DE NOVO ASSEMBLY OF A ZOANTHID TRANSCRIPTOME FOR THE STUDY OF CORALLICOLIDS

by

Jade Shivak

B.Sc., The University of British Columbia, 2021

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Botany)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

February 2023

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

DE NOVO ASSEMBLY OF A ZOANTHID TRANSCRIPTOME FOR THE STUDY OF CORALLICOLIDS

submitted by Jade Shivak in partial fulfilment of the requirements for

the degree of Master of Science

in Botany

Examining Committee:

Patrick Keeling, Professor, Botany, UBC
Supervisor

Laura Parfrey, Associate Professor, Zoology/Botany, UBC
Supervisory Committee Member

Brian Leander, Professor, Zoology/Botany, UBC
Supervisory Committee Member

Patrick Martone, Professor, Botany, UBC
Additional Examiner

Abstract

The discovery of widespread apicomplexan symbionts of anthozoans, known as corallicolids, has raised many questions about their evolutionary history, physiology, and ecology. However, gathering molecular data from these unicellular eukaryotes is challenging due to their small size, low abundance in host tissue, and inconsistent prevalence in host populations. This task is further complicated by the presence of contamination from host genetic material in the samples. *Parazoanthus swiftii* is a corallicolid host with a high infection rate and no dinoflagellate symbionts making it a potentially useful model for this association. The objective of my thesis is to develop baseline genetic resources for the study of *P. swiftii* and its symbionts using transcriptomics. Several *P. swiftii* transcriptomes were sequenced, assembled, and filtered, and BUSCO analysis detected 88% of proteins from a set of conserved orthologs. Although less than half of the transcripts were functionally annotated, 56% of predicted proteins were present across all of the samples. Two of the five samples had a diverging expression profile and lower relative expression of anthozoan transcripts. The dataset also contained 1059 potential apicomplexan genes, including three that were lineage-specific. This transcriptome will aide in the study of corallicolids and their host associations, as well as add to the number of genetic resources available for zoanthids.

Lay Summary

Corallicolids are unicellular organisms that live inside the guts of animals such as sea anemones, corals, and zoanthids. They are part of the phylum Apicomplexa, which is largely comprised of parasites, including the causative agent of malaria in humans. Since corallicolids were discovered, many questions have been raised about their evolutionary relationship to other apicomplexans and their symbiotic relationship with their hosts. This project involved the creation of a genetic dataset that characterized the gene expression of *P. swiftii*, a zoanthid that is known to contain corallicolids. The transcriptome contained sequences from other organisms, but was relatively complete and consistent across samples. This project will contribute to our understanding of zoanthids and help us study corallicolids because we can differentiate host sequences from corallicolid sequences.

Preface

This work is unpublished and the research did not require ethical approval. Patrick Keeling and I designed the project and outlined the research objectives. Patrick Keeling, Waldan Kwong, and Victoria Jacko-Reynolds collected, imaged, and preserved the biological samples. Victoria Jacko-Reynolds and I preformed RNA extraction and prepared the samples for sequencing at the University of British Columbia Sequencing and Bioinformatics Consortium. Waldan Kwong provided the trimmed reads and preliminary assembly. I conducted the data processing and analysis with guidance from Vojtěch Žárský, who developed scripts to parse the outputs of SSU analysis and taxonomic classification. I generated tables and figures based on the data and wrote the manuscript. Patrick Keeling, Mahara Mtawali, Gordon Lax, Corey Holt, and Vojtěch Žárský provided feedback on this manuscript.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BLAST	Basic Local Alignment Search Tool
Bp	Base pair
BUSCO	Benchmarking universal single-copy orthologs
COG	Clusters of Orthologous Genes
DNA	Deoxyribonucleic acid
EST	Expressed Sequence Tag
KEGG	Kyoto Encyclopedia of Genes and Genomes
Mb	Mega-base pair, 1 million bp
mRNA	Messenger ribonucleic acid
NCBI	National Centre for Biotechnology Information
ORF	Open reading frame
RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
SRA	Sequence Read Archive
SSU	Small subunit rRNA
TMM	Trimmed Mean of M-Values
WGS	Whole Genome Sequencing/Sequences

Acknowledgements

Thank you to Patrick Keeling for giving me the opportunity to learn and grow in a positive and supportive lab environment. I further extend my gratitude to my internal committee members, Laura Parfrey and Brian Leander, for their advice and feedback throughout the program. I am also thankful for my graduate advisor, Liang Song, and my graduate coordinator, Alice Liu, for helping me navigate the administrative components of my degree, which were often more intimidating than all the bioinformatics.

Special thanks to my lab mates at the Keeling Lab for their help and collaboration throughout the past three years, particularly Vojtěch Žárský. This project was made possible, in large part, due to his guidance and expertise. I would also like to thank Victoria Jacko-Reynolds, Waldan Kwong, Liz Cooney, and Gordon Lax for their general guidance and help. Furthermore, I acknowledge the contribution of any algae, microbes, or invertebrates who were dissected, blended, or otherwise harmed during the making of this thesis.

In addition, thank you to my family for their care and support, including my grandfather, who shared with me his love of the natural world. I appreciate my step siblings for embracing my quirks, and my step-father, who keeps a little note in his phone so he can explain my research to people. Finally, it is difficult to overstate the gratitude I have for my mother, who constantly inspires me with her strength and has been there every step of the way.

To my father, for fostering my sense of curiosity.

Chapter 1: Introduction

1.1 Corallicolids

Although coral reef ecosystems are known to support vast amounts of biodiversity in oligotrophic oceans, multicellular flora and fauna account for only a portion of the inhabitants of a reef. Bacteria, viruses, and protists play a pivotal role in these ecosystems by participating in nutrient cycling, influencing marine food webs, and causing disease. The photosymbiosis between anthozoans and their algal symbionts has been well-characterized due to its role in coral bleaching (van Oppen and Blackall 2019). During the genotyping of coral-associated algae in 2002, another eukaryotic symbiont was discovered and identified as belonging to a parasitic group of protists known as apicomplexans (Toller *et al.* 2002). This symbiont was observed at multiple disparate sites up to a depth of 1400 m via microbiome and metagenome surveys, but it was not until recently that the cells were imaged and characterized (Mathur *et al.* 2018; Šlapeta *et al.* 2013; Kwong *et al.* 2019; Vohsen *et al.* 2020). Ultimately, they were formally described as members of the new order Corallicolida (Kwong *et al.* 2021).

The phylum Apicomplexa is largely comprised of obligate intracellular parasites, including the causative agents of human diseases, such as malaria and toxoplasmosis (Votýpka *et al.* 2017). A defining feature of apicomplexans is the apical complex, which is a group of cytoskeletal and secretory structures that aide in host cell invasion (Hu *et al.* 2006). Organelles in the apical complex known as micronemes secrete adhesin proteins that interact with receptors on the host cell membrane (Boucher and Bosch 2015; Soldati *et al.* 2001). These proteins play a role in the

glideosome, which uses an actomyosin motor to pull back on bridging proteins connected to the adhesins in order to propel the cell forwards (Keeley and Soldati 2004).

Another notable characteristic common to apicomplexans is a relict plastid known as the apicoplast. The ancestors of these microbes were photosynthetic, but have since lost the ability to photosynthesize during the transition to heterotrophy. Although apicoplasts have a reduced genome, they continue to aide in metabolic processes with the help of nuclear-encoded genes by synthesizing fatty acids, iron-sulfur clusters, isoprenoids, and tetrapyrroles (McFadden and Yeh 2017).

Corallicolids share many characteristics with other apicomplexans, including the apicoplast. While the corallicolid apicoplast has yet to be imaged, molecular sequencing of the plastid genome has revealed chlorophyll biosynthesis genes under strong purifying selection (Kwong *et al.* 2019). This is unusual for a non-photosynthetic organism and suggests that these genes may serve an additional function. Additionally, it could indicate that corallicolids are in the midst of an evolutionary transition towards parasitism (Keeling *et al.* 2021). The effect of corallicolids on their hosts is still ambiguous, but one study posited that they may be opportunistic pathogens, only affecting their hosts under stressful conditions (Keeling *et al.* 2021). Recently, a preprint manuscript found significantly increased corallicolid abundance in corals with the highest heat-stress susceptibility (Bonacolta *et al.* 2022). While it is not known whether corallicolids infect any non-anthozoan hosts, a group of blood fish parasites formed a clade with corallicolids in an 18S rRNA phylogeny, suggesting that they may have more close relatives to be discovered (Hayes and Smit 2019). The phylogenetic position of corallicolids relative to other

apicomplexans is currently unresolved with only their organelles and SSU rRNA having been sequenced so far (Kwong *et al.* 2021; Vohsen *et al.* 2020).

The need for more sequencing data from corallicolids presents the challenge of gathering good-quality samples. Corallicolids are found in parasitophorous vacuoles within the host cells. Specifically, they are located in mesenterial filaments, which are white threads of tissue in the gastrovascular cavity that possess specialized cells for defense and digestion (Brusca *et al.* 2016). There is currently no method to detect corallicolids in anthozoans without sequencing data or microscopy. The symbionts are not ubiquitously present in anthozoans, even among members of the same population (Kwong *et al.* 2019). They also tend to be found in low abundance relative to the host tissue which further complicates the process of sequencing (Mathur *et al.* 2018). Using homology-based searches to identify sequences is often limited because they rely on genetic resources from related organisms, which are generally more abundant for animals than protozoa (Armengaud *et al.* 2014). Furthermore, their position as intracellular symbionts means that all samples will contain a large proportion of host sequences. The identification of host sequences is necessary to facilitate the process of differentiating host and symbiont sequences.

