TRANSCRIPTOME ASSEMBLY AND CHARACTERIZATION OF THE BENTHIC ANNELID *PARAMPHINOME JEFFREYSII*

An Undergraduate Research Scholars Thesis

by

IRENE MARTINEZ

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 M. Labonté
Dr. Elizabeth Borda

May 2019

Major: Marine Biology

TABLE OF CONTENTS

ABSTRACT

Transcriptome Assembly and Characterization of the Benthic Annelid *Paramphinome jeffreysii*

Irene Martinez
Department of Marine Biology
Texas A&M University

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

Species of the phylum Annelida have been essential as model organisms in the studies of biology, neurobiology, evolution, ecology, and phylogenomics. Prior work using genomics and transcriptomics has provided new insights into the evolution of Annelida, such as phylogenetic relationships, life history, and lifestyle adaptations. Little biological information is known about the amphinomid *Paramphinome jeffreysii*. Although transcriptome data have been available since 2014, complete annotations of gene content is not publically available. The objective of this research is to annotate the transcriptome of *P. jeffreysii* in order to contribute towards a more comprehensive understanding of the biological pathways, cellular components, and molecular functions of the species. Cellular and metabolic processes, as well as biological regulation were among the top biological processes discovered, while binding, catalytic activity, and transporter activity were among the top molecular functions found. The top-hit species included a brachiopod, *Lingula anatina*, as well as *Capitella teleta* (second top hit), and several mollusks, highlighting the lack of available comparable annotated data for Amphinomida, and Annelida in general within public databases. Therefore, the continued exploration of transcriptomics in non-model organisms, such as *P. jeffreysii*, allows for continuing comparative research.

ACKNOWLEDGEMENTS

I would like to thank my research advisors, Drs. Jessica Labonté and Elizabeth Borda, as well as Dr. Orissa Moulton for their guidance and support throughout the course of this research. I want to thank my predecessors, Giovanni Madrigal and Arianna Bartlett, for their help with my questions. I also want to give special thanks to Yui Matsumoto – without her encouragement, patience, support, and guidance, this thesis would not have been possible.

Thanks go to my friends and my fellow members of the Labonté Viral Ecology Lab for their assistance and constant support. I also want to extend my gratitude to the Louis Stokes Alliance for Minority Participation and Texas Sea Grant for allowing me this opportunity.

Finally, thanks to my mother for her endless encouragement and advocacy during this process.

CHAPTER I

INTRODUCTION

Background

Annelida is a highly diverse phylum, belonging to a branch of the Tree of Life within the protostome group Lophotrochozoa, including taxa that have the trochophore larval stage or a lophophore feeding apparatus in adults (Halanych et al., 1995; Mushegian, 2007). With over 16,500 described species in Annelida, this taxonomic group is comprised of segmented worms that are found worldwide in many habitats, including damp terrestrial, aquatic, and marine environments (Struck, 2013; Struck et al., 2011). Earthworms and leeches are the most recognizable annelids; however, the majority of the known diversity for Annelida lies among marine polychaetes (Rousset, Pleijel, Rouse, Erséus, & Siddall, 2007).

Annelids exhibit a diversity of morphological features and a wide range of feeding strategies, such as suspension feeding, scavenging, active predation, and parasitically living on other metazoans (McHugh, 2000). Annelida also exhibits a wide array of development mechanisms and reproductive strategies, including both sexual and asexual reproduction, internal and external fertilization, brooding, and viviparism (Ferrier, 2012; Mehr et al., 2015). Annelida has been considered an ideal group in understanding major transitions in evolution of segmentation and the nervous system, transitions to terrestrial lifestyle, diversifications of larval types (Mehr et al., 2015; Weigert et al., 2014), and as models in evolutionary developmental studies to discover ancestral characteristics among bilaterians (Struck et al., 2011; Weigert et al., 2014).

Amphinomida, a clade of Annelida, comprises approximately 200 described species from 25 genera and is divided into two families: Euphrosinidae and Amphinomidae, which is further divided into the subfamilies Archinominae and Amphinominae (Borda, Kudenov, Bienhold, & Rouse, 2012; Mehr et al., 2015; Verdes, Simpson, & Holford, 2017) (Figure 1). Commonly referred to as "fireworms", Amphinomida consists of polychaetes that are unique with calcareous chaetae, that in some species are used as a defense mechanism, capable of piercing the soft tissue of predators, causing a painful sting (Ahrens et al., 2013; Borda et al., 2013; Verdes et al., 2017). They are distributed globally and common in intertidal, continental shelf communities, shallow, tropical reef and chemosynthetic environments (Borda et al., 2013; Mehr et al., 2015). Some species of Amphinomidae are capable of regenerating lost segments, and asexual reproduction (Ahrens, Kudenov, Marshall, & Schulze, 2014; Yáñez-Rivera & Méndez, 2014).

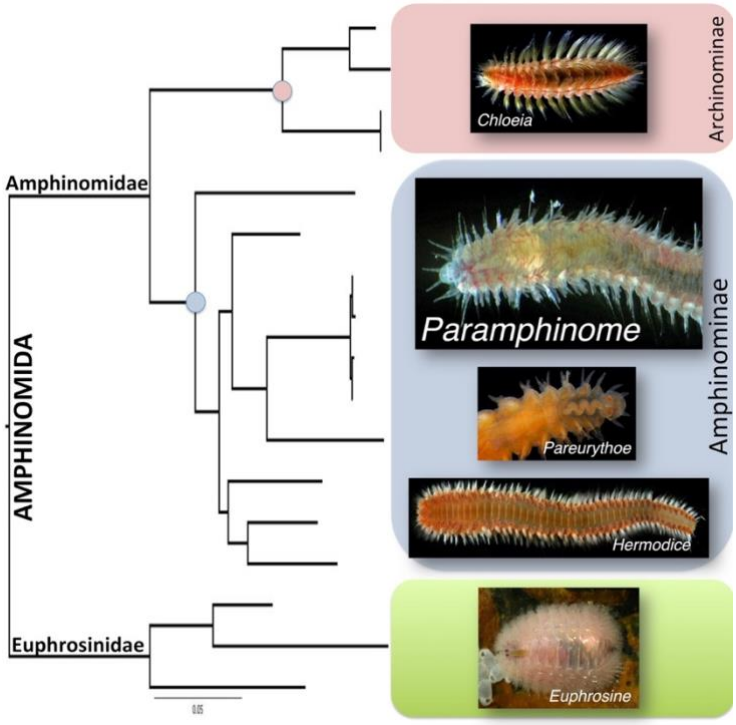

Figure 1. Phylogenetic tree of Amphinomida. The evolutionary relationship of five representative amphinomids across the three main clades, Amphinominae (*Paramphinome jeffreysii* [the focus of this study], *Hermodice carunculata*, and *Pareurythoe californica*), Euphrosinidae (*Euphrosine capensis*), and Archinominae (*Chloeia pinnata*). Image adapted from Borda et al. 2015.

