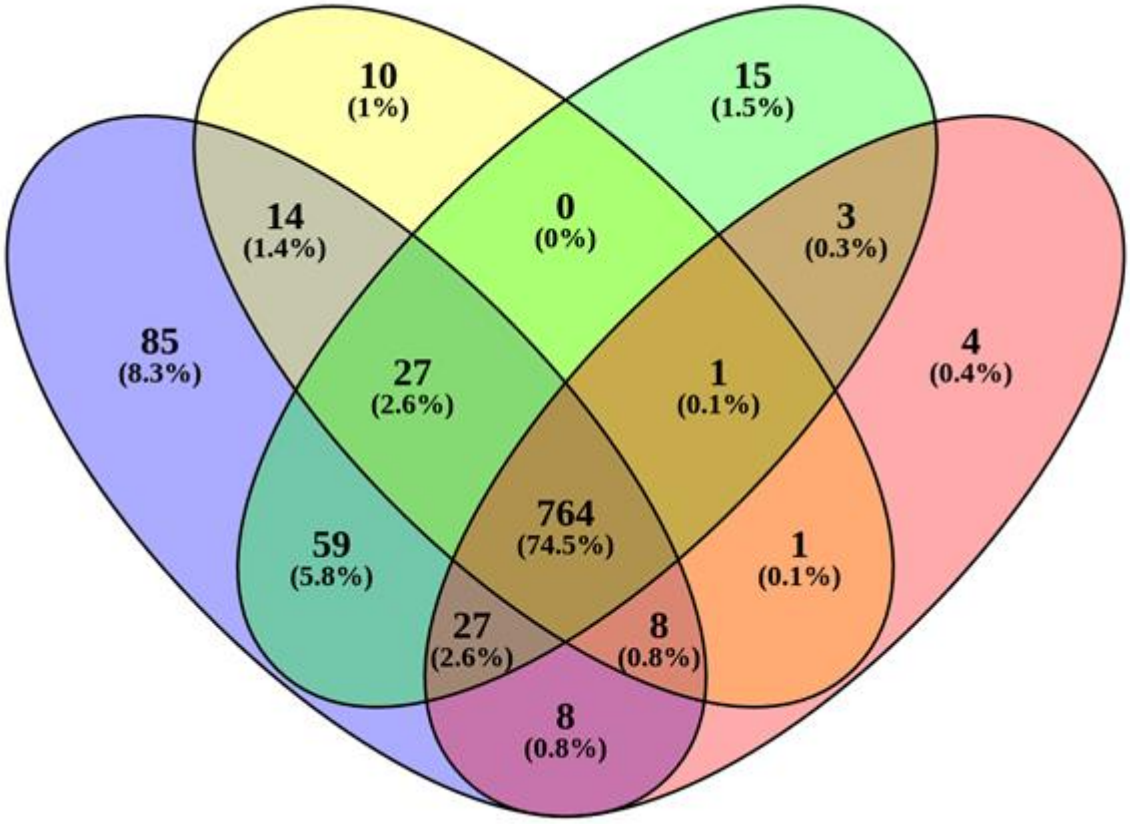


Figure 4. A comparison and visualization of the 1,026 amphinomid proteins, from *P. jeffreysii* (blue), *H. carunculata* (yellow), *C. pinnata* (green), and *E. capensis* (red) which were orthologous to the Lophotrochozoan database of 2,339 genes.

Orthologous Set Annotation

The orthologous set shared with the Lophotrochozoans produced a greater overall coverage of annotation results for the amphinomids compared to the whole transcriptomes (Figure 2, Figure 5). *H. carunculata* had the lowest percentage of annotations compared to the number of proteins, 72%, compared to the other amphinomids at 93.6%, 94.3%, and 90.9%, for *P. jeffreysii*, *C. pinnata*, and *E. capensis*, respectively (Figure 5).

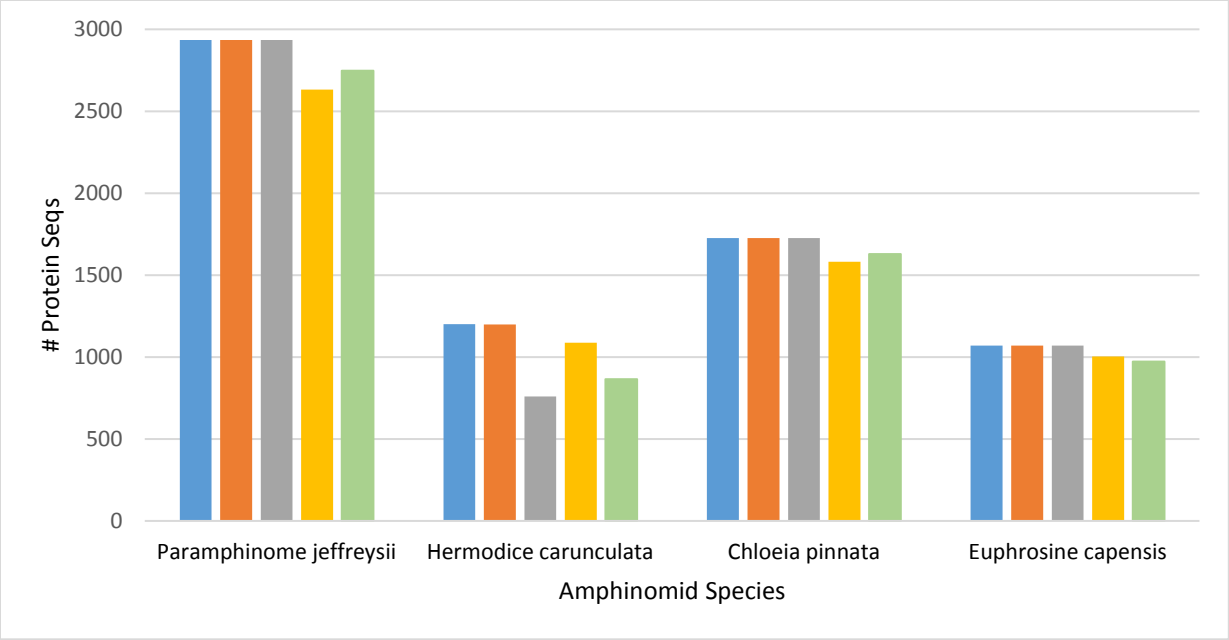
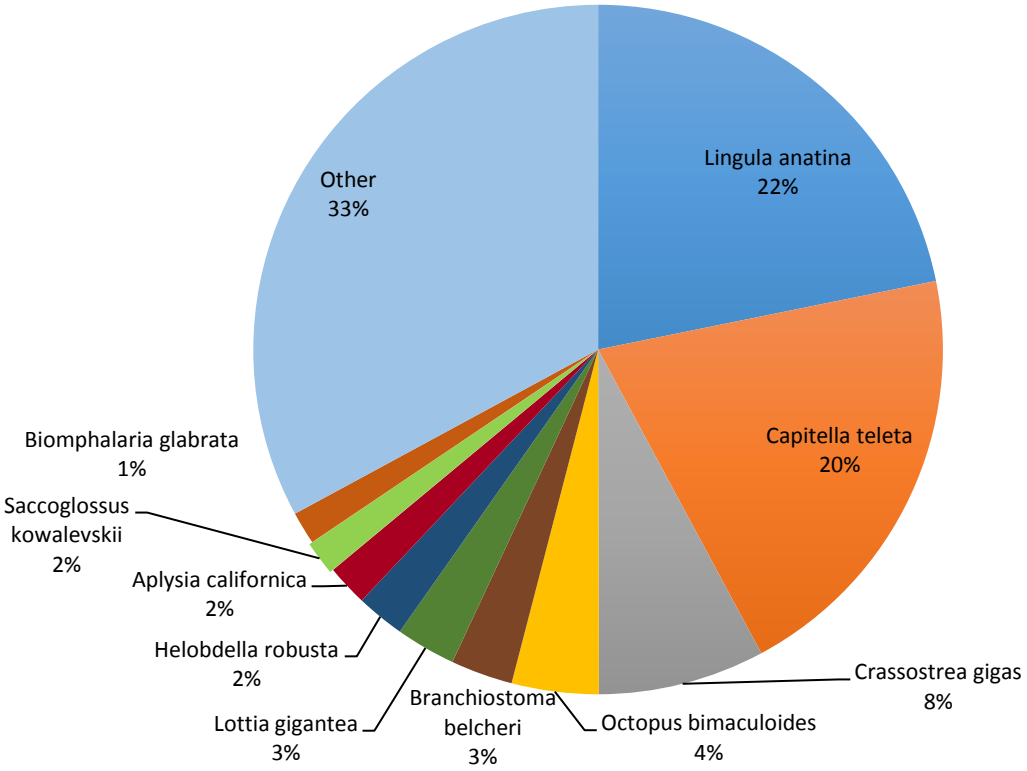