1.2 Golden Zoanthids

Corals, sea anemones, and zoanthids all belong to Anthozoa, a group of marine invertebrates. They maintain a polypoid body plan throughout their life cycle, meaning that they are attached to the substrate with their mouths and tentacles facing the water column (Brusca *et al.* 2016). Unlike corals, zoanthids lack calcareous skeletons for structural support, but instead form

colonies with polyps connected at the base and debris incorporated into their outer body wall (Light 2007). They feed on plankton in the water column, although many zoanthids also possess photosynthesizing algal symbionts similar to those found in coral (Rabelo *et al.* 2014). Several species of zoanthid also live on the exterior of other invertebrates, such as black coral, hydrozoans, and sponges (Reimer *et al.* 2018).

The golden zoanthid, *Parazoanthus swiftii*, is a species of colonial zoanthid commonly found in the Caribbean Sea and Atlantic Ocean (Montenegro 2020; Reimer *et al.* 2017). They have a polyp height of 1-3 mm and colonies tend to grow in linear formations or small clusters (Duchassaing and Michelotti 1860; Swain and Wulff 2007). Like many other zoanthids, *P. swiftii* is epibiotic, meaning that it lives on the surface of other organisms. They are generally associated with one of 46 species of host sponge, but have also been observed on octocorals or non-living substrates (Vaga *et al.* 2020; Crocker and Reiswig 1981; Swain and Wulff 2007). This broad range of hosts is likely a result of the fact that *P. swiftii* is less deeply embedded in the host tissue than other sponge-associated zoanthids and has the unique ability to spread from its host sponge to nearby substrates (Duchassaing and Michelotti 1860; Crocker and Reiswig 1981). *P. swiftii* produces toxins and uses its bright orange colouring to alert potential predators to its presence (West 1976). As a result, the sponge benefits from decreased predation and the zoanthid gains a substrate with little competition. Sponges draw water through pores on their surface during filter feeding and zoanthids can take advantage of the resulting water currents by feeding on the zooplankton they convey (Montenegro-González and Acosta 2010; Roopnarine and Hertog 2012). One study of the Water Factory study site in Curaçao found that *P. swiftii*

associated with the host sponge *Topsentia ophiraphidites* were most abundant at a depth of 20-30 m (Reimer *et al.* 2018).

P. swiftii harbors a corallicolid that is closely related to *Anthozoaphila gnarlus*, one of three described corallicolid species which are largely differentiated using SSU rRNA phylogenies (Kwong *et al.* 2021). Their abundance makes them a good study system because they are easily accessible. They also lack algal symbionts, which could minimize the presence of non-metazoan contamination in samples (Reimer *et al.* 2014). Notably, preliminary studies have suggested that corallicolids are found very consistently in *P. swiftii* samples (Kwong *et al.* 2021).

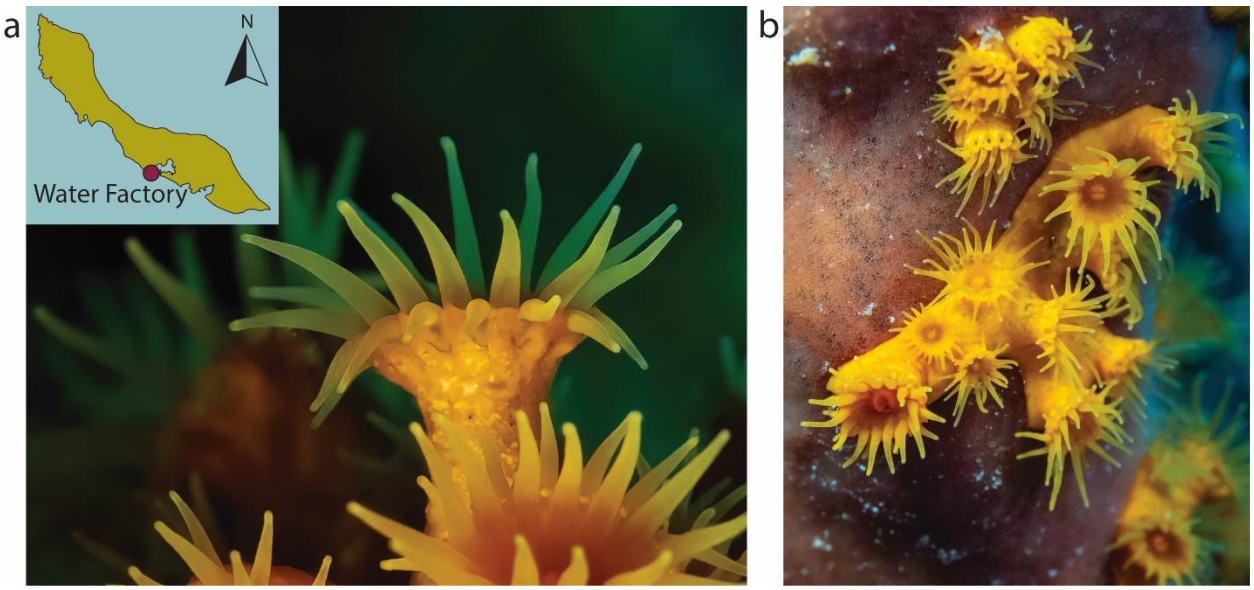

Figure 1. Golden zoanthid images and map of collection site. A) Lateral view of a golden zoanthid polyp along with an inset map of Curaçao indicating the location of the study site, Water Factory. B) Image of golden zoanthid colony on the host sponge *Topsentia ophiraphidites*.

Zoanthids are a particularly under-sampled group of cnidarians and they lack publicly available molecular data. Out of the 29,850 cnidarian experiments published on NCBI's Sequence Read Archive, only 119 are from zoanthids, whereas 22,152 are from stony corals. Out of the zoanthid datasets, none have published annotations and just four species have a published transcriptome (Agarwala *et al.* 2018). The transcriptomes for *Zoanthus natalensis* and *Palythoa variabilis* were analyzed to identify peptides related to venom and toxin production (Liao *et al.* 2019, Huang *et al.* 2016). *P. variabilis* sequences were further investigated to find prospective biopharmaceutical enzymes (Morlighem *et al.* 2018). In addition, Huang *et al.* (2017) went beyond examining protein coding genes and identified over 10 thousand long non-coding RNAs in the transcriptomes of *P. variabilis* and *Palythoa caribaeorum*. The transcriptome of *Parazoanthus axinellae* was sequenced as part of a large phylogenomic dataset, but it is relatively small and was discarded from the final analysis (Simon *et al.* 2017). While *P. swiftii* has an unannotated genome assembly available, it has no sequenced transcriptome (Agarwala *et al.* 2018).

1.3 Objective

The objective of my thesis is to assemble and annotate a *de novo* transcriptome for the corallicolid host, *P. swiftii.* To accomplish this, RNA-Seq data was collected from five *P. swiftii* polyps and combined to generate an assembly. The assembly was then taxonomically classified and filtered to create a host transcriptome, which was evaluated based on completeness. Finally, gene expression was compared between samples and the dataset was searched for corallicolid genes. The development of a *P. swiftii* transcriptome contributes to the genetic resources available for zoanthids and can be used in future studies to identify host sequences in corallicolid datasets.

Chapter 2: Methods

2.1 Sample Collection and Sequencing

Five *P. swiftii* samples were collected at the Water Factory Dive Site in Curaçao (12°06'26.2"N 68°57'00.3"W) from October 9th to 15th, 2021 (Table 1). They were located at a depth of 15 to 23 m in approximately 29ºC seawater. To preserve the tissue during transportation, whole polyps were crushed using a mortar and pestle, suspended in 50 μL of RNAlater, and stored at -20^{0}C. RNA from the samples was then precipitated, washed, and resuspended according to the TRIzol extraction protocol. The University of British Columbia Sequencing and Bioinformatics Consortium prepared the Illumina TruSeq mRNA stranded libraries, which were amplified with 15 cycles of PCR during the final step. The libraries were then sequenced on the Illumina NextSeq platform with a paired-end 150 bp read length.

Sample	Site	Isolation Date	Tissue	Paired-end Reads
Zoanthid 1	Water Factory	2021-10-09	Whole Polyp	27,756,679
Zoanthid 4	Water Factory	2021-10-10	Whole Polyp	17,575,392
Zoanthid 5	Water Factory	2021-10-15	Whole Polyp	28,971,199
Zoanthid 6	Water Factory	2021-10-14	Whole Polyp	25,438,667
Zoanthid 7	Water Factory	2021-10-14	Whole Polyp	23,502,218

Table 1. Metadata for *P. swiftii* samples collected in Curaçao. The information associated with zoanthid sample collection and library size in paired-end reads is listed for each sample.

2.2 Data Analysis

2.2.1 Assembly

The demultiplexed Illumina reads were first trimmed using Trim Galore! version 0.6.5 to remove the Universal Illumina Adapters, filter out reads with a length of less than 50 bp, and trim off base pairs with a Phred score of less than 20 (Martin 2011). Then the forward and reverse reads from all five samples were used to generate a single assembly using rnaSPAdes version 3.15.1 with k-mer lengths of 49 and 73 (Bushmanova *et al.* 2019). rnaSPAdes also calculated the length, coverage, and GC content for each transcript, which was used during taxonomic classification. The average read length, total assembly length, and number of reads mapped to the assembly were used to estimate of the transcriptome's overall base coverage with the Lander/Waterman equation (Lander & Waterman 1988).

2.2.2 Annotation

SSU rRNA sequences were identified in the transcripts using Infernals's cmscan program with the Rfam database (Madeira *et al.* 2022; Kalvari *et al.* 2021). Then, SSU sequences were selected and searched against the SILVA SSU database version 138.1 using BLASTN (Quast *et al.* 2013). The results of this search were imported into R statistical software version 4.2.1 for plotting (R Core Team 2017).

Likely protein-coding sequences over 50 amino acids long were obtained from the assembly using Transdecoder to predict open reading frames (Haas *et al.* 2013, Haas 2021). They were evaluated based on their identities in a BLASTP search (e-value 1e-3) against the UniProt reference proteomes and a conserved domain search against the Pfam database using hmmscan

(Batemen *et al.* 2023; Mistry *et al.* 2021; Madeira *et al.* 2022). Open reading frames that overlapped with the previously identified SSU rRNA genes were removed.