Recent studies on amphinomids have primarily been centered on the taxonomy, higher-level phylogenetic and phylogenomic relationships, population genetics, and systematics (Ahrens et al., 2013; Andrade et al., 2015; Borda et al., 2012; Borda et al., 2013; Borda et al., 2015; Struck, 2013; Struck et al., 2011; Sun & Li, 2017; Verdes et al., 2017; Weigert et al., 2016; Weigert et al., 2014). Transcriptomics analyses have been performed for a few members of Amphinomida – *Hermodice carunculata* (Mehr et al., 2015; Verdes et al., 2017), *Paramphinome jeffreysii* (Verdes et al., 2017) and *Eurythoe complanta* (Weigert et al., 2016). Previous transcriptomic work has included *Paramphinome jeffreysii* to better understand deep level annelid relationship, and was also examined for the presence specific gene groups, such as venom/toxin genes or immunity toll-like receptors (Halanych & Kocot, 2014; Verdes et al., 2017; Weigert et al., 2014). Despite this work, complete annotated transcripts are not available, limiting comparative data and limiting the biological information for poorly known amphinomids, such as *P. jeffreysii*.

Objective

The objective of this research is to annotate the transcriptome for *P. jeffreysii*, in order to contribute towards a more comprehensive understanding of the biological processes, cellular components, and molecular functions that occur in this species.

CHAPTER II

METHODS

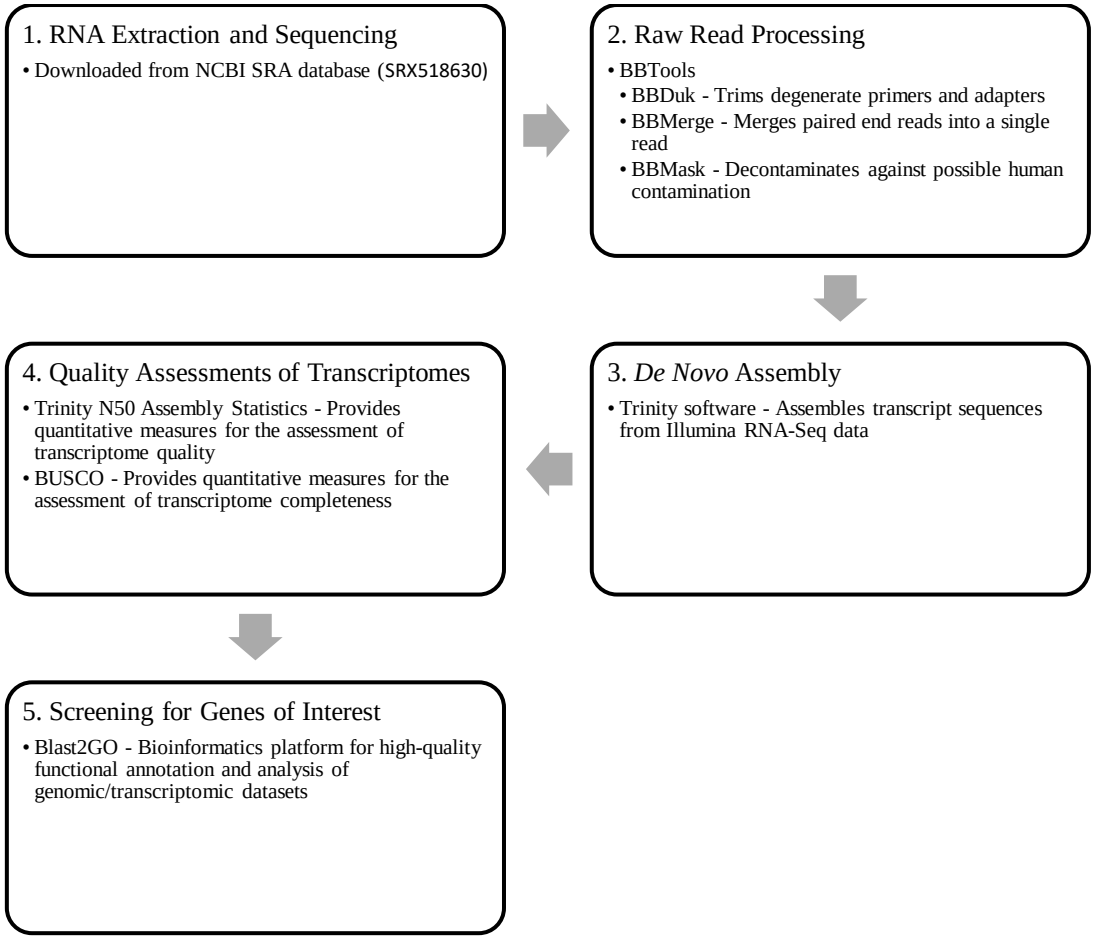

Figure 3. Workflow of bioinformatics methods.

Data Download, Decontamination, Quality Check, and *De Novo* Assembly

An overall workflow of the research methods described can be found above in Figure 2. Transcriptome files for *Paramphinome jeffreysii* were downloaded from the NCBI database (SRA Accession number: SRX518630). Raw reads were processed using BBTools (Bushnell, Rood, & Singer, 2017). The reads were first trimmed based on quality using BBDuk with the following parameters: kmer length of 31, hamming distance of 1, and force trim of 5. Primers

and adapters were then removed from the unmatched reads using BBDuk with the following parameters: kmer length of 23, hamming distance of 1, kmer trimming to the right, mink of 11, quality trim to the right, quality trim of 20, and specified to trim adapted based on pair overlap. The paired-end reads were then merged into single reads using BBMerge. BBMask was then used on the merged reads in order to remove possible human contamination and prevent false-positive matches in any highly conserved regions of the transcriptome.

The pre-processed reads were *de novo* assembled into contigs (defined as a contiguous region of DNA formed from overlapping reads representing part of a transcript) with Trinity (Haas et al., 2013) using default parameters. N50 assembly statistics were generated with the trinitystats.pl script within Trinity. N50 statistics represent the minimum contig length that accounts for 50% of the total assembly size. N50 allows quality determination since larger contigs are synonymous with better quality data.

Two transcriptome quality assessments were also performed using the BUSCO v2.0 software (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015) on both the whole *P. jeffreysii* transcriptome assembly, as well as the filtered assembly, using the default settings (expectation value of e^{-3}) for both the Metazoa and Eukaryota databases (files: metazoan_odb9 and eukaryota_odb9 downloaded from the software website (http://busco.ezlab.org)). BUSCO v2.0 software (Simão et al., 2015) was used in order to estimate the completeness of the transcriptomes, in regard to gene content. This method screens for sets of genes that represent the completeness of a transcriptome while considering the rare gene duplications and/or loss events that can occur during speciation. BUSCOs are displayed as either complete (match scores and length alignments fall within an expected range), fragmented (match scores are within the expected range, but length alignments are not within the expected range), or missing (no

significant matches or the match score is below the fragmented threshold). If there was poor coverage during the capture of the gene expression profile, then there will be a large portion of BUSCOs missing. Only contigs greater than 400 bp were kept for further analyses because most short contigs are uninformative.