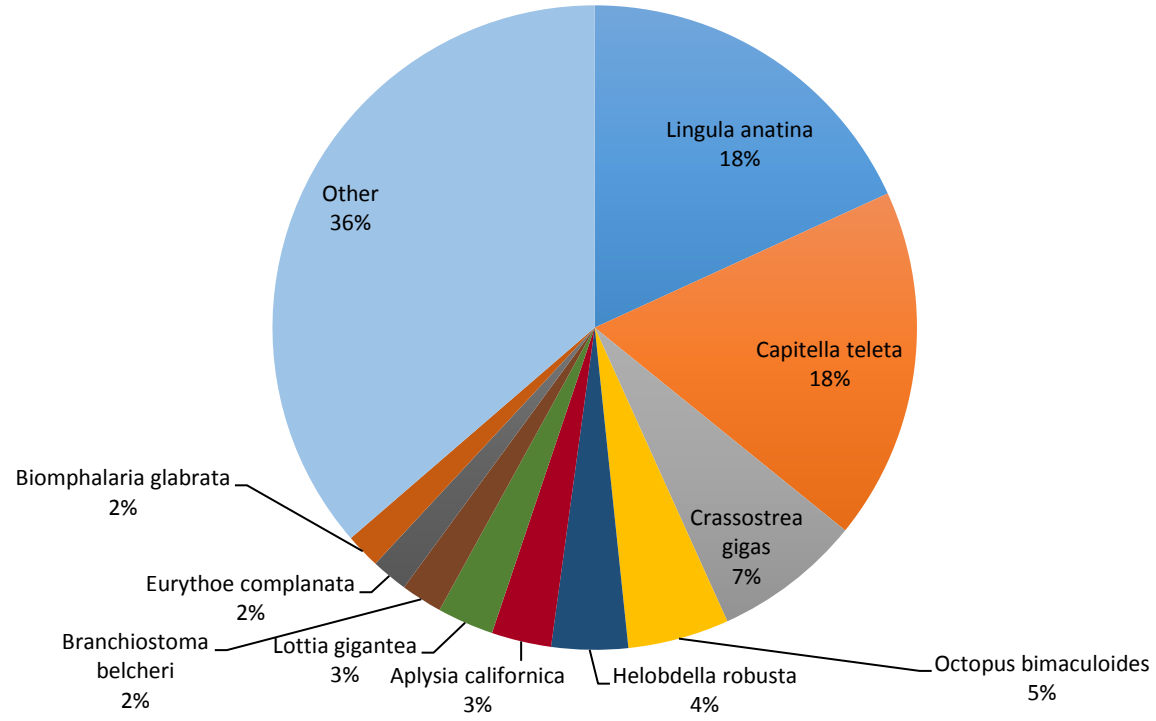
Figure 5. Gene ontology (GO) and annotation results of putative proteins using the Blast2Go pipeline for the orthologous proteins to Lophotrochozoa. For comparison, total protein sequences (with identical sequences removed) are shown in blue. BLASTp (orange) and InterProScan (grey) were two databases used to assign putative gene ontology. Blast2Go has functions including mapping (yellow) and annotation (green) for the recovery and assignment of GO terms.

The top taxonomic source for blast hits included *Lingula anatina*, a brachiopod, to *P. jeffreysii*, *H. carunculata*, and *E. capensis*. Conversely, *C. pinnata* had a top species BLAST hit to an annelid, *Capitella teleta* (Figure 6). Some top 10 species BLAST hits were unique to one amphinomid, including *Limulus Polyphemus* (Atlantic horseshoe crab) to *C. pinnata* (Figure 6C). *Eurythoe complanata*, an amphinomid fireworm, was only in the top 10 species BLAST hits to *H. carunculata* and *E. capensis* (Figure 6B, Figure 6D).

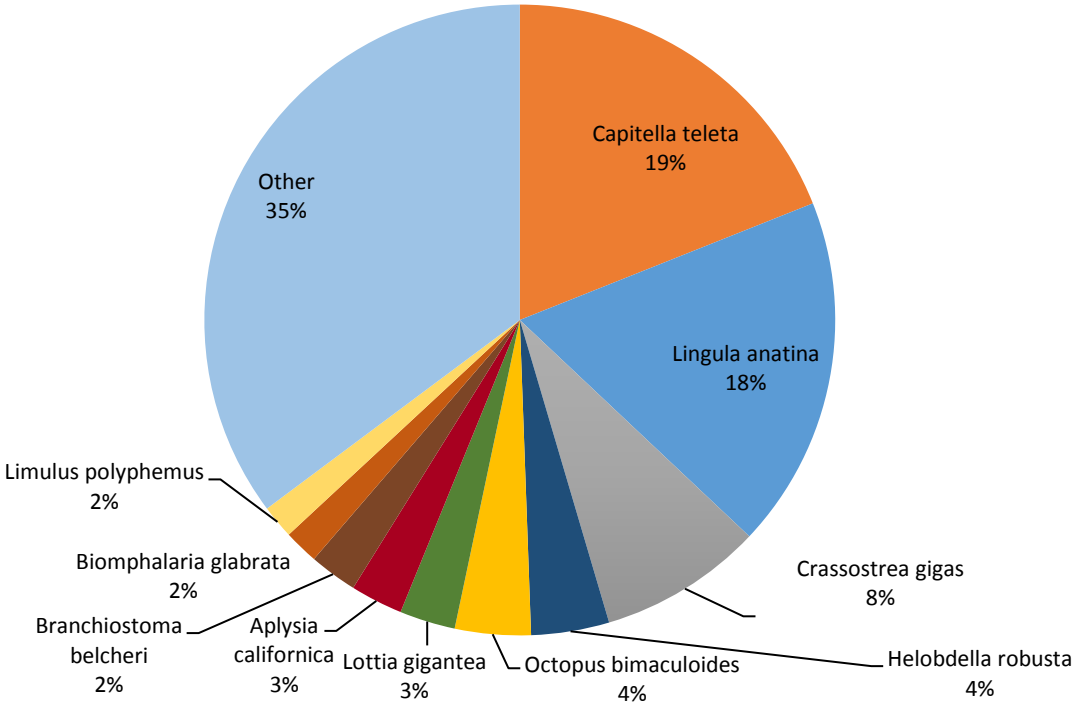
6A


Lingula anatina
22%
Capitella teleta
20%
Crassostrea gigas
8%
Octopus bimaculoides
4%
Branchiostoma belcheri
3%
Lottia gigantea
3%
Helobdella robusta
2%
Aplysia californica
2%
Saccoglossus kowalevskii
2%
Biomphalaria glabrata
1%
Other
33%