Proteins were then classified based on BLASTP searches and anything with a non-anthozoan best hit was filtered out (McGinnis *et al.* 2004). Unidentified sequences were kept if they had a GC-content that fell between the 10^{th} (0.38) and 90^{th} (0.50) percentiles of the identified anthozoan genes. Functional annotations were also generated for each transcript using EggNOG-mapper, including KEGG and COG categories (Cantalapiedra *et al.* 2021). The filtered set of host proteins were assessed for completeness using a BUSCO version 5.2.2 search against the metazoan_odb10 database with 954 orthologs (Manni *et al.* 2021). To compare the *de novo* transcriptome's completeness to other cnidarian datasets, BUSCO scores for free-living Cnidarian proteomes were recorded from the UniProt database (Bateman *et al.* 2023). Genes that were identified as apicomplexan during the classification step or hit to an apicomplexan gene during functional annotation were compiled and investigated separately. There was a highly-expressed ubiquitin transcript that was an outlier and the taxonomy was relatively unclear, so it was removed from the apicomplexan dataset before plotting.

2.2.3 Expression Analysis

Each individual transcriptome had its reads aligned to the assembly with bwa and indexed with SAMtools (Li *et al.* 2009). The resulting bam files were then imported into Rsubread's featureCounts function in R alongside the annotations for each peptide (Liao *et al.* 2019). EdgeR was used to apply TMM normalization factors, normalize by library sizes, and identify genes that had low expression levels across all samples (Robinson *et al.* 2010). The multidimensional

scaling plot was created with the limma package's plotMDS function using the normalized dataset with a pairwise comparison method (Ritchie *et al.* 2015). The ggplot, ggVennDiagram, ggExtra, and pheatmap packages were also used during plotting (Wickham 2016; Gao 2022; Attali and Baker 2022; Kolde 2019).

2.2.4 Characterizing Available Zoanthid Resources

To gather data on zoanthid genetic resources, metadata for RNA-Seq and WGS datasets for organisms in the order Zoantharia was obtained from the SRA Run Selector on NCBI (Agarwala *et al.* 2018). Any organellar genomes or datasets using non-random genomic sequence capture methods, such as selection for Ultra-Conserved Elements, were excluded. Each genome and transcriptome with an assembly available was queried with corallicolid SSU rRNA sequences to detect corallicolid presence. Three corallicolid sequences were used: *Corallicolida aquarius* sequence MH304758.1, *Anthozoaphila gnarlus* MW192642.1, and Apicomplexa sp. Type N AF238264.1. Sequences with a 90% or greater identity were then searched against the whole NCBI nucleotide database. If the sequence had a top result matching corallicolid SSU rRNA sequences, it was recorded for the table.

Chapter 3: Results and Discussion

3.1 *P. swiftii* Assembly is Similar to Other Zoanthid Assemblies

RNA sequencing was performed on five *P. swiftii* samples, producing a total of 123 million read-pairs with an average read length of 148 bp. The reads were then combined and assembled into a 155 Mb assembly with 245,367 contigs (Table 2). The number of reads and contigs are on the same order of magnitude as most other available zoanthid transcriptomes (Table 2). The average contig length in the *P. swiftii* assembly was 631 bp and the N50 value is 926 bp. While the *P. axinellae* transcriptome had far fewer reads and contigs, it was generated using EST rather than RNA-Seq and was discarded in the original study due to low completeness (Table 2, Simon *et al.* 2017). 99.1 % of the raw reads mapped back to the *de novo P. swiftii* assembly, resulting in an overall base coverage of 227x. A higher proportion of reads were mapped compared to the other zoanthid transcriptomes (Table 2). While this could be the result of differences in filtering steps, a high proportion of mapped reads supports the quality of the assembly (Raghavan *et al.* 2022). Overall, the transcriptome was broadly similar to the transcriptomes of related organisms.

Study	Current study	Simon *et al.* 2017	Huang *et al.* 2016	Huang *et al*. 2017	Liao *et al*. 2019
organism	*Parazoanthus swiftii*	*Parazoanthus axinellae*	*Palythoa variabilis*	*Palythoa caribaeorum*	*Zoanthus* sp. *natalensis*
assay type	RNA-Seq	EST	RNA-Seq	RNA-Seq	RNA-Seq
reads	123 Million	15,564	68 Million	118 Million	375 Million
contigs	245,367	10,190	276,526	136,654	225,236
mean contig length	631	NA	395	874	697
% reads mapped	99.1	NA	87.22	89.63	>80

Table 2. Transcriptome assembly statistics for *P. swiftii* and other zoanthids. The study is listed for each organism along with the type of assay. The number of reads (single-end) or read pairs (paired-end), number of contigs, mean contig length in base pairs, and percentage of reads mapped back to the assembly are also listed for each dataset.

3.2 SSU Sequences Reveal Taxonomic Diversity

Ribosomal RNA (rRNA) is highly conserved across cellular life and widely used to identify organisms. Before delving into a broad classification of the contigs, SSU rRNA sequences were isolated from the *P. swiftii* transcriptome in order to gather rough estimates for the taxonomic composition of the samples. There were 4919 rRNA sequences recovered from the search, including 1758 from prokaryotes, 2640 from anthozoans, and 521 from other eukaryotes (Figure 2). Despite the abundance of unique sequences, the number of unique taxonomic assignments was only 1221 in total. Within the other eukaryotes, 20 SSU sequences were identified as sponges and 9 were identified as corallicolids (Figure 2).

The presence of bacteria and other eukaryotes is to be expected because whole polyps were ground up in the samples. As a result, RNA from any organism living in or on them would be sequenced along with the tissue. Despite this, the majority of anthozoan SSU sequences had higher coverage than other organisms (Figure 2). Compared with the *P. variabilis* transcriptome,

this dataset has far more species identified (Morlighem *et al.* 2018). However, the *P. variabilis* tissue was cut up and washed with distilled water, so the difference may be partially due to discrepancies in sample preparation (Huang *et al*. 2016).

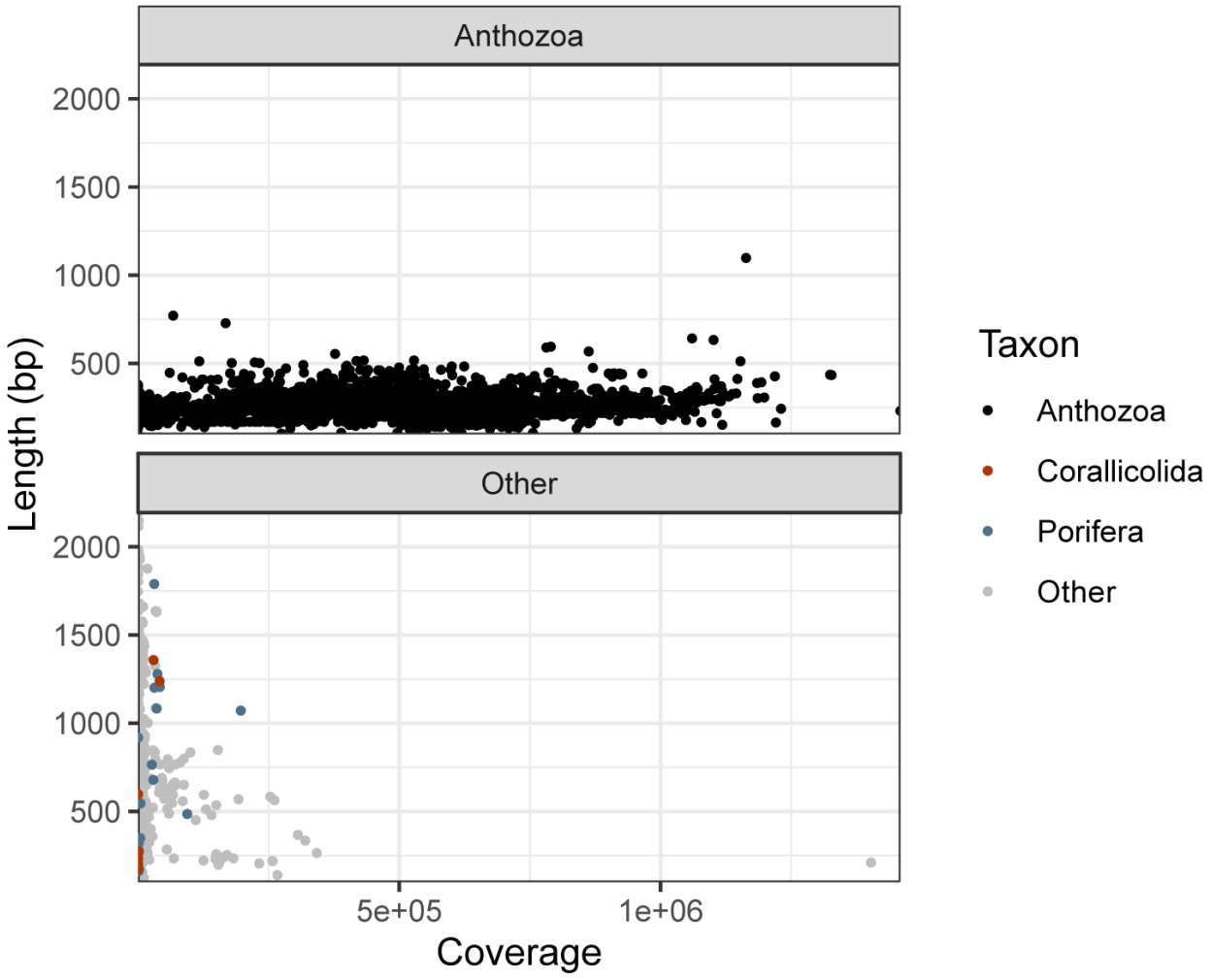

Figure 2. SSU sequences identified in the *P. swiftii* transcriptome. rRNA sequence length in bp and coverage is plotted, with taxa represented by different colours. Corallicolida and Porifera points were brought to the front to avoid being covered by taxa marked as 'Other'.