Identification of Coding Regions and Annotation

Candidate open reading frames (ORFs) – the part of a reading frame that has the ability to be translated, therefore putative genes – were predicted using the annotation pipeline in Blast2GO PRO program (Conesa et al., 2005; Götz et al., 2008). Further annotations included comparison of the ORFs to various databases [NCBI non-redundant database, InterPro, and Gene Ontology (GO)] in order to assign functional terms to the transcriptome assembly. BLASTx was performed via CloudBlast against NCBI's Non-Redundant (NR) database using the following parameters: expectation value (E-value) of 10^{-3}, word size of 3 and high scoring segment pair (HSP) length cutoff of 33, and the top 20 hits were saved. InterPro Scan, which conducts domain-based searches, was conducted with the default settings, was performed against the following databases: CDD, HAMAP, HMMPanther, HMMPfam, HMMPIR, FPrintScan, BlastProDom, ProfileScan, HMMTigr, PatternScan. The results from the two annotation methods, InterPro Scan and Blast2GO, were then merged together in order to confirm results and add new GO annotations to the transcriptome.

Further understanding of the specific physiological components found in *P. jeffreysii* is allowed by placing the data into categories based on broader GO terms. The Blast2GO Pro output is based on three GO categories: biological processes, molecular functions, and cellular components. Biological processes are comprised of recognized series of molecular events and networks important to the functionality of the cell, organs, or organism as a whole. Molecular

function category is comprised of specific gene functions, such as molecular binding and enzyme catalysis. Lastly, subcellular and macromolecular structures where the gene products can be found are encompassed in the cellular component category.

CHAPTER III

RESULTS

Quality Assessment of Assembly

The paired-end, raw data files contained 14,418,967 and 15,308,126 reads. BBDuk removed 12,188 reads (0.04%) and 9,740 reads (0.03%) of contaminated reads, respectively. BBMerge joined 4,539,741 (34.21%) while 8,728,016 (65.78%) remained ambiguous and 64 (0.00%) had no solution for the first merged file. The second file joined 4,849,488 (34.05%) while 9,392,193 (65.94%) remained ambiguous and 344 (0.00%) had no solution. BBMask masked 40,655,284 (2.78%) and 41,514,306 (2.69%) base pairs. See Table 1 for a summary of the combined statistics.

Table 1. BBTools table of the combined statistics for *P. jeffreysii* (SRA# SRX518630)

Total Reads	29,727,093
Contaminated Reads	21,928
Joined Reads	9,389,229
Ambiguous Reads	18,120,209
No Solution Reads	408
Total Bases Masked	5.47%

After the paired-end reads were merged and *de novo* assembled with Trinity, 110,337 contigs were generated, with 68,879 of the genes were identified with unique transcript identifier prefixes. N50 assembly statistic report showed that 50% of the total assembly is encompassed in assembled contigs above 758 bp, while 30% of the total assembly was in contigs above 1,197 bp, and 10% of the total assembly was in contigs above 2,025 bp. After filtering for contigs longer

than 400 bp, 54,144 contigs were removed and 56,193 assembled transcripts remained. Of these, Trinity recorded 28,355 as unique transcripts. The N50 assembly statistic report showed that 50% of the total assembly was in contigs above 995 bp, 30% of the total assembly was in contigs above 1,427 bp, and 10% of the total assembly was in contigs above 2,222 bp. The statistical analysis saw an increase in mean contig length (from 589.45 to 881.16 bp) and the median contig length (406 to 695 bp) (Table 2). A distribution of the contig length can be visualized in Figure 3. The majority of the contigs contained less than 1,000 bp, which aligned with the N50 statistic of a mean contig length of 881 bp.

Table 2. Trinity *de novo* assembly statistics.

	Unfiltered	**Filtered (contigs >400bp)**
# of transcripts	110,337	56,193
# of unique transcripts	68,879	28,355
% GC	43.80	44.30
Contig N10	2,025	2,222
Contig N30	1,197	1,427
Contig N50	758	995
Median contig length	406	695
Mean contig length	589.45	881.16
Minimum length	202	400
Maximum length	8,097	8,097
Total assembled bases	65,038,316	49,515,008

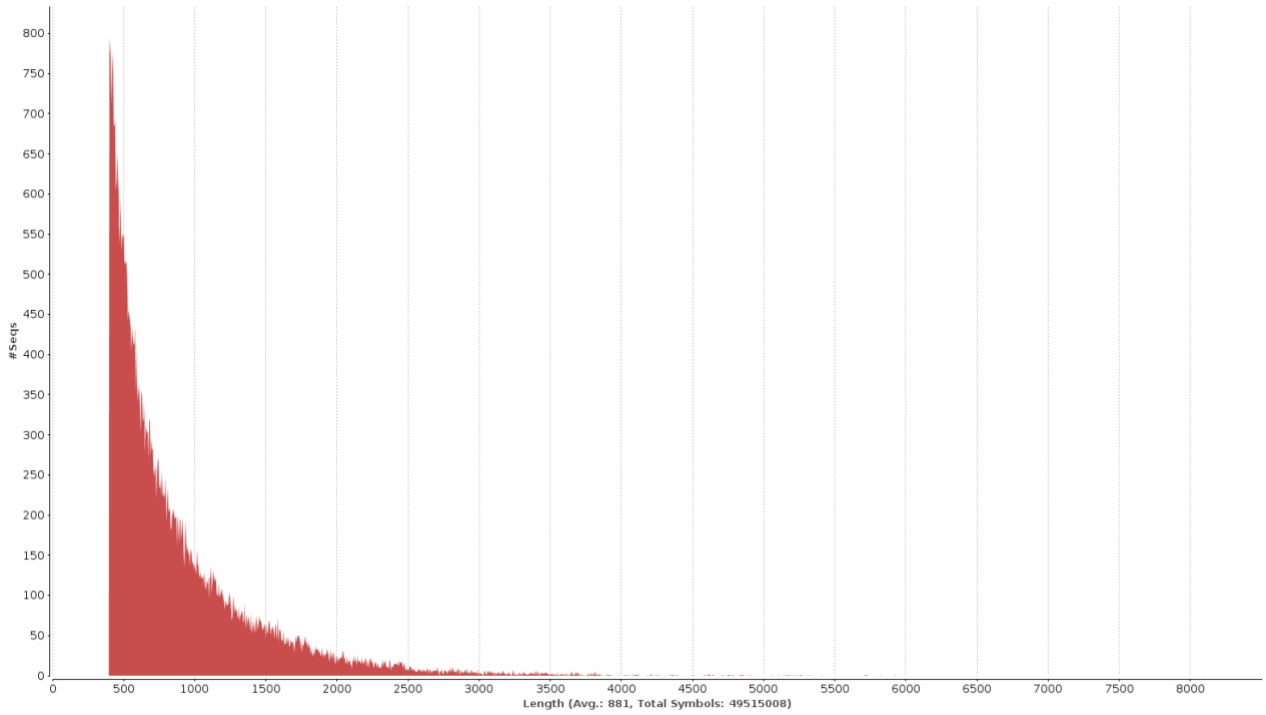

Figure 3. Contig distribution by length, post-filter.