6B


Lingula anatina
18%
Capitella teleta
18%
Crassostrea gigas
7%
Octopus bimaculoides
5%
Helobdella robusta
4%
Aplysia californica
3%
Lottia gigantea
3%
Branchiostoma belcheri
2%
Eurythoe complanata
2%
Biomphalaria glabrata
2%
Other
36%

6C



6D

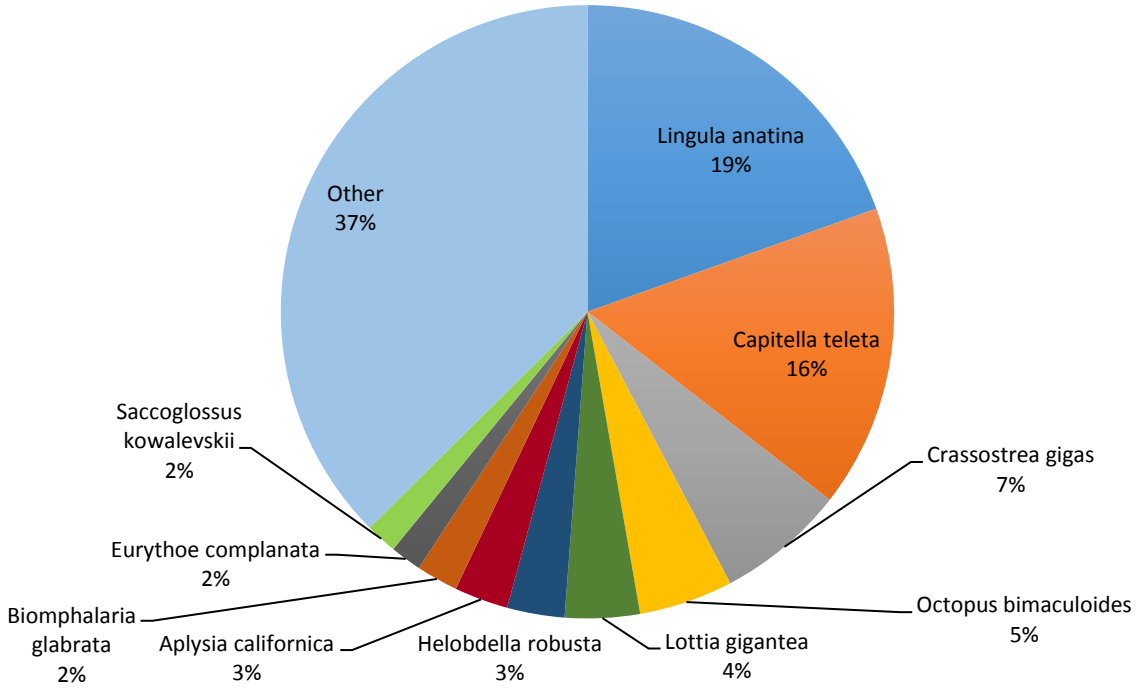


Figure 6. The top BLAST species hit for *P. jeffreysii* (A), *H. carunculata* (B), *C. pinnata* (C), and *E. capensis* (D).

The results of the top 50 GO-Slim biological processes show the most frequent biological processes in the orthologous set based on GO terms. Among *P. jeffreysii, H. carunculata, C. pinnata,* and *E. capensis*, there are slight variations in the rank of some biological processes, for example, the top biological process in *P. jeffreysii*, *C. pinnata*, and *E. capensis* is "cellular protein modification process" but in *H. carunculata*, the top biological process is ribosome biogenesis (Figure 7, Appendix 1A, 1B, 1C). Overall, 42 out of the top 50 categories are shared (Figure 7, Appendix 1A, 1B, 1C).

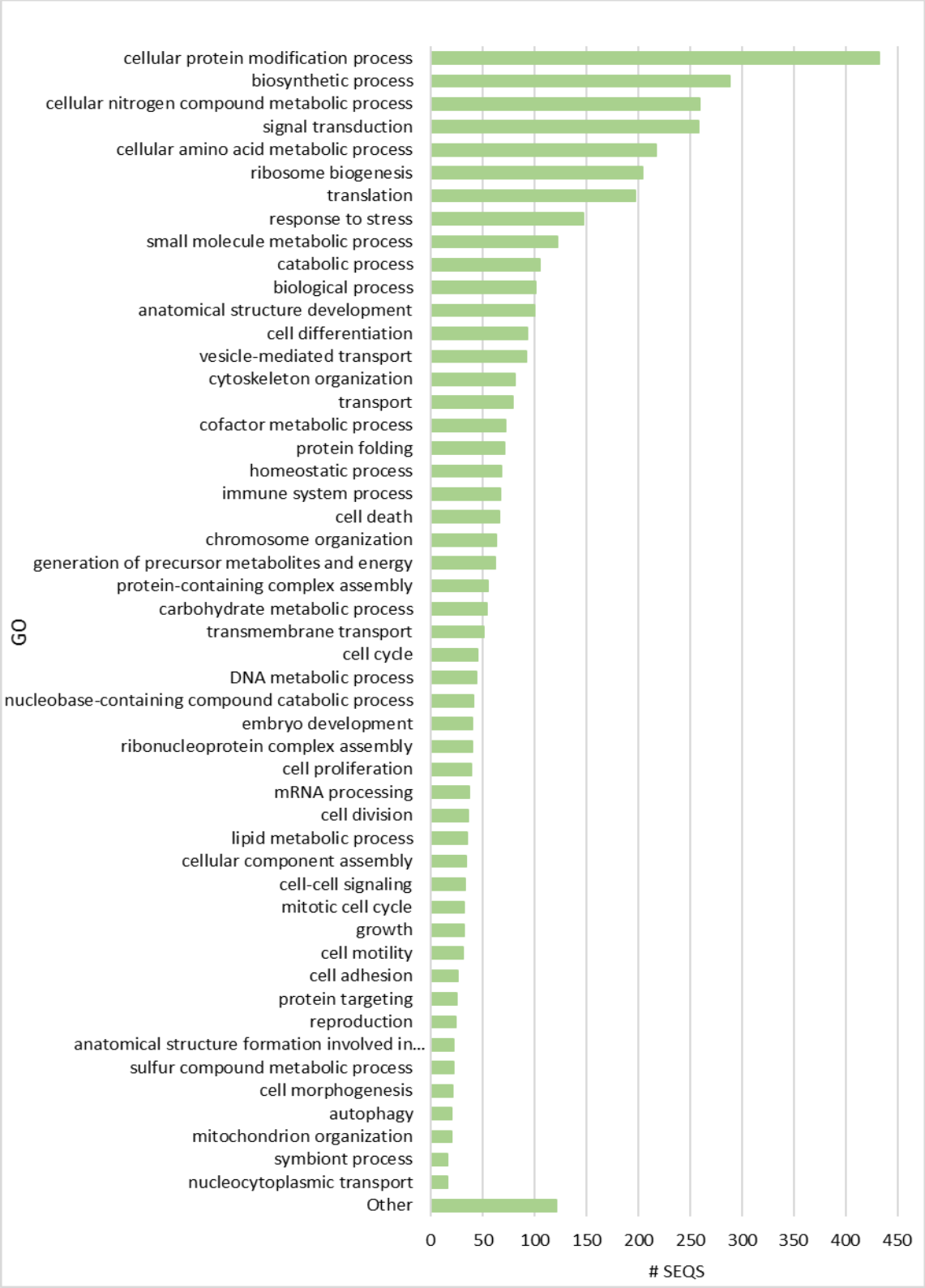


Figure 7. The Generic GO-Slim results showing the top 50 biological processes for *P. jeffreysii*.