3.3 Taxonomic Classification was Used to Identify Anthozoan Genes

Following rRNA analysis, the open reading frames of protein-coding transcripts were predicted. Potential open reading frames in the assembly were evaluated against protein and conserved domain databases. There were 199,428 predicted proteins with a mean length of 127 amino acids. This relatively short length is the consequence of having selected a short minimum ORF length of 50 amino acids during prediction. The proteins were taxonomically classified and filtered to create a subset of probable host genes for downstream analysis. Only 11% of proteins could be classified and of these, 53% were eumetazoan, 2.5% were poriferan, and 0.62% were apicomplexan (Figure 3).

Although not all sequences were assigned a taxonomic classification, the predicted proteins classified as anthozoan were used to determine the range of GC-content of the host. Since GC-content has been shown to vary between different taxa (Šmarda *et al.* 2014; Romiguier *et al.* 2010; Wu *et al.* 2010), unidentified sequences that fell within the 10^{th} (0.38) and 90^{th} (0.50) percentiles of known anthozoan sequences were included in the filtered host dataset. The sequences falling outside of this range were discarded along with any non-anthozoan identified sequences. This resulted in 42% of the total sequences being removed to create the filtered host dataset.

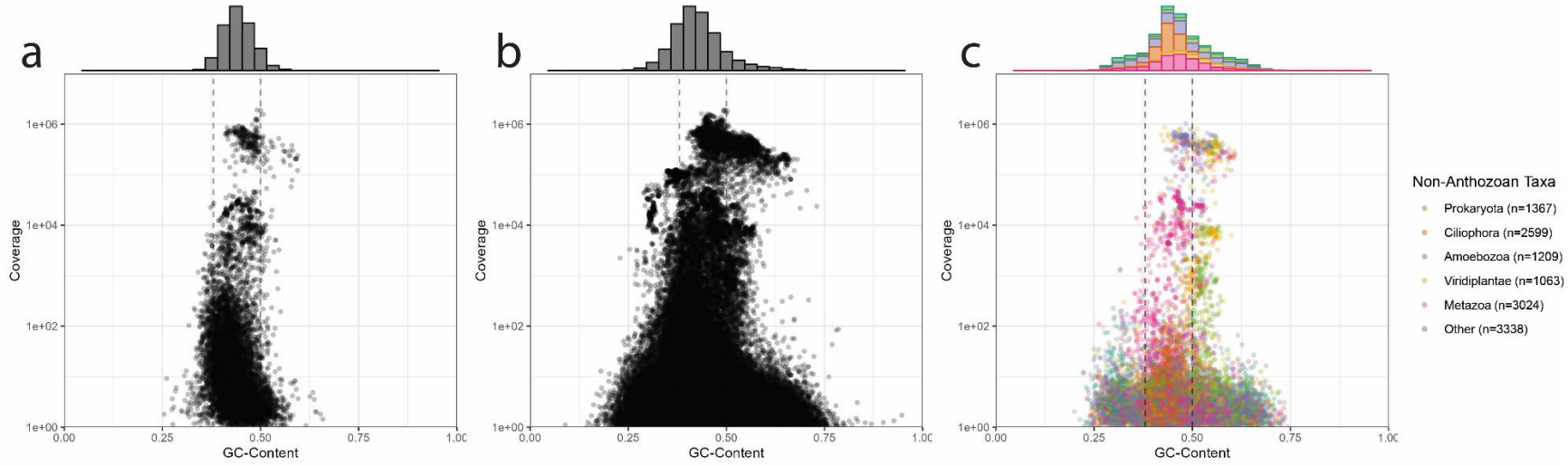

Figure 3. Taxonomic classification of the *P. swiftii* transcriptome. Each plot shows the coverage and GC-Content of transcripts in the *P. swiftii* assembly. A histogram is included at the top of each plot to illustrate the point density at a given GC value. Vertical dashed lines mark the 10th and 90th percentiles of anthozoan GC-content (0.50 and 0.38). The plots contain a) Anthozoan-identified sequences, b) unidentified sequences, and c) non-anthozoan identified sequences. Non-anthozoan sequences are coloured by taxonomic group and the size of each group is listed in the legend.

When filtering transcripts from *de novo* assemblies, there is a challenging tradeoff between the removal of extraneous sequences and the retention of desired sequences (Freedman *et al.* 2021). Although the filtered host dataset was only used to examine gene expression and completeness, the GC-content filtration method has several limitations. Firstly, assuming that the classified sequences are representative of the unclassified sequences, the method would remove the 20% of unidentified host genes which lie outside the 10^{th} and 90^{th} percentiles of GC-content. While GC-content can vary between organisms, phylogenetic relationships are not the sole predictors of GC-content (Lightfield *et al.* 2011; Foerstner *et al.* 2005). Of the identified transcripts, most bacterial transcripts fell outside the host GC range, whereas ciliate sequences were generally inside the host GC range. Thus, it is likely that this method was better at filtering out bacterial transcripts than ciliate transcripts. In the future, it may be useful to employ more nuanced approaches in order to generate precise taxonomic subsets, such as using a cutoff for coverage to eliminate noise (Freedman *et al.* 2021). Multiple annotation methods could also be applied to taxonomic classification. For example, including both nucleotide and protein homology searches could increase the proportion of sequences classified. A previous study detected 13,240 long non-coding RNAs in the *Palythoa caribaeorum* transcriptome, so it is likely that non-coding RNA sequences are present in other zoanthid transcriptomes (Huang *et al.* 2017). Other available zoanthid transcriptomic projects did not perform taxonomic classification aside from the removal of dinoflagellate symbiont reads, and it is possible that this step is not strictly necessary (Huang *et al.* 2016; Huang *et al.* 2017; Liao *et al.* 2019). In fact, Morlighem *et al.* (2018) examined potential biopharmaceuticals from the 'holo-transcriptome' of P. variabilis, which included both the host and associated organisms.

3.4 The Completeness Estimate of the *P. swiftii* Transcriptome is 72%

One of the most common metrics for evaluating the completeness of *de novo* transcriptomes is to examine groups of orthologous genes that are expected to be present in the organism. The filtered host dataset was analyzed for completeness by observing the presence of orthologs common to all metazoans and assigning the predicted proteome a BUSCO score (Figure 4). This showed that 72% of the expected orthologs were complete, and of those only 12% were duplicated (Figure 4). UniProt reference proteomes are curated to include high-quality proteomes that represent certain taxonomic groups (The UniProt Consortium 2023). Compared with the BUSCO scores of reference cnidarian proteomes, the *de novo P. swiftii* predicted proteome had a higher proportion of fragmented proteins. This is unsurprising because the reference proteomes are the result of in-depth whole genome sequencing projects and include model organisms such as *Hydra* and *Nematostella* (Bateman *et al.* 2023; Watanabe *et al.* 2009). Nevertheless, the *P. swiftii* transcriptome was only missing 12% of the expected orthologs, which is comparable to the proportion of missing orthologs in the reference proteomes (Figure 4).

The proteins were functionally annotated following the evaluation of completeness. Out of the 199,428 predicted proteins in the unfiltered assembly, 87,989 (44%) had an ortholog in the EggNOG database. The filtered assembly contained 125, 983 sequences with 51,936 (41%) EggNOG annotations. This proportion is similar to the predicted proteome of *Zoanthus natalensis*, in which 78,658 (43%) of the 183,215 predicted proteins were functionally annotated (Liao *et al.* 2019). The most highly represented KEGG categories in the *P. swiftii* predicted proteome were enzymes, membrane trafficking proteins, and chromosome associated proteins (Table A1). The number of predicted proteins is far higher than the number of proteins in the

reference cnidarian proteomes, which range from 21,298 to 25,411 proteins (Bateman *et al.* 2023).

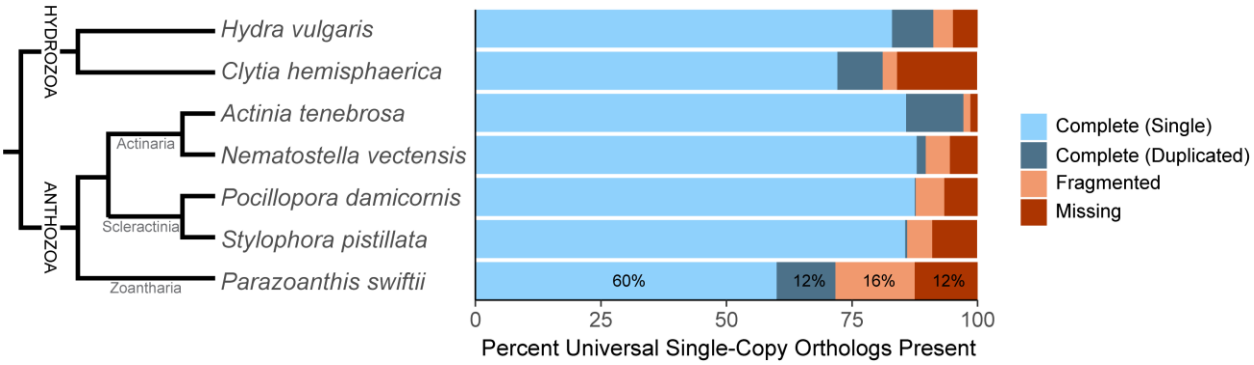

Figure 4. BUSCO comparisons between free-living cnidarian proteomes. The percentage of complete, fragmented, and missing single-copy orthologs present in each proteome are indicated on the x-axis. A schematic tree along the y-axis illustrates the taxonomic relationships between the species corresponding to each proteome. The *P. swiftii* predicted proteome was generated in this study.

There are several reasons why this *P. swiftii* predicted proteome could be larger than that of related reference proteomes. Firstly, the transcriptome was assembled using sequences from five individuals, so polymorphisms within the golden zoanthid population, alternative splicing events, and heterozygosity could result in multiple assemblies of the same gene. In addition, the settings used when predicting open reading frames allowed for proteins as short as 50 amino acids. While choosing a short cut-off may add to the number of good-quality proteins predicted, it can also introduce inaccurate predictions thus inflating the number of proteins present (Haas 2021). Finally, the presence of reads from sponges and other organisms in the filtered dataset may have led to an increased number of genes. Since the sample was filtered based on GC-content and

whole polyps were ground up during sample preparation, it is likely that some non-host proteins remained.