The results from the BUSCO analysis showed that there was ~50% completeness of the transcriptome for *P. jeffreysii*. However, as there was also ~26% of the transcriptome that was identified as fragmented, the transcriptome could be considered at least ~70-75% complete.

Table 3. BUSCO results for *P. jeffreysii*. The whole transcriptome assembly was run against the Metazoan database and the Eukaryota database.

	Metazoa Database (Total BUSCOs: 978)		**Eukaryota Database (Total BUSCOs: 303)**	
	#	**%**	**#**	**%**
Complete BUSCOs	150	49.5%	508	51.9%
Fragmented BUSCOs	80	26.4%	252	25.8%
Missing BUSCOs	73	24.1%	218	22.3%

Functional Annotation and Visualization of Characterization Output

For the annotation step, an E-value cutoff of $1e^{-3}$ was used; this is considered a standard

and acceptable but liberal threshold (De Wit et al., 2012). The E-value distribution for *P. jeffreysii* shows that a large number of BLAST hits were below the E-value of $\leq 1e^{-180}$ (Figure 4). In total, there are 21,380 number of genes with a GO annotation (Figure 5). The results from merging of the two annotation methods can be quantitatively visualized below (Figure 6). Specifically, there were a total of 62,616 mapped GO terms prior to the InterProScan (IPS) merge and 76,104 GO terms after the merge. 62,296 terms were confirmed, while 10,638 terms were considered too general (i.e. too broad of a GO term) and were not included.

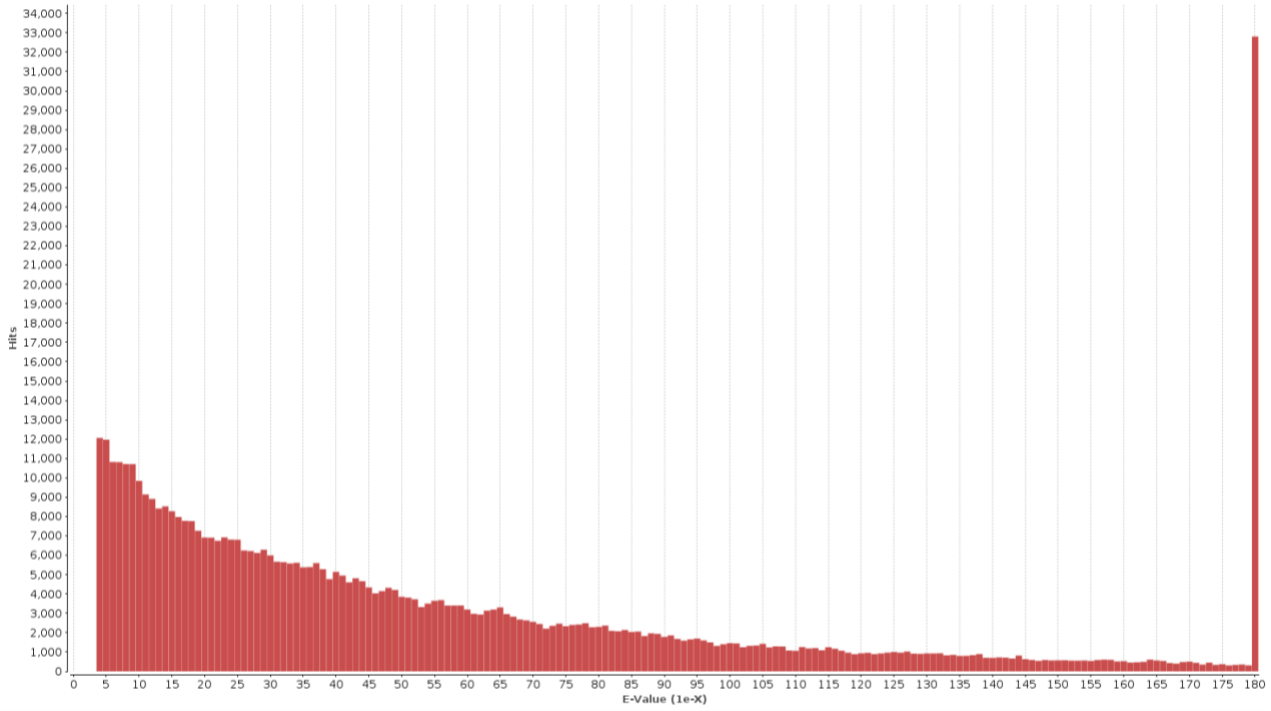

Figure 4. E-value distribution of *P. jeffreysii* transcriptome, post-filter.

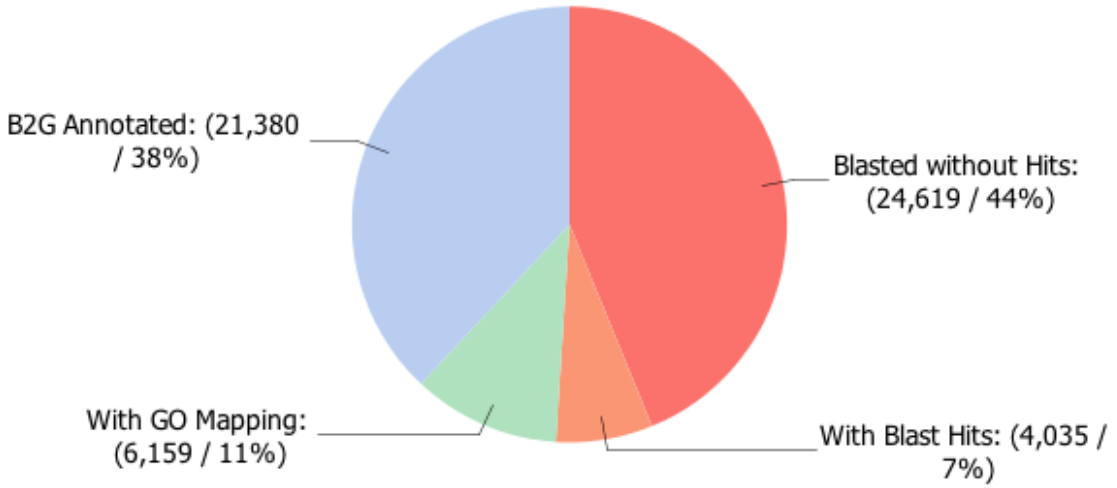

Figure 5. Data distribution of transcripts. “With Blast hits” represents the number of successful sequences after BLAST. “With GO mapping” represents the number of successful sequences with after Mapping step. “B2G Annotated” represents the number of successfully annotated sequences.