CHAPTER IV

DISCUSSION

This study has expanded upon knowledge of Amphinomida and Lophotrochozoa by providing the annotations of their shared orthologous genes for future phylogenomic analysis. The orthologous set for the amphinomids and Lophotrochozoans were designated by gene numbers from the Lophotrochozoan orthologous gene sequence supplementary file from Weigert et al. 2014, which included a single sequence per species representative. The genes in the Lophotrochozoan database were published without annotations, because the purpose of their study was to determine phylogenetic relationships within Annelida (Weigert et al. 2014). The next step beyond determining phylogenetic relationships is to characterize and annotate the transcriptomes for comparison between different taxa to identify shared and unique biological pathways. The gene descriptions generated from the amphinomid data in Blast2Go has been supplemented to the list of Lophotrochozoan orthologs, for example, gene 21885 has a gene description of a vigilin protein which is present in the four amphinomids and Lophotrochozoa (Appendix 4). This gene, along with the other 763 orthologs, can be used for comparison between distantly related taxa to uncover information about differences in structure and function (Figure 5).

Although *de novo* assembly is a reliable method for transcriptome assembly, particularly for non-model organisms, a major limitation was residual sequence redundancy relative to other methods including genome-guided assembly (Huang et al. 2016). Redundancy can be seen by the discrepancy in the number of initial BLAST hits to the lophotrochozoan database (See Table 3, Homologous Hits (≥30% Seq Similarity) and the actual number of unique sequences (See Table

3, Actual Protein Hits). Similarly, when evaluating the individual gene files, there was the presence of multiple similar sequences for a single gene. For example, Appendix 3 shows an alignment for gene 21889 (39s ribosomal mitochondrial-like) which contains multiple sequences from each Amphinomid.

The amphinomid proteins with hits to gene 21889 also show similar amino acid sequences. For example, the protein sequence Cpinn_149868, differs from Cpinn_149867 because it has an additional six amino acids at the beginning of the sequence (Appendix 3). Also, proteins which were aligned to a reference gene aligned to different sections of the gene, and overlapped in certain regions, suggesting incomplete and/or missing fragments. For example, alignment of sequence Hcaru_5437 with other sequences begins at amino acid position 59 (Appendix 3). The sequence similarity and presence of more than one sequence per gene for each amphinomid, is likely due to alternative splicing resulting in multiple isoforms (Breitbart et al. 1987). Given the high level of redundancy, this study highlights the importance of filtering redundant data earlier in the workflow, after gene assembly. Pipelines have been developed, including DRAP (*de novo* RNA-seq Assembly Pipeline), which allows for the reduction of similar sequences after Trinity assembly, which would have aided this study (Cabau et al. 2016). As a result, future work will include the elimination of similar sequences after assembly and identification of a single homologous sequence for each gene per species for downstream alignment and phylogenomic analysis. This will allow for further identification, characterization, and comparison of key shared biological pathways within Amphinomida to Lophotrochozoan relatives.

Despite the potential for thousands of candidate genes among the four datasets, there was, on average, only a 32% recovery of orthologs, and further reduced to an average of 764 genes

that were shared among the four species (Figure 4). The amphinomid orthologous set only had 1,046 protein hits to the 2,339 Lophotrochozoan genes with estimated transcriptome completeness varying between 8.2 and 36.2% (Figure 4, Table 1). These low recoveries are possibly due to sampling error outside of our control. The specific site of muscle tissue extracted by the Halanych Lab at Auburn University was not shared, but the species are variable by size and may have influenced the transcriptome completeness depending on if whole animal or specific sections (i.e. posterior segments) were extracted.

The previously published transcriptome of *H. carunculata* used 58,454 ORFs in which 32,500 were assigned ISP IDs, our data showed 12,186 ORFs (See Table 2, "Identical Sequences Removed") and 18,773 confirmed ISP GOs (Mehr et al. 2015). For comparison, Appendix 2 shows the number of contigs per contig length (bp), which corresponds to the published transcriptome majority contig lengths between 200-600 bp (Mehr et al. 2015). Mehr and colleagues (2015) reduced their contigs by selecting ORFs longer than 200 amino acids; in our data, we opted for sequence lengths >100 amino acids to avoid a reduction to the working dataset. Annotations of proteins in various categories have been published for *H. carunculata* (i.e. immune response genes, reproduction genes, potential phylogenetic markers), using Blast2Go, but our results lacked many of these results (Mehr et al. 2015).

A higher percentage of annotations to total proteins was seen in the orthologous set shared with Lophotrochozoa versus the whole transcriptomes of each Amphinomid (Figure 2, Figure 5). This is expected because the orthologous proteins conserved between distantly related taxa are necessary for life and are more likely to have been previously studied. The IPS annotations indicate that although *H. carunculata* had lower IPS GOs before and after merging with BLAST GO terms compared to E. capensis, which had the lowest transcriptome

completeness (Table 1), that IPS was able of retrieve a higher amount of confirmed IPS GOs compared to *E. capensis* (Figure 3). This highlights the importance of using multiple databases besides NCBI for annotation, since around 11% of proteins in the whole transcriptome H. carunculata had no results and the merging function serves to add and validate GO terms (Figure 2, Figure 3).

The top 10 species BLAST hit reveals differences in the source of BLAST hits for the amphinomids (Figure 6). Interestingly, *Eurythoe complanata*, an amphinomid fireworm, was only in the top 10 species BLAST hits to *H. carunculata* and *E. capensis* (Figure 6B, Figure 6D). This could indicate a lack of data for *E. complanata*, and/or a lowering in the rank of top species BLAST hits if dominant isoforms had BLAST hits to organisms other than *E. complanata*. Further investigation to see what proteins hit to certain organisms could reveal interesting shared characteristics between the distantly related taxa, for example, only *C. pinnata* had BLAST hits to *Limulus Polyphemus*, or the Atlantic horseshoe crab (Figure 6C).

The study of transcriptomics in non-model organisms has its limitations due to the lack of closely related reference genomes not only in terms of annotation, but also in analysis of the transcriptome completeness. The closest relatives available as the reference lineage was either *Caenorhabditis*, which is distantly related to annelids, or members of Arthropoda (i.e. *Diptera*, *Endopterygota*, *Hymenoptera*). This study demonstrates that non-annelid references in available databases limits reliable quantification of transcriptome data completeness (Table 1). Low completeness in our data may also be a result of the initial tissue sampling method or other laboratory error beyond our control.

Large-scale genomic data for amphinomids is limited, and only available as raw sequencing reads (Sequence Read Archive: *Pareurythoe californica* (SRX965727),

Paramphinome jeffreysii (SRX518630), *Hermodice carunculata* (SRX194586)). Some simply lack sequence gene identities (i.e. NCBI, Expressed Sequence Tags: *Eurythoe complanata*), thus limiting gene discovery and the ability to build comparable datasets for insights into the evolution of genes of interest. Investing time into annotating genomic datasets will provide the ability to understand biological, cellular, and molecular processes in amphinomid species in relation to Lophotrochozoa, for example. The current study is a major step forward in overcoming this barrier and offers a workflow for a) gene annotation and identity b) shared gene isolation among differential assembly file sizes, which establishes a foundation for continuing research on the transcriptomics of Amphinomida.