3.5 Most Genes were Expressed Across All Samples

The gene expression of the filtered host transcriptome highlighted the similarities between samples. 56% of genes were present across all samples, with zoanthids 1 and 6 having the highest number of unique sequences (Figure 5a). Samples 1 and 6 also clustered together on the principal component analysis (Figure 5b). This suggests a degree of homogeneity in gene expression between individuals, which is to be expected for members of the same population. The top 50 most-expressed host genes with a protein length of over 100 amino acids were compared across samples based on expression level (Figure A2). Samples 1, 5, and 6 shared their most highly expressed genes, whereas samples 4 and 7 had very divergent expression patterns. In particular, genes involved in information storage and processing had higher expression in samples 1, 5, and 6 (Figure A2).

While examining the variation between samples, it is also important to consider non-anthozoan reads. Using the sequences that were assigned during taxonomic classification, the relative proportion of each group was compared between samples in the unfiltered assembly. The proportion of genes detected from each taxonomic group was relatively consistent between samples (Figure A1). In contrast, there was a high degree of between-sample variation in relative gene expression by different taxa (Figure A1). Some of the differences between samples may be accounted for by intraspecific variation or different physiological processes. For example, the

increased proportion of non-anthozoan reads in samples 4 and 7 could be due to the presence of partially digested food in the gastrovascular cavity.

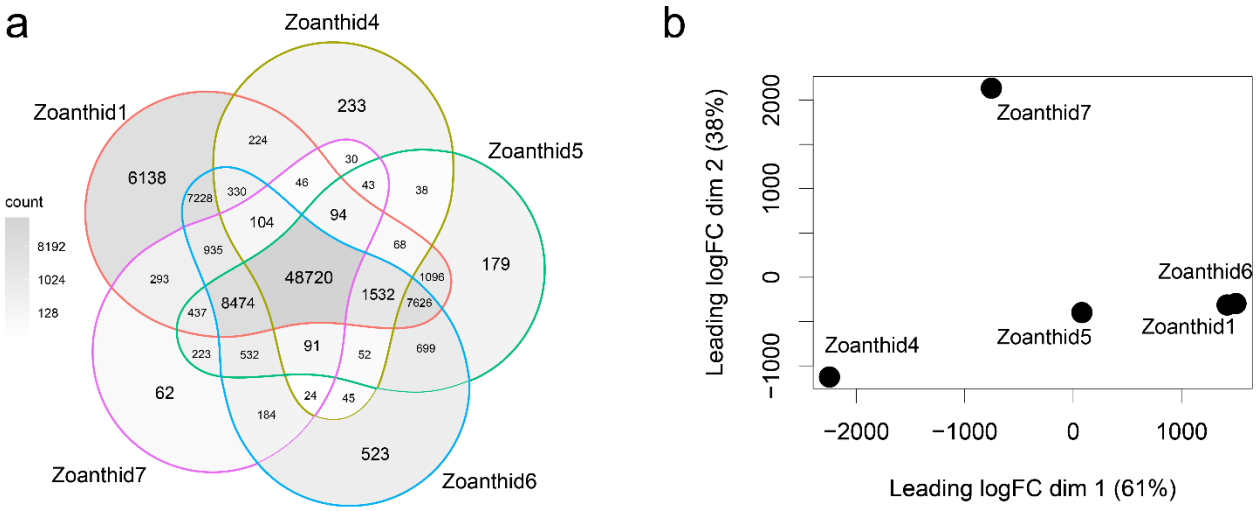

Figure 5. Examining *P. swiftii* gene expression between samples. a) The Venn Diagram represents sequences that are present in all samples overlapping in each segment. The segment shade corresponds to the number of transcripts in the segment. b) Principal Component Analysis of normalized gene expression. The log-fold change in expression between samples is shown on the first two principal components and the proportion of variation explained by each principal component is indicated on its respective axis.

3.6 Apicomplexan Genes were Discovered in the Transcriptome

In addition to the apicomplexan rRNA sequences described earlier, a number of potential apicomplexan proteins were identified during taxonomic classification and functional annotation. Taxonomy of functionally annotated sequences was taken from the top ortholog hit in the EggNOG database, whereas the taxonomic classification step identified proteins based on BLAST hits in the UniProt database. There were a total of 1059 potential apicomplexan genes, including 1030 identified during functional annotation and 142 identified during taxonomic

classification. 412 potential apicomplexan genes were detected in all five samples, which suggests that every zoanthid sampled contained corallicolids. Although the functional annotations were intended for identifying orthologs and are likely less precise for taxonomy assignment, 113 genes were common to both search methods. That being said, it is possible that some highly divergent or conserved proteins have inaccurate taxonomic assignments.

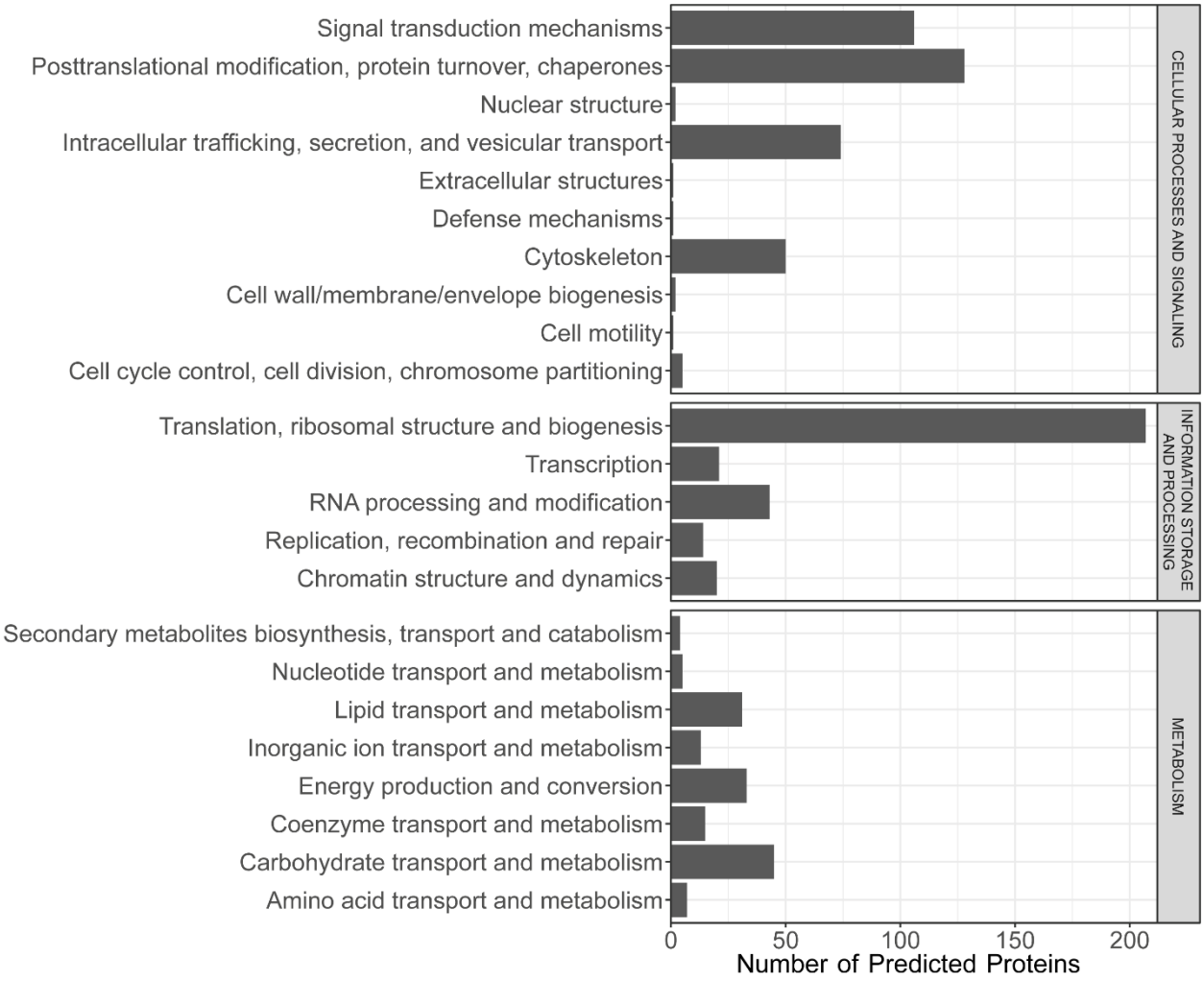

Figure 6. Apicomplexan proteins identified in the *P. swiftii* transcriptome. Bar plot of genes identified as Apicomplexan during database searches are counted by their assigned COG category. The bar length indicates the number of predicted proteins in that category.

By analyzing the KEGG categories, 54 ribosome proteins and 45 proteins related to exosomal function were detected (Table A1). The most highly represented COG category was translation, ribosomal structure, and biogenesis, followed by posttranslational modification and signal transduction mechanisms (Figure 6). There were also cytoskeletal proteins, such as actin, tubulin subunits, and dynein. Interestingly, three lineage-specific apicomplexan proteins were identified, including two microneme proteins and one glideosome-associated protein. Gliding motility has previously been observed in corallicolids and the presence of glideosome proteins indicates that some motile life stages were also present in the golden zoanthid tissue (Kwong *et al.* 2021).