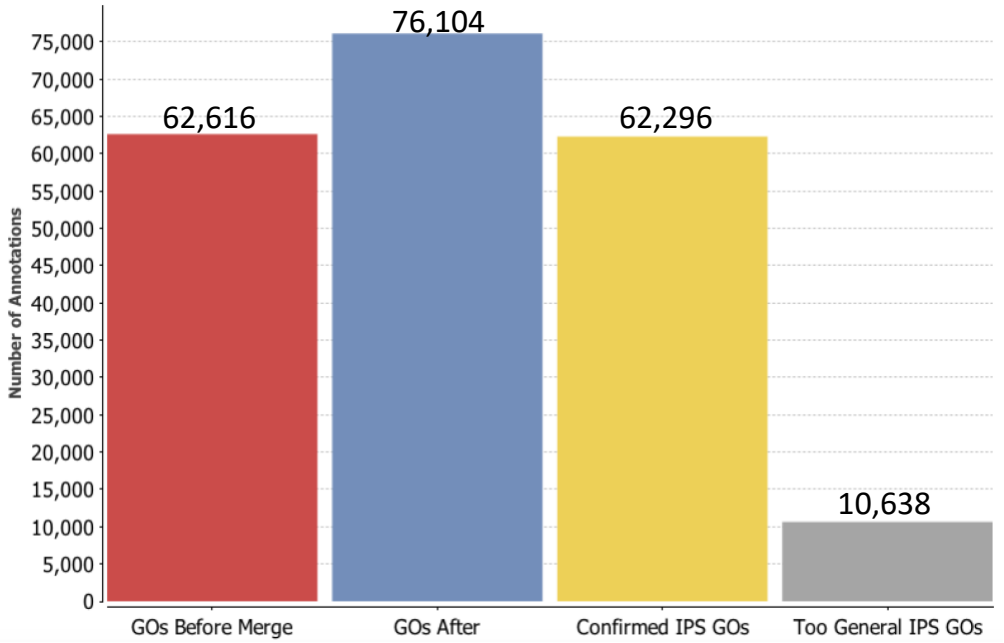

Figure 6. Number of gene ontology terms before and after InterPro Scan (IPS) merging with GO annotations.

Twenty-nine species resulted as top-hits to *P. jeffreysii*, with a brachiopod, an annelid, a chordate, and seven mollusks among the top ten hits (Figure 6). There were 24,935 total GO terms that were mapped with *Lingula anatina*, a brachiopod, as the top hit (6,269, or 25%); and the annelid *Capitella teleta* (5,968, or 24%) as the second. Of the twenty-nine top-hit species, mollusks accounted for nine (7,486 or 30%). Annelida accounted for four of the top-hit species (6,954, or 28%), but only one was among the top 10 hits. Other representatives included Cnidaria (five species – 833, or 3%); Echinodermata (three species – 649, or 3%); Chordata and Arthropoda (both with two species represented – 1,318, or 5% and 487, or 2%, respectively), and Hemichordata (594, or 2%), Priapulida (211, or 0.8%), and Platyhelminthes (134, or 0.5%) (all with one represented species). Fourteen of the twenty-nine taxa belonged to Lophotrochozoa, accounting for 20,709, or 83% (nine mollusks, four annelids, and one brachiopod). The lack of fungal, protists, or members of the microbial community confirmed there was no contamination in the dataset.

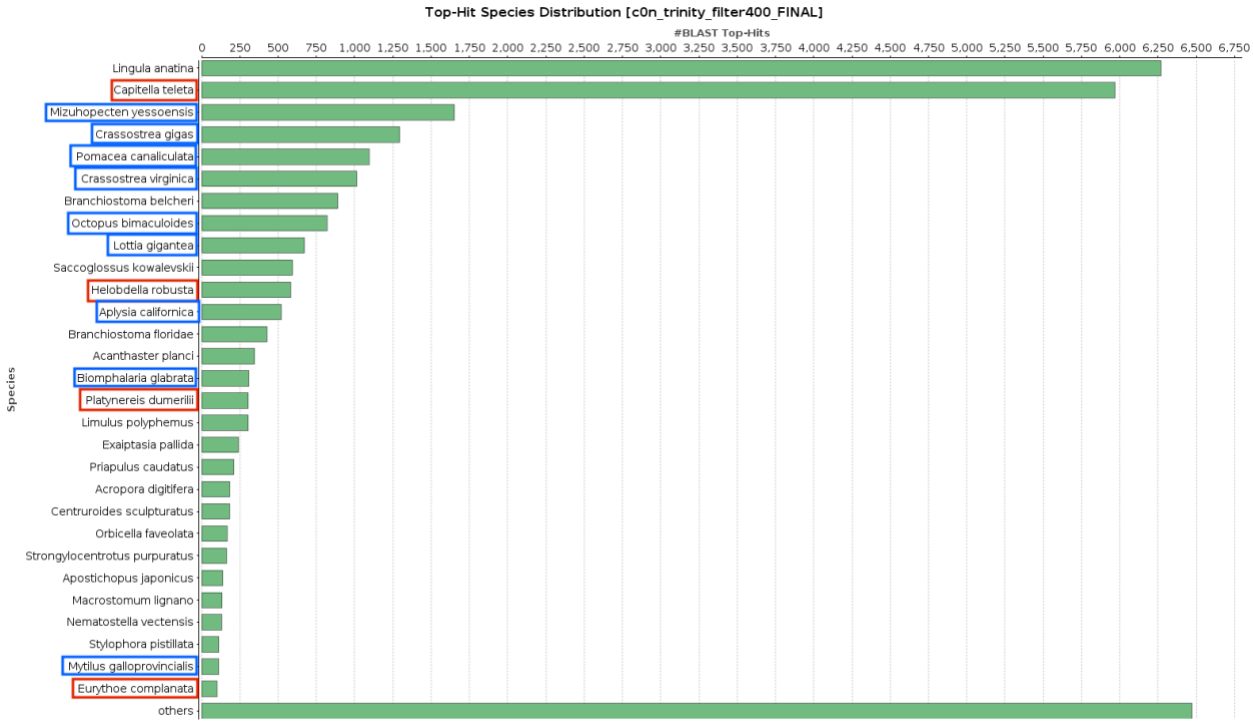

Figure 6. Top-hit species distribution of transcripts after assembly was filtered for contigs >400 bp. The highlighted species are representatives from Annelida (red) and Mollusca (blue), showcasing the lack of comparable annotated data for amphinomids, and Annelida in general.

A breakdown for each GO category can be found in Figures 7-9. Gene expression (2,501, or 10%) and transport (2,256, or 9%) were the top two biological process (Total: 25,092 GO terms) (Figure 7). Protein binding was the top hit for molecular function with 5,293, or 20%, of the 25,844 GO terms that were mapped as a molecular function, while hydrolase activity (3,701, or 14%) was the second top molecular function found (Figure 8). The top hit was cytoplasmic part with 3,487, or 25%, of the 13,783 GO terms that were mapped as a cellular component, while genes in the protein-containing complex (3,215, or 23%) was the second top cellular component found (Figure 9).