REFERENCES

Ahrens JB, Borda E, Barroso R, Campbell AM, Wolf A, Nugues M, Paiva P, Rouse GW, Schulze A (2013) The curious case of Hermodice carunculata (Annelida: Amphinomidae): evidence for genetic homogeneity throughout the Atlantic Ocean and associated basins. Mol Ecol 22:2280–2291

Ahrens JB, Kudenov JD, Marshall CD, Schulze A (2014) Regeneration of posterior segments and terminal structures in the bearded fireworm, Hermodice carunculata (Annelida: Amphinomidae). J Morphol 275:1103–1112. doi:10.1002/jmor.20287

Altschul S, Gish W, Miller W, Myers E, Lipman D (1990) Basic local alignment search tool. J Mol Biol 215:403–410. doi:10.1016/S0022-2836(05)80360-2

Andrade SC, Novo M, Kawauchi GY, Worsaae K, Pleijel F, Giribet G, Rouse GW (2015) Articulating "Archiannelids": Phylogenomics and Annelid Relationships, with Emphasis on Meiofaunal Taxa. Mol Bio Evol 32:2860-2875. doi:10.1093/molbev/msv157

Bonyadi-Naeini A, Rastegar-Pouyani N, Rastegar-Pouyani E, Glasby CJ, Rahimian H (2017) Intertidal polychaetes from Abu Musa Island, Persian Gulf, with a new species from the Perinereis cultrifera species complex (Annelida: Nereididae). J Mar Biol Assoc UK 1-12.

Borda E, Kudenov JD, Beinhold C, Rouse GW (2012) Towards a revised Amphinomidae (Annelida, Amphinomida): description and affinities of a new genus and species from the Nile Deep-sea Fan, Mediterranean Sea. Zool Scripta 40:307–325

Borda E, Kudenov JD, Blake JA, Chevaldonné P, Desbruyères D, Hourdez S, Fabri M-C, Pleijel F, Schulze A, Shank T, Wilson NG, Rouse GW (2013) Cryptic species of Archinome (Annelida:Amphinomida) from vents and seeps. Proc R Soc London Biol 280:20131876.

Borda E, Yáñez-Rivera B, Ochoa GM, Kudenov JD, Sanchez-Ortiz C, Schulze A, Rouse G (2015) Revamping Amphinomidae (Annelida: Amphinomida), with the inclusion of Notopygos. Zool Scripta 44:324–333

Breitbart RE, Andreadis A, Nadal-Ginard B (1987) Alternative Slicing: A Ubiquitous Mechanism for the Generation of Multiple Protein Isoforms from Single Genes. Annu Rev Biochem 56:467-495. doi:10.1146/annurev.bi.56.070187.002343

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797. doi:10.1093/nar/gkh340

Ferrier DEK (2012) Evolutionary crossroads in developmental biology: annelids. Development 139:2643-2653. doi:10.1242/dev.074724

Götz S et al. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 36:3420-3435

Haas BJ et al. (2013) De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat Protoc 8:1494-1512. doi:10.1038/nprot.2013.084

Huang X, Chen XG, Armbruster PA (2016) Comparative performance of transcriptome assembly methods for non-model organisms. BMC Genomics 17:523. doi:10.1186/s12864-016-2923-8

Huang Y, Niu B, Gao Y, Fu L, Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26:680

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Mentjies P, Drummond A (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647-1649

Mehr S, Verdes A, DeSalle R, Sparks J, Pieribone V, Gruber DF (2015) Transcriptome sequencing and annotation of the polychaete Hermodice carunculata (Annelida, Amphinomidae). BMC Genomics 16:445. doi:10.1186/s12864-015-1565-6

Nakamura K, Tachikawa Y, Kitamura M, Ohno O, Sugnumac M, Uemura D (2008) Complanine, and inflammation-inducing substance isolated from the marine fireworm *Eurythoe complantata.* Org Biomol Chem 6:2058-2060. doi:10.1039/b803107j

Rouse GW, Pleijel F (2001) Polychaetes (Polychaetes). Oxford University Press, New York

Schulze A, Grimes C, Rudek T (2017) Tough, armed and omnivorous: Hermodice carunculata (Annelida: Amphinomidae) is prepared for ecological challenges. J Mar Biol Assoc UK 97:1075-1080. doi:10.1017/S0025315417000091

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210-3212. doi:10.1093/bioinformatics/btv351

Struck TH, Paul C, Hill N, Hartmann S, Hösel C et al. (2011) Phylogenomic analyses unravel annelid evolution. Nature 472:95-98. doi:10.1038/nature09864

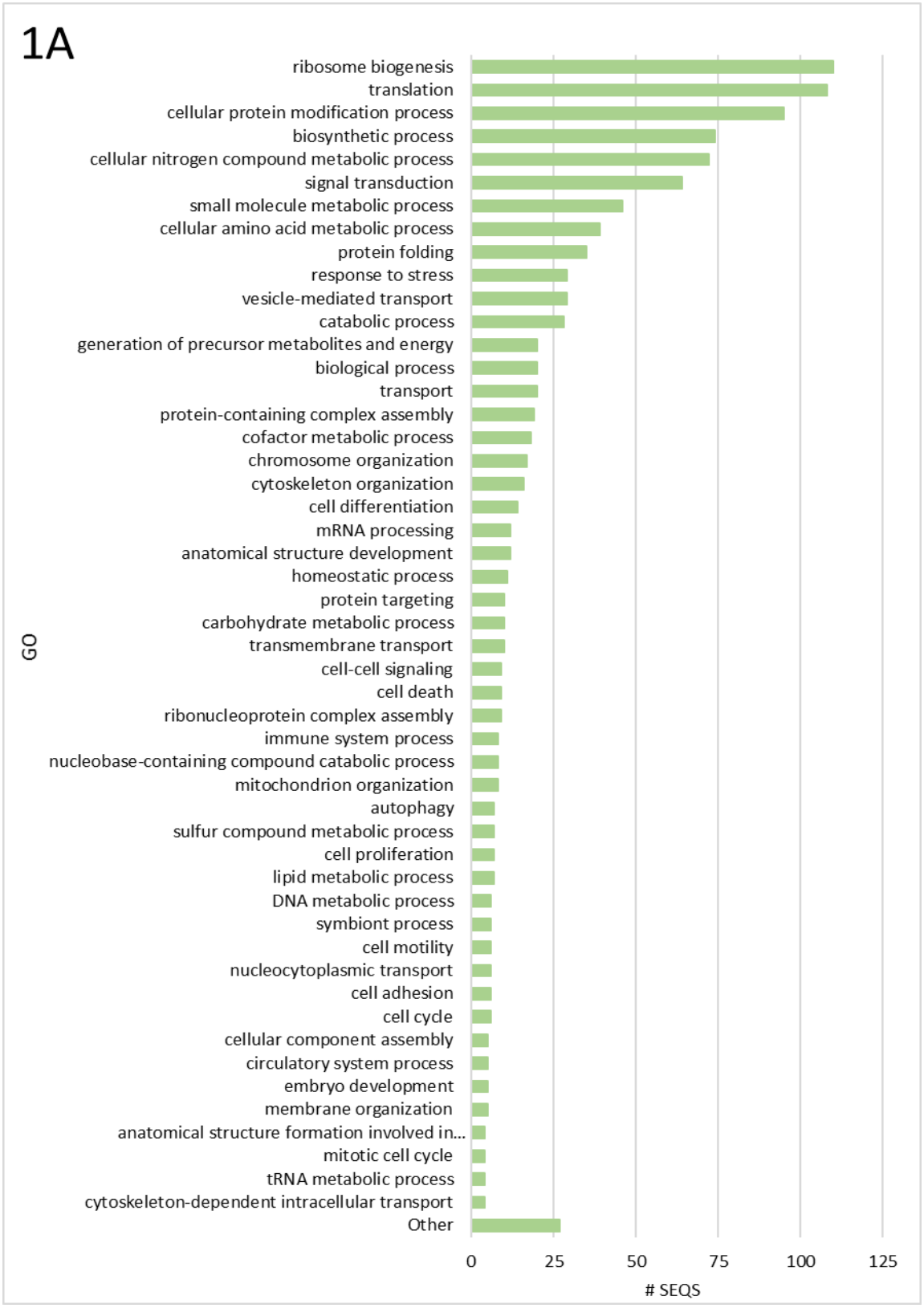
Sun Y, Li X (2017) A new genus and species of bristle worm from Beibu Gulf, South China Sea (Annelida, Polychaeta, Amphinomidae). ZooKeys 708:1–10