3.7 Other Zoanthid Assemblies Contained Corallicolid rRNA

In order to investigate the prevalence of corallicolids in publicly available assemblies, SSU rRNA sequences from the symbionts were searched against 20 WGS assemblies and 5 transcriptome assemblies (Table 3). While corallicolid SSU rRNA sequences were absent from all of the transcriptomes, they were observed in 5 genomes (Table 3). There was one matching contig detected in each genome assembly of *Bergia* sp. *'catenularis,' Isaurus tuberculatus*, and *Palythoa grandis. Parazoanthus atlanticus* and *Parazoanthus swiftii* genomes had 4 and 3 contigs, respectively. The detection of corallicolids in the *P. swiftii* genome align with the observation of apicomplexan genes in the *de novo* transcriptome assembly. Unfortunately, the metadata associated with the genome sequences on NCBI is incomplete and the sequences have not yet been linked to a publication, so their origin is not clear. Kwong *et al.* (2019) screened 11 aquarium zoanthids for corallicolids and were able to detect them in 7 individuals, including two from the genus *Palythoa.* Despite the lower proportion of corallicolid presence here, corallicolids

were also detected in a *Palythoa* genome assembly. The difference in corallicolid prevalence may be accounted for by the previous study's use of SSU rRNA amplicon sequencing rather than the WGS and RNA-Seq techniques used to generate the zoanthid assemblies (Kwong *et al.* 2019).

		Genome			Transcriptome		
Family	**Genus**	**Absent**	**Present**	**NA**	**Absent**	**Present**	**NA**
Epizoanthidae	*Epizoanthus*	5	0	4	0	0	0
	Paleozoanthus	0	0	1	0	0	0
Hydrozoanthidae	*Hydrozoanthus*	1	0	3	0	0	0
Parazoanthidae	*Antipathozoanthus*	2	0	2	0	0	0
	Bergia	0	1	1	0	0	0
	Churabana	0	0	1	0	0	0
	Parazoanthus	1	2	1	0	0	1
	Umimayanthus	2	0	2	0	0	0
Sphenopidae	*Palythoa*	1	1	5	3	0	0
Zoanthidae	*Isaurus*	0	1	0	0	0	0
	Zoanthus	3	0	6	2	0	0
Total		15	5	26	5	0	1

Table 3. Corallicolid presence in publicly available zoanthid assemblies. Corallicolid SSU rRNA detection in genomes and transcriptomes available for Zoantharians from the NCBI Sequence Read Archive. The number of WGS and transcriptomes is shown for each genus in the database along with the number of assembled transcriptomes containing corallicolid SSU rRNA. The final column for each group shows the number of datasets that were not assembled (NA), which were excluded from the rRNA search.

Chapter 4: Conclusion and Future Directions

In this thesis I describe the *de novo* assembly and annotation of the *P. swiftii* transcriptome. Upon sequencing the RNA of whole golden zoanthid polyps, the reads were assembled into 245,367 contiguous transcripts. The assembly was similar to previous zoanthid assemblies, aside from one EST study of another *Parazoanthus* species which had a relatively low read count. By examining the rRNA present in the samples, it was revealed that the majority of high-coverage sequences belonged to anthozoans. As expected, rRNA belonging to the zoanthids' host sponge and corallicolid symbionts was also detected. In addition to the *P. swiftii* transcriptome, corallicolid rRNA was present in five of twenty zoantharian genome assemblies available on NCBI, suggesting that there is still much to learn about the distribution of corallicolids in zoanthids. Given the multiple organisms detected in the *P. swiftii* samples, the whole assembly was then taxonomically classified into anthozoan and non-anthozoan sequences based on homology protein searches. While the majority of sequences were not classified due to low similarity to reference proteomes, the identified host proteins were used to further filter the transcriptome based on GC-content.

The predicted proteome of *P. swiftii* was compared with cnidarian reference proteomes found in the UniProt database, and while the number of fragmented genes was higher in the *P. swiftii* transcriptome, it had a similar proportion of missing genes. Only 41% of the genes in the predicted proteome were functionally annotated, but 56% were expressed across all samples. 1059 apicomplexan genes were detected in the dataset, the most abundant of which were related to translation.

This study added to the limited number of available zoanthid transcriptomes and can be used in future zoanthid phylogenies. Additionally, the potential corallicolid hosts identified using NCBI assemblies could be examined to confirm the presence of symbionts. Subsequent studies will be able to apply this dataset to corallicolid transcriptomes in order to differentiate host and symbiont transcripts. Highly expressed genes, such as ribosomal genes, from both the symbiont and the host could be used as a benchmark to estimate the relative proportion of each organism in future datasets. Given the small proportion of symbionts in this bulk-sequencing transcriptome, it is likely that many of the corallicolid genes detected are among the most highly-expressed. As such, I predict that these genes will also be identified in future corallicolid sequencing projects, especially single-cell transcriptomes.

References

Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, Bolton E, Bourexis D, Brister JR, Bryant SH, Canese K, et al. 2018. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 46(D1):D8–D13. doi:10.1093/nar/gkx1095.

Armengaud J, Trapp J, Pible O, Geffard O, Chaumot A, Hartmann EM. 2014. Non-model organisms, a species endangered by proteogenomics. *J Proteomics*. 105:5–18. doi:10.1016/J.JPROT.2014.01.007.

Attali D, Baker C. 2022. ggExtra: Add Marginal Histograms to "ggplot2", and More "ggplot2" Enhancements. https://cran.r-project.org/package=ggExtra.

Bateman A, Martin M-J, Orchard S, Magrane M, Ahmad S, Alpi E, Bowler-Barnett EH, Britto R, Bye-A-Jee H, Cukura A, *et al*. 2023. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* 51(D1):D523–D531. doi:10.1093/NAR/GKAC1052.

Bonacolta AM, Miravall J, Gómez-Gras D, Ledoux J-B, López-Sendino P, Garrabou J, Massana R, Campo J del. 2022. Apicomplexans predict thermal stress mortality in the Mediterranean coral *Paramuricea clavata*. [Preprint]*bioRxiv*:2022.11.23.517658. doi:10.1101/2022.11.23.517658.

Boucher LE, Bosch J. 2015. The apicomplexan glideosome and adhesins – Structures and function. *J Struct Biol.* 190(2):93–114. doi:10.1016/J.JSB.2015.02.008.

Brusca, RC., Giribet, G., Moore, W., Shuster, SM. (2016). Invertebrates (4th ed.). Sunderland (MA): Sinauer Associates, Inc. p. 266-326

Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. 2019. rnaSPAdes: a *de novo* transcriptome assembler and its application to RNA-Seq data. *Gigascience*. 8(9):1–13. doi:10.1093/GIGASCIENCE/GIZ100.

Cantalapiedra CP, Herņandez-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol*. 38(12):5825–5829. doi:10.1093/MOLBEV/MSAB293.

Crocker LA, Reiswig HM. 1981. Host Specificity in Sponge-Encrusting Zoanthidea (Anthozoa: Zoantharia) of Barbados, West Indies. *Mar Biol*. 65:231–236.

Duchassaing P, Michelotti J. 1860. Mémoire sur les coralliaires des Antilles. *Mémoires l'Academie des Sciences de Turin*. 2(19):279–365.

Foerstner KU, von Mering C, Hooper SD, Bork P. 2005. Environments shape the nucleotide composition of genomes. *EMBO Rep*. 6(12):1208–1213. doi:10.1038/SJ.EMBOR.7400538.

Gao C-H. 2022. ggVennDiagram: A "ggplot2" Implement of Venn Diagram. https://cran.r-project.org/package=ggVennDiagram.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. *De novo* transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nat Protoc*. 8(8):1494–1512. doi:10.1038/NPROT.2013.084.

Haas BJ. 2021. TransDecoder [Internet]. Github. [cited 2023 Jan 19] Available from: https://github.com/TransDecoder/TransDecoder/wiki.

Hu K, Johnson J, Florens L, Fraunholz M, Suravajjala S, DiLullo C, Yates J, Roos DS, Murray JM. 2006. Cytoskeletal Components of an Invasion Machine—The Apical Complex of *Toxoplasma gondii*. *PLOS Pathog*. 2(2):e13. doi:10.1371/JOURNAL.PPAT.0020013.

Huang C, Morlighem JÉR, Zhou H, Lima ÉP, Gomes PB, Cai J, Lou I, Pérez CD, Lee SM, Rádis-Baptista G. 2016. The Transcriptome of the Zoanthid *Protopalythoa variabilis* (Cnidaria, Anthozoa) Predicts a Basal Repertoire of Toxin-like and Venom-Auxiliary Polypeptides. *Genome Biol Evol*. 8(9):3045. doi:10.1093/GBE/EVW204.

Huang C, Morlighem JRL, Cai J, Liao Q, Perez CD, Gomes PB, Guo M, Rádis-Baptista G, Lee SMY. 2017. Identification of long non-coding RNAs in two anthozoan species and their possible implications for coral bleaching. *Sci Reports* 2017 71. 7(1):1–18. doi:10.1038/s41598-017-02561-y.

Kalvari I, Nawrocki EP, Argasinska J, Quinones-Olvera N, Finn RD, Bateman A, Petrov AI. 2018. Non-coding RNA analysis using the Rfam database. *Curr Protoc Bioinforma.* 62(1):e51. doi:10.1002/CPBI.51.

Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, Griffiths-Jones S, Toffano-Nioche C, Gautheret D, Weinberg Z, et al. 2021. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* 49(D1):D192–D200. doi:10.1093/NAR/GKAA1047.

Keeley A, Soldati D. 2004. The glideosome: a molecular machine powering motility and host-cell invasion by Apicomplexa. *Trends Cell Biol*. 14(10):528–532. doi:10.1016/J.TCB.2004.08.002.

Keeling PJ, Mathur V, Kwong WK. 2021. Corallicolids: The elusive coral-infecting apicomplexans. *PLOS Pathog.* 17(9):e1009845. doi:10.1371/JOURNAL.PPAT.1009845.

Kolde R. 2019. pheatmap: Pretty Heatmaps. https://cran.r-project.org/package=pheatmap.

Kwong WK, del Campo J, Mathur V, Vermeij MJA, Keeling PJ. 2019. A widespread coral-infecting apicomplexan with chlorophyll biosynthesis genes. *Nat* 2019 5687750. 568(7750):103–107. doi:10.1038/s41586-019-1072-z.