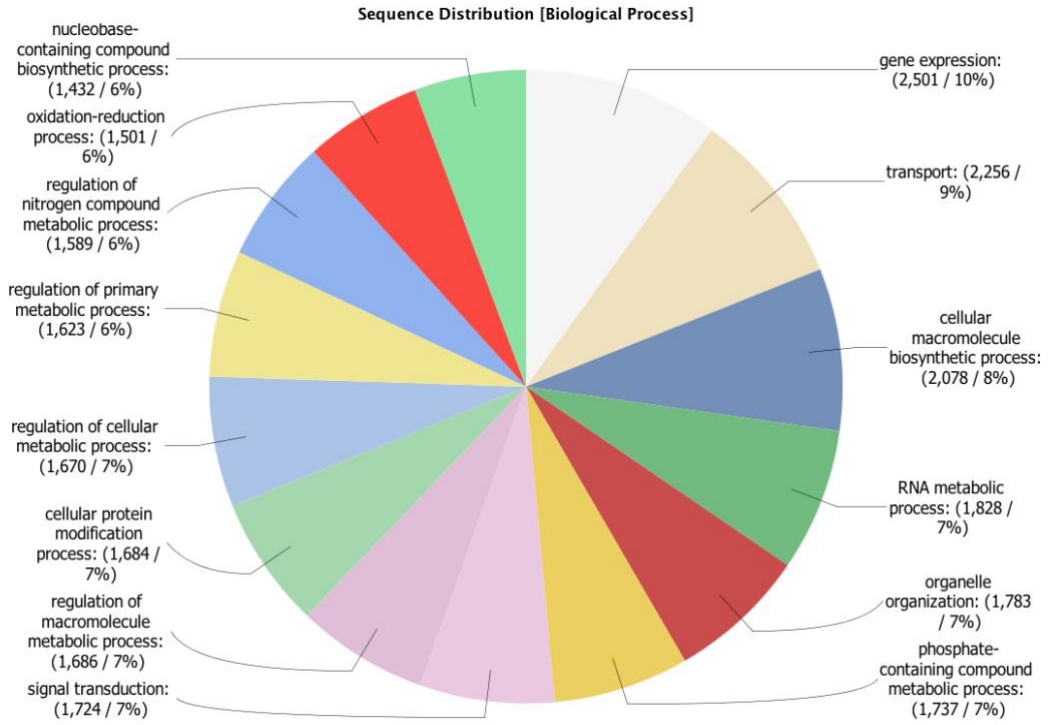

Figure 7. Sequence distribution breakdown across all Gene Ontology levels for biological processes.

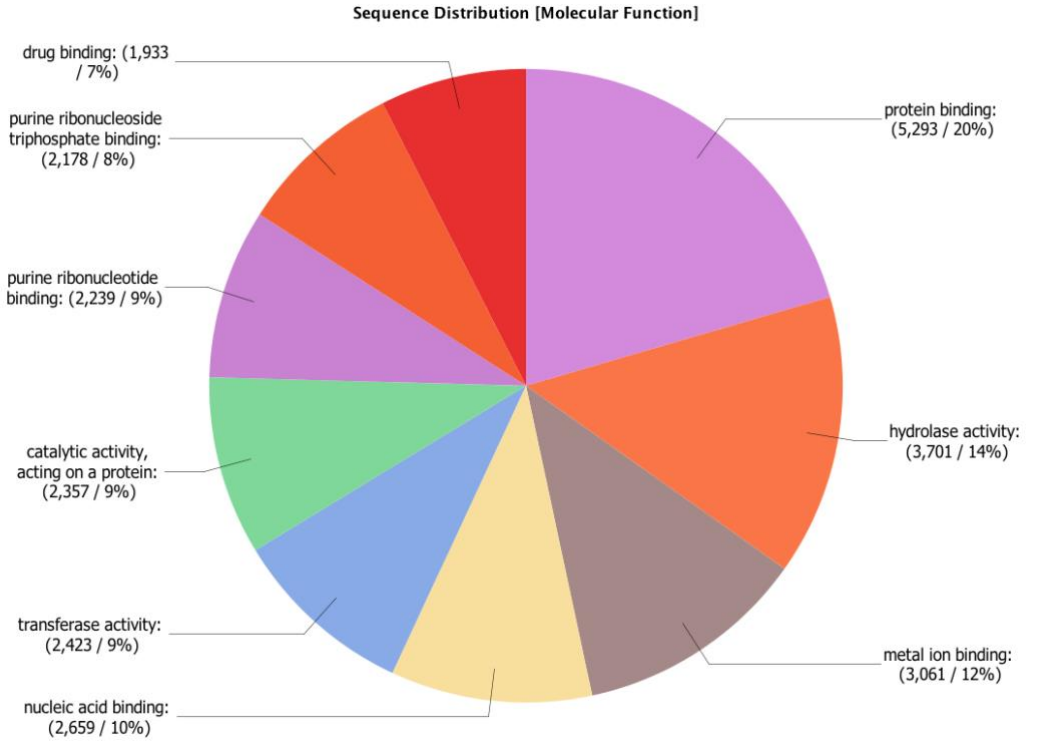

Figure 8. Sequence distribution breakdown across all Gene Ontology levels for molecular functions.

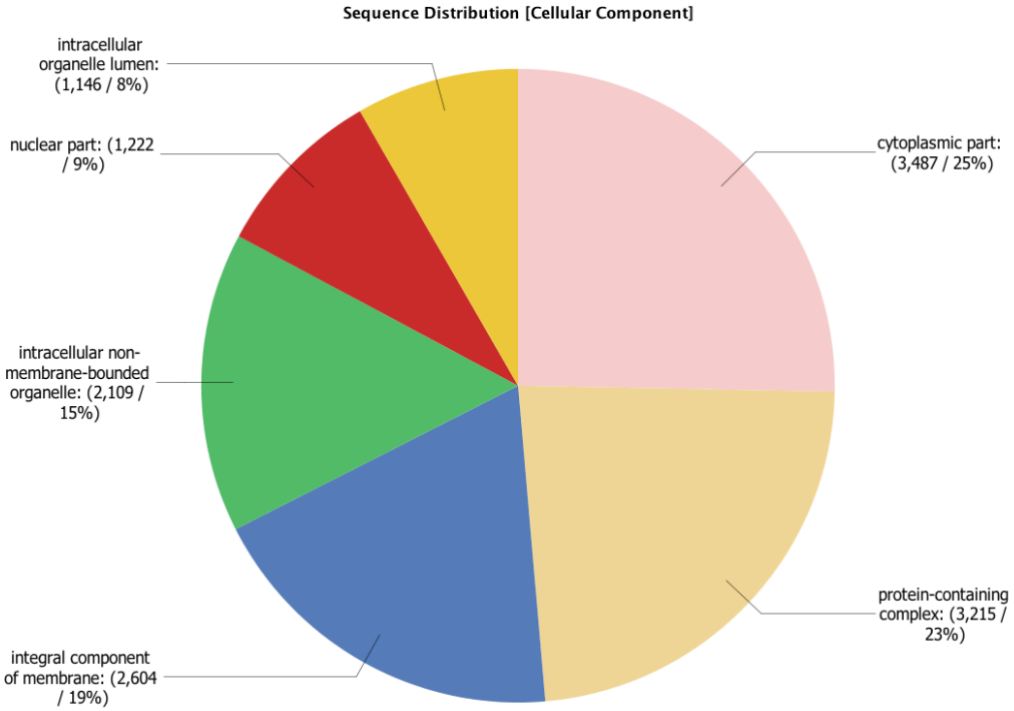

Figure 9. Sequence distribution breakdown across all Gene Ontology levels for cellular components.