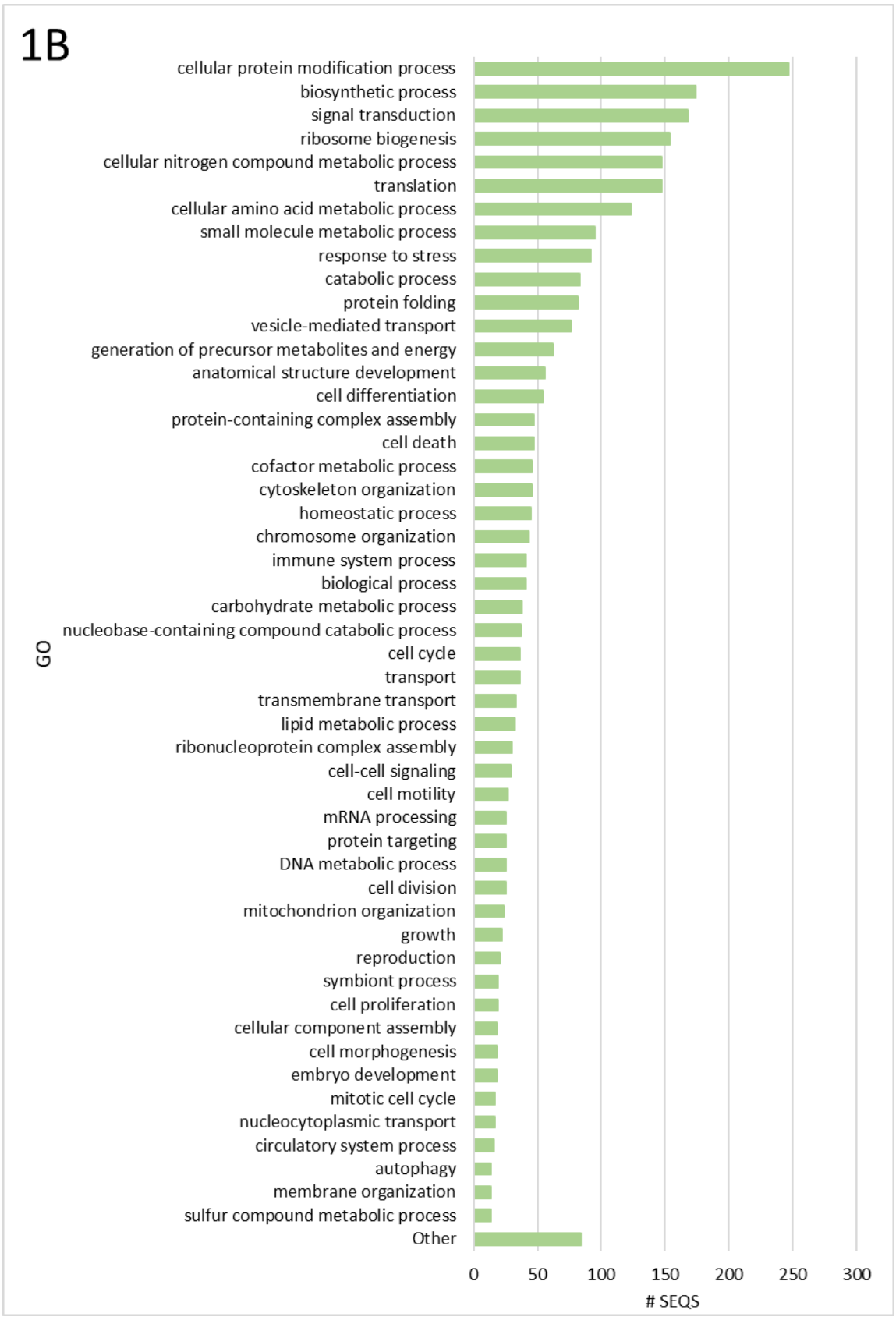
Verdes A, Simpson D, Holford M (2018) Are Fireworms Venomous? Evidence for the Convergent Evolution of Toxin Homologs in Three Species of Fireworms (Annelida, Amphinomidae) Genome Biol Evol 10:249-268. doi:10.1093/gbe/evx279