Kwong WK, Irwin NAT, Mathur V, Na I, Okamoto N, Vermeij MJA, Keeling PJ. 2021. Taxonomy of the Apicomplexan Symbionts of Coral, including Corallicolida ord. nov., Reassignment of the Genus *Gemmocystis*, and Description of New Species Corallicola aquarius gen. nov. sp. nov. and *Anthozoaphila gnarlus* gen. nov. sp. nov. *J Eukaryot Microbiol*. 68(4):e12852. doi:10.1111/JEU.12852.

Lander ES, Waterman MS. 1988. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics*. 2(3):231–239. doi:10.1016/0888-7543(88)90007-9.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 25(16):2078–2079. doi:10.1093/BIOINFORMATICS/BTP352.

Liao Q, Gong G, Poon TCW, Ang IL, Lei KMK, Siu SWI, Wong CTT, Rádis-Baptista G, Lee SMY. 2019. Combined transcriptomic and proteomic analysis reveals a diversity of venom-related and toxin-like peptides expressed in the mat anemone *Zoanthus natalensis* (Cnidaria, Hexacorallia). *Arch Toxicol*. 93(6):1745–1767. doi:10.1007/S00204-019-02456-Z/.

Light SF. 2007. The Light and Smith Manual: Intertidal Invertebrates from Central California to Oregon. *University of California Press.* 4: 179:180.

Lightfield J, Fram NR, Ely B. 2011. Across Bacterial Phyla, Distantly-Related Genomes with Similar Genomic GC Content Have Similar Patterns of Amino Acid Usage. *PLoS One.* 6(3):e17677.

Madeira F, Pearce M, Tivey ARN, Basutkar P, Lee J, Edbali O, Madhusoodanan N, Kolesnikov A, Lopez R. 2022. Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Res. 50(Web Server issue): W276-W279. doi:10.1093/NAR/GKAC240.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol.* 38(10):4647–4654. doi:10.1093/MOLBEV/MSAB199.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. 17(1):10–12. doi:http://dx.doi.org/10.14806/ej.17.1.200. http://journal.embnet.org/index.php/embnetjournal/article/view/200.

Mathur V, del Campo J, Kolisko M, Keeling PJ. 2018. Global diversity and distribution of close relatives of apicomplexan parasites. *Environ Microbiol*. 20(8):2824–2833. doi:10.1111/1462-2920.14134.

McFadden GI, Yeh E. 2017. The apicoplast: now you see it, now you don't. *Int J Parasitol.* 47(2–3):137–144. doi:10.1016/J.IJPARA.2016.08.005.

McGinnis S, Madden TL. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* 32(Web Server issue). doi:10.1093/NAR/GKH435.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. 2021. Pfam: The protein families database in 2021. *Nucleic Acids Res*. 49(D1):D412–D419. doi:10.1093/NAR/GKAA913.

Montenegro-González J, Acosta A. 2010. Habitat preference of Zoantharia genera depends on host sponge morphology. *Univ Sci.* 15(2):110–121.

Montenegro J, Hoeksema BW, Santos MEA, Kise H, Reimer JD. 2020. Zoantharia (Cnidaria: Hexacorallia) of the Dutch Caribbean and One New Species of *Parazoanthus*. *Diversity* 12(5):190.

Morlighem JRL, Huang C, Liao Q, Gomes PB, Pérez CD, Prieto-da-Silva ÁR de B, Lee SMY, Rádis-Baptista G. 2018. The Holo-Transcriptome of the Zoantharian *Protopalythoa variabilis* (Cnidaria: Anthozoa): A Plentiful Source of Enzymes for Potential Application in Green Chemistry, Industrial and Pharmaceutical Biotechnology. *Mar Drugs* 16(6):207. doi:10.3390/MD16060207.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41(D1):D590–D596. doi:10.1093/NAR/GKS1219.

R Core Team. 2017. R: A Language and Environment for Statistical Computing. https://www.r-project.org/.

Rabelo EF, Rocha LL, Colares GB, Bomfim TA, Nogueira VLR, Katzenberger M, Matthews-Cascon H, Melo VMM. 2014. *Symbiodinium* diversity associated with zoanthids (Cnidaria: Hexacorallia) in Northeastern Brazil. *Symbiosis*. 64(3):105–113. doi:10.1007/S13199-014-0308-9/FIGURES/3.

Raghavan V, Kraft L, Mesny F, Rigerte L. 2022. A simple guide to de novo transcriptome assembly and annotation. *Brief Bioinform.* 23(2):1–30. doi:10.1093/BIB/BBAB563.

Reimer JD, Albinsky D, Yang SY, Lorion J. 2014. Zoanthid (Cnidaria: Anthozoa: Hexacorallia: Zoantharia) species of coral reefs in Palau. Mar Biodivers. 44(1):37–44. doi:10.1007/S12526-013-0180-5/TABLES/2

Reimer JD, Lorion J, Irei Y, Hoeksema BW, Wirtz P. 2017. Ascension Island shallow-water Zoantharia (Hexacorallia: Cnidaria) and their zooxanthellae (*Symbiodinium*). *J Mar Biol Assoc United Kingdom*. 97(4):695–703. doi:10.1017/S0025315414000654.

Reimer JD, Wee HB, García-Hernández JE, Hoeksema BW. 2018. Zoantharia (Anthozoa: Hexacorallia) abundance and associations with Porifera and Hydrozoa across a depth gradient on the west coast of Curaçao. *Systematics and Biodiversity*. 16(8):820–830. doi:10.1080/14772000.2018.1518936.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. {limma} powers differential expression analyses for {RNA}-sequencing and microarray studies. *Nucleic Acids Res.* 43(7):e47. doi:10.1093/nar/gkv007.

Robinson M, McCarthy D, Smyth D. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 26:139–140.

Romiguier J, Ranwez V, Douzery EJP, Galtier N. 2010. Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. *Genome Res.* 20(8):1001. doi:10.1101/GR.104372.109.

Roopnarine PD, Hertog R. 2013. Detailed Food Web Networks of Three Greater Antillean Coral Reef Systems: The Cayman Islands, Cuba, and Jamaica. *Dataset Pap Ecol*. 2013:1–9. doi:10.7167/2013/857470.

Simion P, Philippe H, Baurain D, Jager M, Richter DJ, Di Franco A, Roure B, Satoh N, Quéinnec É, Ereskovsky A, *et al*. 2017. A Large and Consistent Phylogenomic Dataset

Supports Sponges as the Sister Group to All Other Animals. *Curr Biol.* 27(7):958–967. doi:10.1016/j.cub.2017.02.031.

Šlapeta J, Linares MC. 2013. Combined Amplicon Pyrosequencing Assays Reveal Presence of the Apicomplexan "type-N" (cf. *Gemmocystis cylindrus*) and *Chromera velia* on the Great Barrier Reef, Australia. *PLoS One*. 8(9):e76095. doi:10.1371/JOURNAL.PONE.0076095.

Šmarda P, Bureš P, Horová L, Leitch IJ, Mucina L, Pacini E, Tichý L, Grulich V, Rotreklová O. 2014. Ecological and evolutionary significance of genomic GC content diversity in monocots. *Proc Natl Acad Sci USA.* 111(39):E4096–E4102. doi:10.1073/PNAS.1321152111/SUPPL_FILE/PNAS.1321152111.SD04.TXT.

Soldati D, Dubremetz JF, Lebrun M. 2001. Microneme proteins: structural and functional requirements to promote adhesion and invasion by the apicomplexan parasite *Toxoplasma gondii*. *Int J Parasitol*. 31(12):1293–1302. doi:10.1016/S0020-7519(01)00257-0.

Swain TD, Wulff JL. 2007. Diversity and specificity of Caribbean sponge–zoanthid symbioses: a foundation for understanding the adaptive significance of symbioses and generating hypotheses about higher-order systematics. *Biol J Linn Soc*. 92(4):695–711. doi:10.1111/J.1095-8312.2007.00861.X.

Toller WW, Rowan AR, Knowlton AN. 2002. Genetic evidence for a protozoan (phylum Apicomplexa) associated with corals of the *Montastraea annularis* species complex. *Coral Reefs* 21:143-146. doi:10.1007/s00338-002-0220-2.

Vaga CF, Santos MEA, Migotto AE, Reimer J, Kitahara M V. 2020. Octocoral-associated *Parazoanthus cf. swiftii* from the southwestern Atlantic. *Mar Biodivers*. 50(2):1–7. doi:10.1007/S12526-020-01041-3/FIGURES/3.

van Oppen, M. J. H. and Blackall, L. L. 2019. Coral microbiome dynamics, functions and design in a changing world. *Nat. Rev. Microbiol*. 179 17, 557–567 (2019).

Votýpka J, Modrý D, Obornik M, Šlapeta J, Lukeš J. 2017. Apicomplexa. Handbook of Protists Second Ed. *Springer International Publishing* :567–624. doi:10.1007/978-3-319-28149-0_20/FIGURES/24.

Watanabe H, Hoang VT, Mättner R, Holstein TW. 2009. Immortality and the base of multicellular life: Lessons from cnidarian stem cells. *Semin Cell Dev Biol*. 20(9):1114–1125. doi:10.1016/J.SEMCDB.2009.09.008.

Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York.*

Wu H, Zhang Z, Hu S, Yu J. 2012. On the molecular mechanism of GC content variation among eubacterial genomes. *Biol Direct.* 7(1):1–16. doi:10.1186/1745-6150-7-2/TABLES/6.