Figure 10 summarizes the GO categories by biological processes (BP), molecular functions (MF), and cellular components (CC) as displayed by Level 2 (BP, MF, and CC and Level 1). Gene ontology levels are dependent on one's data and, thus, differs from research to research. Genes used for cellular processes, metabolic processes, and biological regulations were the top three biological processes found for *P. jeffreysii*. Binding and catalytic and transporter activity were the top three hits for molecular functions, while the majority of the gene activity takes place in the cell or a cell part.

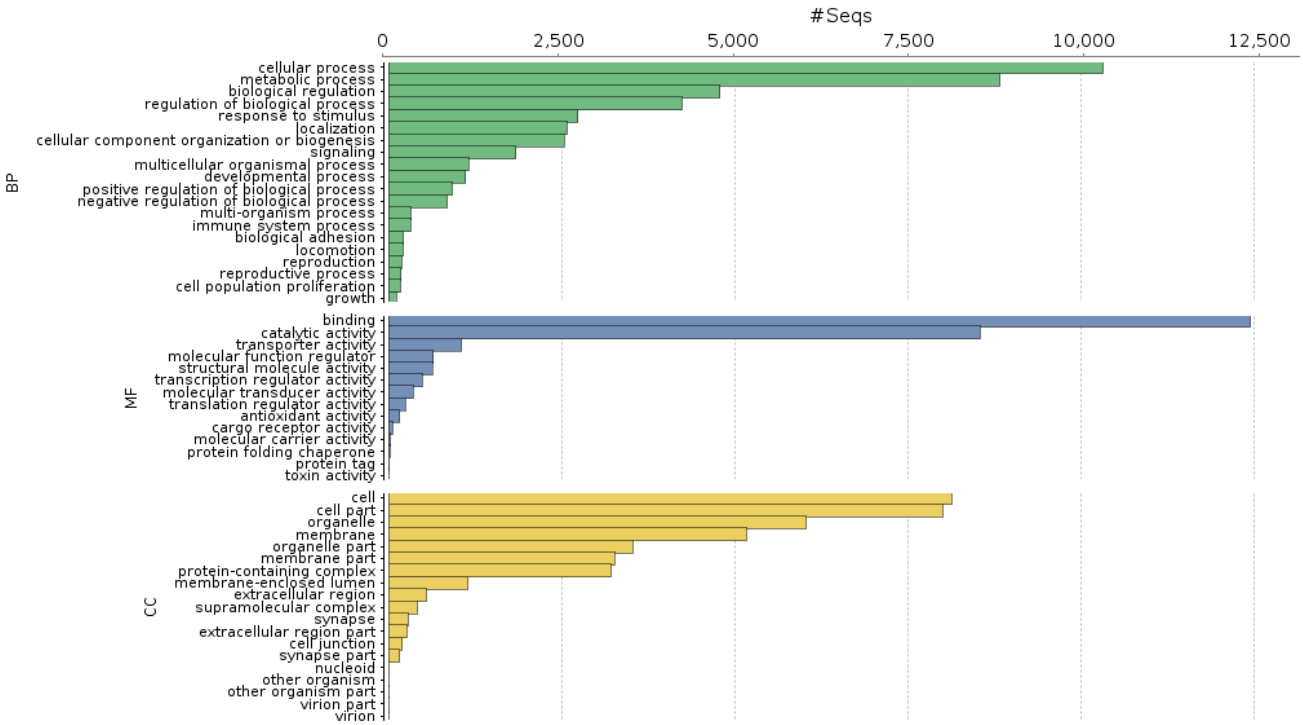

Figure 10. Summary of Gene Ontology (GO) distribution (Level 2) of biological process (green), molecular function (blue), and cellular component (yellow).

CHAPTER IV

DISCUSSION

Transcriptome of *P. jeffreysii*

Here, the publicly available transcriptome for *P. jeffreysii* was assembled and annotated. The top BLAST hits were mainly represented by a brachiopod and a distantly related annelid (>5,800) BLAST top hits), with several species of mollusks also represented (<1,700 BLAST hits). Though there are sequence annotations for *Hermodice carunculata* (Mehr et al., 2015; Verdes et al., 2017), the data is not available for comparison, thus resulting in no hits with *P. jeffreysii*. The only amphinomid among the top hits was *Eurythoe complanata* (29th on the list), which yielded ~100 BLAST hits to *P. jeffreysii*. The limited number of species hits for Annelida highlights the need for more annotated genomic/transcriptomic data in sequence databases, such as GenBank and UniProt.

When compared to the results obtained by Mehr (2015), who sequenced and annotated the transcriptome for *Hermodice carunculata*, housekeeping genes found for *P. jeffreysii* belonged to ATP synthase, while *H. carunculata* also contained phosphofructokinase (PFK) and catalase (CAT. Signaling pathways genes that were found were activin, fringe, jagged, notch, and transforming growth factor (TGF), while *H. carunculata* also contained decapentaplegic (DPP) and notch 2 (Mehr et al., 2015). *P. jeffreysii* had immune response genes, such as caspase, interleukin, toll-like receptors, interferon regulatory factors (IRF), ficolin, antistasin, angiopoietin, the same as *H. carcunculata* (Mehr et al., 2015). Reproduction genes found for both *P. jeffreysii* and *H. carunculata* were attractin, vasa, piwi, zonadhesin, zona pellucida, and spermatogenesis. *H. carunculata* additionally had smaug, nanos, and germ cell-less genes that

are also used for reproduction (Mehr et al., 2015). In Verdes (2017), venom genes were found for *P. jeffreysii* such as serpin, chitinase, metalloproteinase, serine protease, Kunitz, ubiquitin, cysteine-rich secretory proteins (CRISPs), cystatin. *Hermodice carunculata* and *Eurythoe complanata* also contained these venom genes, along with spider neurotoxins such as agatoxin, latrotoxin, and atracotoxin (Verdes et al., 2017), though these spider neurotoxins were not found for *P. jeffreysii*. No light production genes were found for *P. jeffreysii*, but were present in *H. carunculata*.