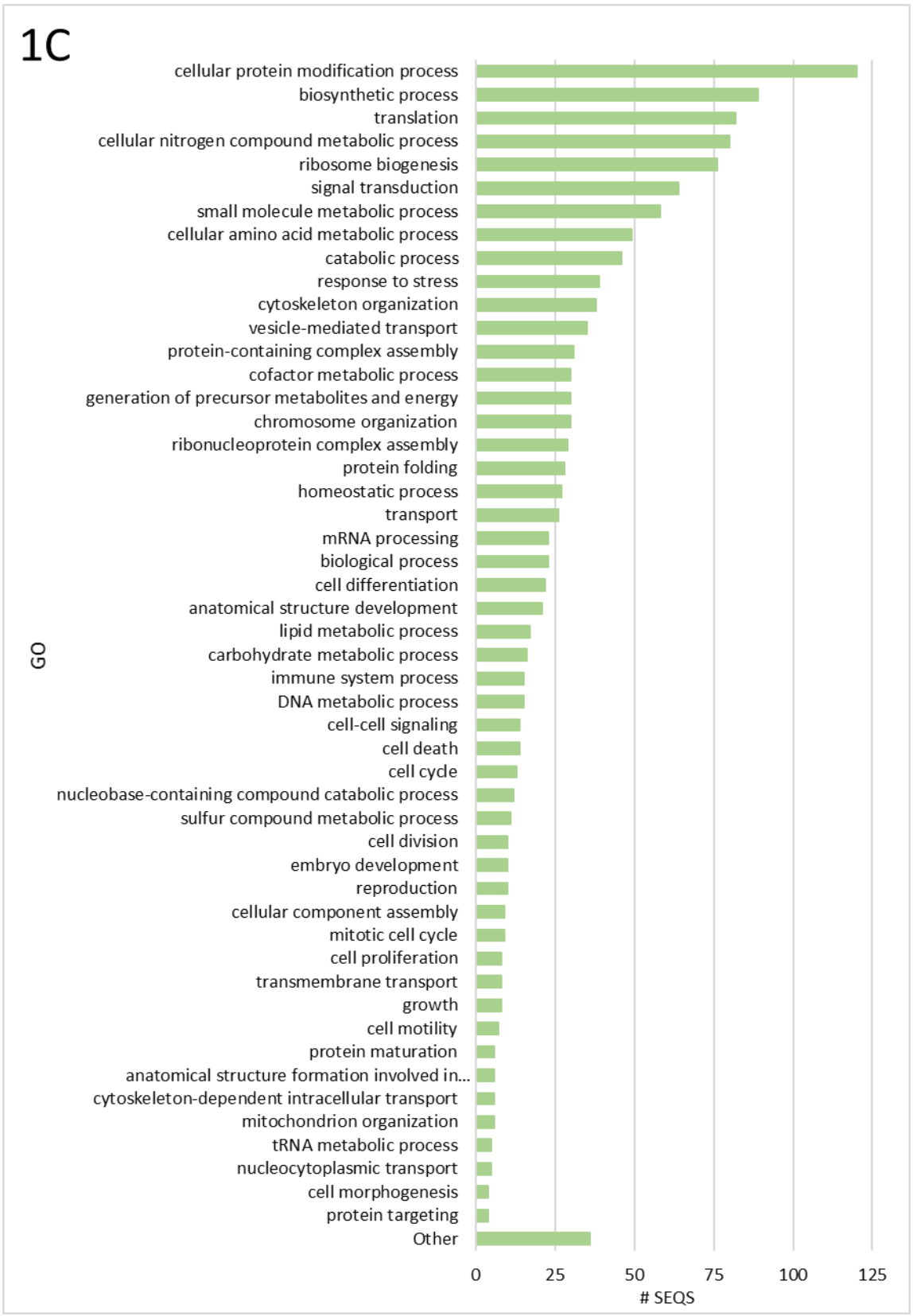
Weidhase M, Bleidorn C, Beckers P, Helm C, (2016) Myoanatomy and anterior muscle regeneration of the fireworm Eurythoe cf. complanata (Annelida: Amphinomidae). J Morphol 277:306–315. doi:10.1002/jmor.20496

Weigert A, Helm C, Meyer M, Nickel B, Arendt D, Hausdorf B, Santos SR, Halanych KM, Purschke G, Bleidorn C, Struck TH (2014) Illuminating the Base of the Annelid Tree Using Transcriptomics. Mol Biol Evol 31:1391–1401. doi:10.1093/molbev/msu080

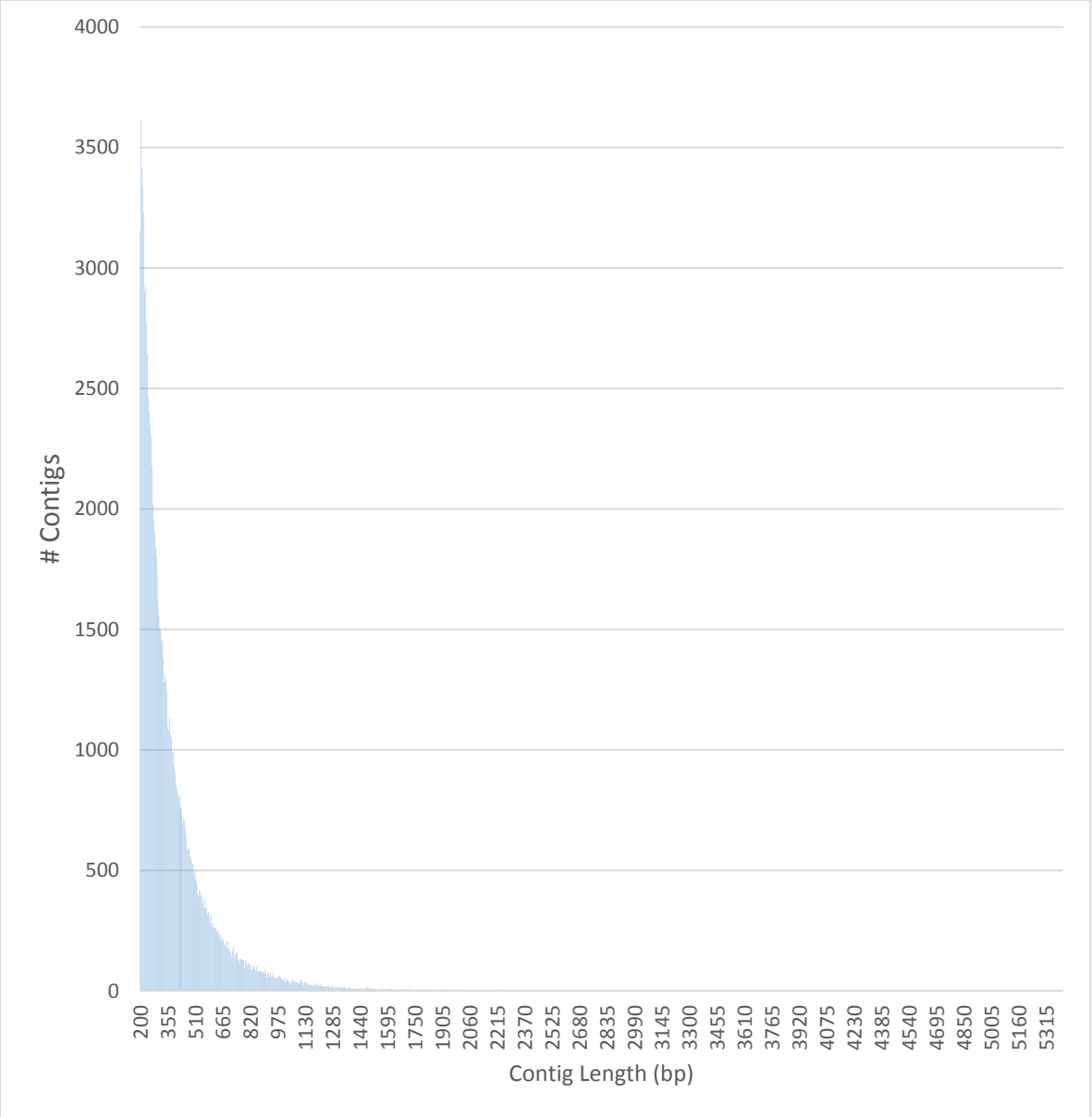
APPENDIX


1A
ribosome biogenesis
translation
cellular protein modification process
biosynthetic process
cellular nitrogen compound metabolic process
signal transduction
small molecule metabolic process
cellular amino acid metabolic process
protein folding
response to stress
vesicle-mediated transport
catabolic process
generation of precursor metabolites and energy
biological process
transport
protein-containing complex assembly
cofactor metabolic process
chromosome organization
cytoskeleton organization
cell differentiation
mRNA processing
anatomical structure development
homeostatic process
protein targeting
carbohydrate metabolic process
transmembrane transport
cell-cell signaling
cell death
ribonucleoprotein complex assembly
immune system process
nucleobase-containing compound catabolic process
mitochondrion organization
autophagy
sulfur compound metabolic process
cell proliferation
lipid metabolic process
DNA metabolic process
symbiont process
cell motility
nucleocytoplasmic transport
cell adhesion
cell cycle
cellular component assembly
circulatory system process
embryo development
membrane organization
anatomical structure formation involved in...
mitotic cell cycle
tRNA metabolic process
cytoskeleton-dependent intracellular transport
Other
GO
0
25
50
75
100
125
SEQS