Appendices

Appendix A - Supplementary Material

A.1 Supplementary Figures

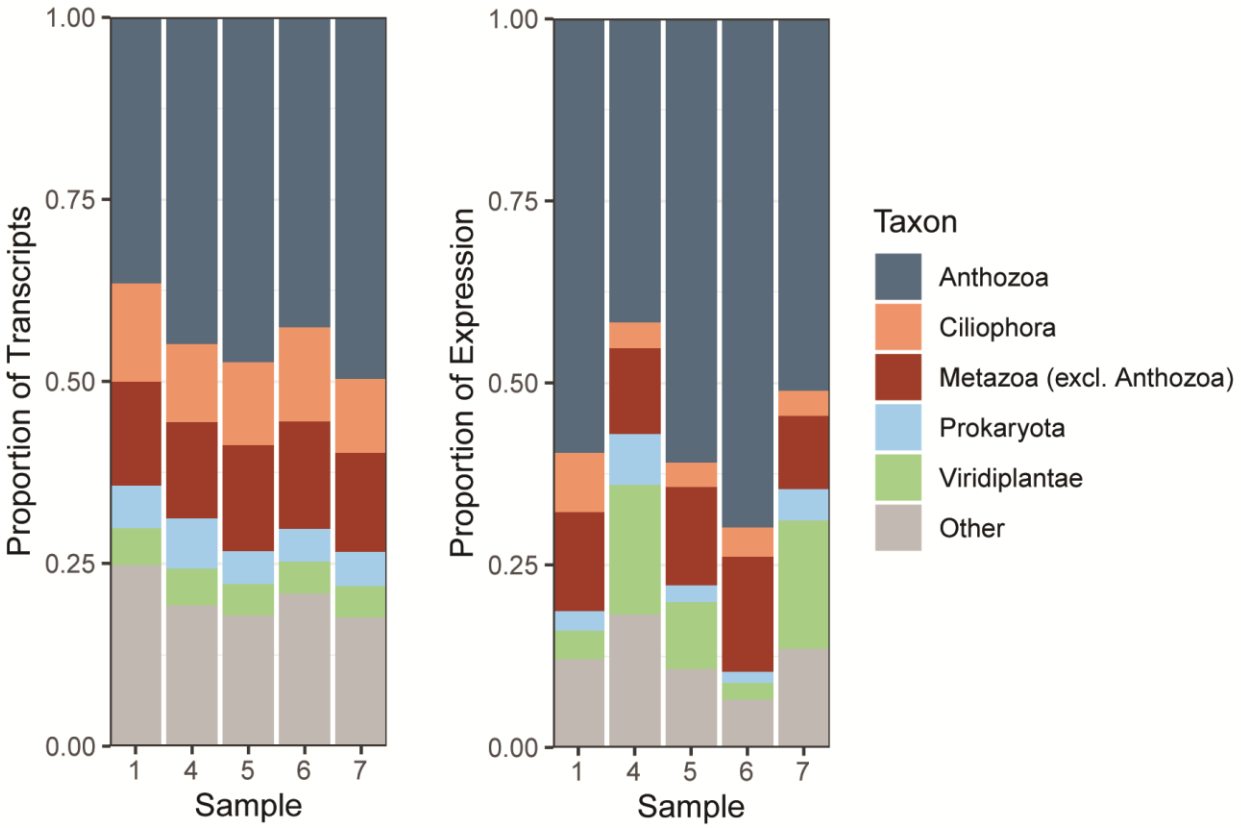

Figure A1. The proportion of taxa identified in each sample. The figure only includes the 11% of transcripts which were taxonomically classified. Bar height in the first panel indicates the proportion of the total number of transcripts detected in each sample. In the second panel, bar height shows the proportion of total expression in each sample. Different bar colours are used to represent the taxonomic groups.

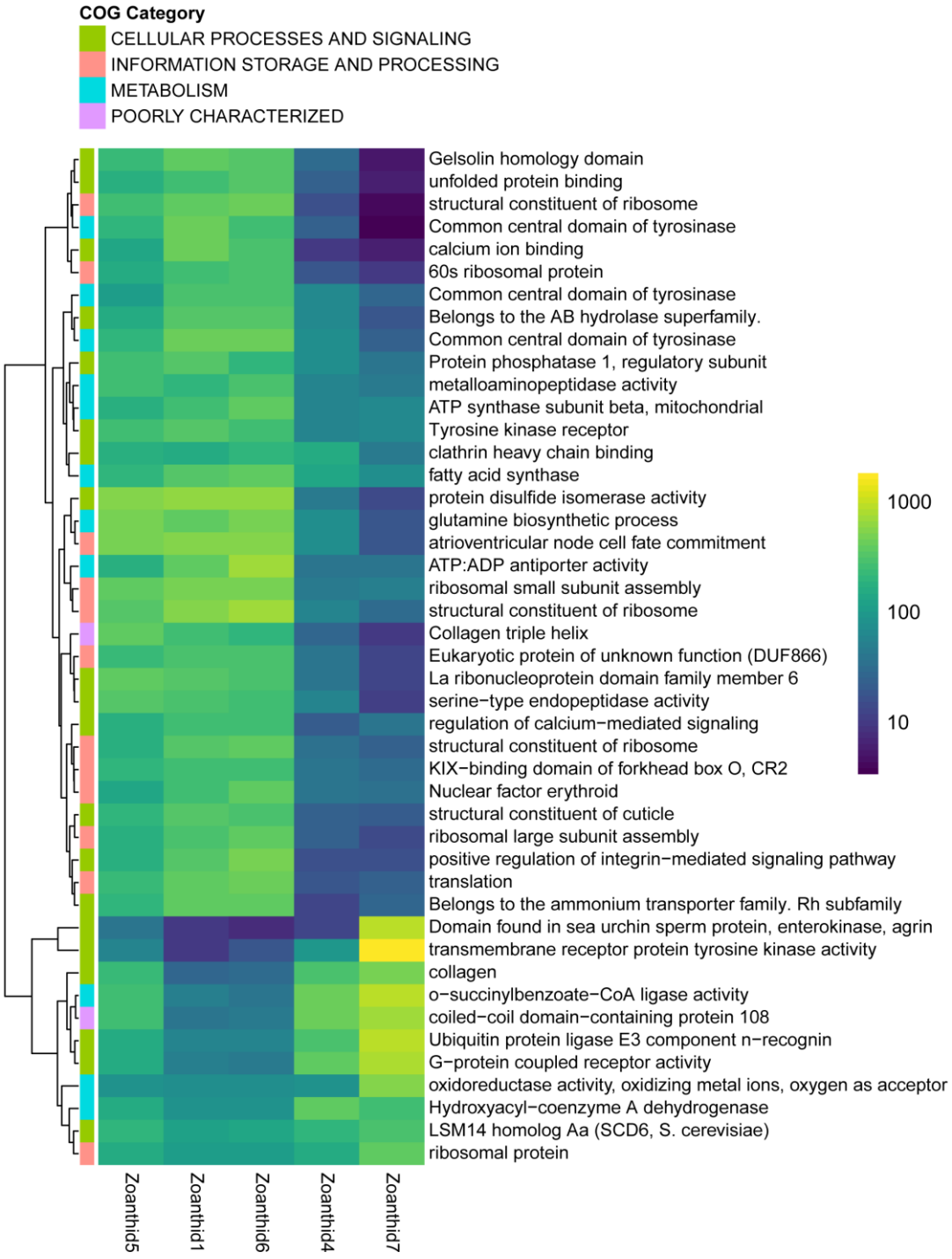

Figure A2. Golden zoanthid gene expression heatmap. Heatmap showing the top 50 most expressed genes which were identified as Anthozoan during classification and had a protein length of over 100 amino acids. The heatmap colour scale indicates the normalized expression level for each gene. Dendrograms show clusters for the y-axis. The colours for the COG group are displayed at the tips of the dendrogram.

A.2 Supplementary Tables

ID	Name	Host	Apicomplexan
ko01000	Enzymes	2231	123
ko04131	Membrane trafficking	602	44
ko03036	Chromosome and associated proteins	466	30
ko04121	Ubiquitin system	371	14
ko04147	Exosome	300	45
ko03019	Messenger RNA biogenesis	236	37
ko01002	Peptidases and inhibitors	247	11
ko02000	Transporters	233	10
ko03041	Spliceosome	215	28
ko03029	Mitochondrial biogenesis	221	15
ko01001	Protein kinases	218	9
ko03400	DNA repair and recombination proteins	209	13
ko03009	Ribosome biogenesis	202	14
ko03011	Ribosome	150	54
ko03000	Transcription factors	184	17
ko01009	Protein phosphatases and associated proteins	184	14
ko03021	Transcription machinery	170	14
ko04812	Cytoskeleton proteins	168	11
ko03037	Cilium and associated proteins	171	5
ko03110	Chaperones and folding catalysts	136	18
ko03016	Transfer RNA biogenesis	120	16
ko03032	DNA replication proteins	86	6
ko04030	G protein-coupled receptors	77	0
ko01003	Glycosyltransferases	75	2
ko04031	GTP-binding proteins	67	7
ko03051	Proteasome	55	19
ko03012	Translation factors	68	0
ko04990	Domain-containing proteins not elsewhere classified	62	0
ko04040	Ion channels	58	0
ko04090	CD molecules	55	0
ko01007	Amino acid related enzymes	40	12
ko01004	Lipid biosynthesis proteins	43	1
ko00536	Glycosaminoglycan binding proteins	35	0
ko02044	Secretion system	21	4
ko04091	Lectins	24	0
ko00199	Cytochrome P450	22	0
ko04515	Cell adhesion molecules	17	0
ko01006	Prenyltransferases	15	1
ko04052	Cytokines and growth factors	11	0
ko00535	Proteoglycans	11	0

ID	Name	Host	Apicomplexan
ko00537	Glycosylphosphatidylinositol-anchored proteins	11	0
ko04054	Pattern recognition receptors	10	0
ko04050	Cytokine receptors	8	0
ko02035	Bacterial motility proteins	7	0
ko00194	Photosynthesis proteins	7	0
ko02048	Prokaryotic defense system	5	1
ko01504	Antimicrobial resistance genes	5	0
ko02042	Bacterial toxins	5	0
ko01011	Peptidoglycan biosynthesis and degradation proteins	3	0
ko01008	Polyketide biosynthesis proteins	3	0
ko02022	Two-component system	2	0
ko03310	Nuclear receptors	1	0

Table A1. Counts of host and apicomplexan genes annotated by KEGG category. The KEGG identification number and name for each group is listed next to the number of genes identified in the host and apicomplexan subsets of the transcriptome.