Transcriptome Quality, Assembly, and Completeness

To improve the assembly statistics for *P. jeffreysii*, contigs shorter than 400 bp were removed due to significant sequence similarity being greatly dependent on the length of the query sequence. Thus, shorter contigs are less likely to have matches against the NR database. The increase in mean contig length between the unfiltered file and the file that contained contigs greater than 400 bp (from 589.45 bp to 881.16 bp, respectively), as well as the increase in the N50 (from 758 bp to 995 bp, respectively), provided better quality results, as any increase in contig length is expected to have a significant and positive effect on the quality of assemblies (Magoč & Salzberg, 2011). BUSCO analyses showed ~50% completeness of the transcriptome, with ~25% considered fragmented, representing ~75% of the transcriptome. A 100% completeness was not expected as different environmental conditions induce different gene expressions. E-values showed high similarity for matches in the NR database.

Conclusion

This transcriptomics study of a non-model organism *P. jeffreysii*, has had its limitations due to the shortage of reference genomes that are closely related, in terms of both annotation and transcriptome completeness analyses. The Sequence Read Archive (SRA) database (NCBI)

stores raw transcriptomic data that can be used to gain a better understanding of the biological pathways and molecular functions for this species, and promote continued comparative genomics research for Annelida, while improving the availability of comparable data. This study has helped to confirm findings found from previous research, such as the finding of immune response genes, as well as venom genes (Mehr et al., 2015; Verdes et al., 2017). It has additionally provided an opportunity to contribute to understanding the molecular makeup of the non-model species *P. jeffreysii*.

REFERENCES

Ahrens, J. B., Borda, E., Barroso, R., Paiva, P. C., Campbell, A. M., Wolf, A., . . . Schulze, A. (2013). The curious case of Hermodice carunculata (Annelida: Amphinomidae): evidence for genetic homogeneity throughout the Atlantic Ocean and adjacent basins. *Molecular Ecology, 22*(8), 2280-2291.

Ahrens, J. B., Kudenov, J. D., Marshall, C. D., & Schulze, A. (2014). Regeneration of posterior segments and terminal structures in the bearded fireworm, Hermodice carunculata (Annelida: Amphinomidae). *Journal of Morphology, 275*(10), 1103-1112.

Andrade, S. C., Novo, M., Kawauchi, G. Y., Worsaae, K., Pleijel, F., Giribet, G., & Rouse, G. W. (2015). Articulating "archiannelids": Phylogenomics and annelid relationships, with emphasis on meiofaunal taxa. *Molecular Biology Evolution, 32*(11), 2860-2875.

Borda, E., Kudenov, J. D., Bienhold, C., & Rouse, G. W. (2012). Towards a revised Amphinomidae (Annelida, Amphinomida): description and affinities of a new genus and species from the Nile Deep-sea Fan, Mediterranean Sea. *Zoologica Scripta, 41*(3), 307-325.

Borda, E., Kudenov, J. D., Chevaldonné, P., Blake, J. A., Desbruyères, D., Fabri, M.-C., . . . Wilson, N. G. (2013). Cryptic species of Archinome (Annelida: Amphinomida) from vents and seeps. *Proceedings of the Royal Society of London B: Biological Sciences, 280*(1770), 20131876.

Borda, E., Yáñez-Rivera, B., Ochoa, G. M., Kudenov, J. D., Sanchez-Ortiz, C., Schulze, A., & Rouse, G. W. (2015). Revamping Amphinomidae (Annelida: Amphinomida), with the inclusion of Notopygos. *Zoologica Scripta, 44*(3), 324-333.

Bushnell, B., Rood, J., & Singer, E. J. P. o. (2017). BBMerge–Accurate paired shotgun read merging via overlap. *PLoS One, 12*(10), e0185056.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. J. B. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics, 21*(18), 3674-3676.

De Wit, P., Pespeni, M. H., Ladner, J. T., Barshis, D. J., Seneca, F., Jaris, H., . . . Palumbi, S. R. (2012). The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. *Molecular Ecology Resources, 12*(6), 1058-1067.

Ferrier, D. E. (2012). Evolutionary crossroads in developmental biology: annelids. *Development, 139*(15), 2643-2653.

Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., . . . Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research, 36*(10), 3420-3435.

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

Halanych, K. M., Bacheller, J. D., Aguinaldo, A., Liva, S. M., Hillis, D. M., & Lake, J. A. (1995). Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. *Science, 267*(5204), 1641-1643.

Halanych, K. M., & Kocot, K. M. (2014). Repurposed transcriptomic data facilitate discovery of innate immunity toll-like receptor (TLR) genes across lophotrochozoa. *The Biological Bulletin, 227*(2), 201-209.

Magoč, T., & Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics, 27*(21), 2957-2963.

McHugh, D. (2000). Molecular phylogeny of the Annelida. *Canadian Journal of Zoology, 78*(11), 1873-1884.

Mehr, S., Verdes, A., DeSalle, R., Sparks, J., Pieribone, V., & Gruber, D. F. (2015). Transcriptome sequencing and annotation of the polychaete Hermodice carunculata (Annelida, Amphinomidae). *BMC Genomics, 16*(1), 445.

Mushegian, A. R. (2007). *Foundations of Comparative Genomics* (1st ed.): Elsevier.

Rousset, V., Pleijel, F., Rouse, G. W., Erséus, C., & Siddall, M. E. (2007). A molecular phylogeny of annelids. *Cladistics, 23*(1), 41-63.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. J. B. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics, 31*(19), 3210-3212.

Struck, T. H. (2013). The impact of paralogy on phylogenomic studies–a case study on annelid relationships. *PLoS One, 8*(5), e62892.

Struck, T. H., Paul, C., Hill, N., Hartmann, S., Hösel, C., Kube, M., . . . Purschke, G. (2011). Phylogenomic analyses unravel annelid evolution. *Nature, 471*(7336), 95.

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.

Verdes, A., Simpson, D., & Holford, M. (2017). Are fireworms venomous? Evidence for the convergent evolution of toxin homologs in three species of fireworms (Annelida, Amphinomidae). *Genome Biology Evolution, 10*(1), 249-268.

Weigert, A., Golombek, A., Gerth, M., Schwarz, F., Struck, T. H., & Bleidorn, C. (2016). Evolution of mitochondrial gene order in Annelida. *Molecular Phylogenetics Evolution, 94*, 196-206.

Weigert, A., Helm, C., Meyer, M., Nickel, B., Arendt, D., Hausdorf, B., . . . Bleidorn, C. (2014). Illuminating the base of the annelid tree using transcriptomics. *Molecular Biology Evolution, 31*(6), 1391-1401.

Yáñez-Rivera, B., & Méndez, N. (2014). Regeneration in the stinging fireworm Eurythoe (Annelida): Lipid and triglyceride evaluation. *Journal of Experimental Marine Biology Ecology, 459*, 137-143.