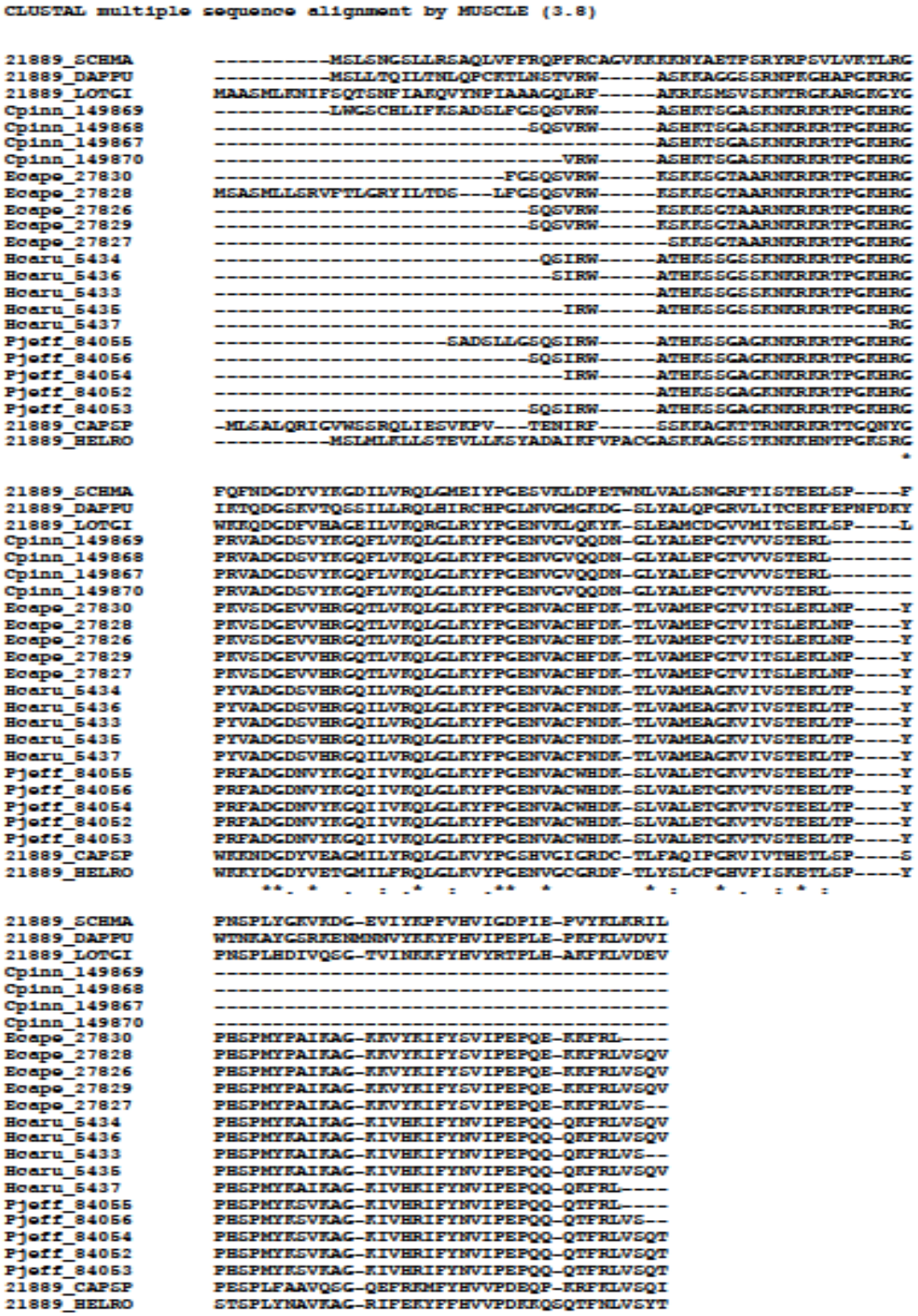

1B
cellular protein modification process
biosynthetic process
signal transduction
ribosome biogenesis
cellular nitrogen compound metabolic process
translation
cellular amino acid metabolic process
small molecule metabolic process
response to stress
catabolic process
protein folding
vesicle-mediated transport
generation of precursor metabolites and energy
anatomical structure development
cell differentiation
protein-containing complex assembly
cell death
cofactor metabolic process
cytoskeleton organization
homeostatic process
chromosome organization
immune system process
biological process
carbohydrate metabolic process
nucleobase-containing compound catabolic process
cell cycle
transport
transmembrane transport
lipid metabolic process
ribonucleoprotein complex assembly
cell-cell signaling
cell motility
mRNA processing
protein targeting
DNA metabolic process
cell division
mitochondrion organization
growth
reproduction
symbiont process
cell proliferation
cellular component assembly
cell morphogenesis
embryo development
mitotic cell cycle
nucleocytoplasmic transport
circulatory system process
autophagy
membrane organization
sulfur compound metabolic process
Other
GO
0
50
100
150
200
250
300
SEQS



Appendix 1. The GO-Slim results showing the top 50 biological processes for *P. jeffreysii* (A), *H. carunculata* (B), *C. pinnata* (C), and *E. capensis* (D).



Appendix 2. The number of contigs per contig length for *H. carunculata*.



Appendix 3. An alignment for the gene 21889 (39s ribosomal mitochondrial-like) containing sequences from amphinomids *P. jeffreysii*, *H. carunculata*, *C. pinnata*, and *E. capensis*, and lophotrochozoans *Helobdella robusta* (HELRO), *Capitella teleta* (CAPSP), *Lottia gigantea* (LOTGI), *Schistosoma mansoni* (SCHMA), and *Daphnia pulex* (DAPPU).

Appendix 4. The first 10 Lophotrochozoan genes. Gene IDs are assigned from Weigert et al. 2014, and supplemented with gene descriptions based on the top BLAST result for the amphinomids. The "X" indicates presence, and blanks indicate absence of that gene for the associated amphinomid species.

Gene ID	Gene Description	*P. jeffreysii*	*H. carunculata*	*E. capensis*	*C. pinnata*
21884	Coatomer subunit zeta-1 isoform x3, Coatomer subunit zeta-1-like, Coatomer subunit zeta-1-like isoform x1"	X	X	X	X
21885	Predicted: vigilin-like, Vigilin isoform x1, Vigilin-like"	X	X	X	X
21886	Luc7 3, Luc7 3 isoform x2, Rna-binding luc7-like 1 isoform x1, Rna-binding luc7-like 2"	X	X	X	X
21887	Potassium channel, Potassium voltage-gated channel subfamily a member 1 isoform x1"	X			X
21888	Dnaj homolog subfamily c member 13-like isoform x1, Dnaj homolog subfamily c member 13-like isoform x2"	X	X	X	X
21889	39s ribosomal mitochondrial-like"	X	X	X	X
21891	Achain crystal structure of ptpn12 catalytic domain, Receptor tyrosine, Receptor tyrosine phosphatase type	X	X	X	X
21893	Gastrula zinc finger -like, Gastrula zinc finger -like isoform x1, Krueppel homolog 1-like, Transcriptional repressor ctcf-like, Zinc finger, Zinc finger, Zinc finger 182-like isoform x1, Zinc finger	X	X	X	X
21895	Cdc42-interacting 4, Formin-binding 1, Formin-binding 1 homolog, Formin-binding 1-like isoform x2, Formin-binding 1-like isoform x3"	X	X	X	X
21896	Probable cytosolic iron-sulfur assembly ciao1 homolog"	X			