Transcriptomic evidence that enigmatic parasites *Polypodium hydriforme* and Myxozoa are cnidarians

Elizabeth S. Chang

Submitted to the graduate degree program
in the Department of Ecology and Evolutionary Biology
and the Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of Master of Arts

_____

Chairperson: Dr. Paulyn Cartwright

_____

Dr. Lena Hileman

_____

Dr. Kirsten Jensen

Date Defended: May 29th, 2013

The Thesis Committee for Elizabeth S. Chang certifies that this is the approved version of the follow thesis:

Transcriptomic evidence that enigmatic parasites *Polypodium hydriforme* and Myxozoa are cnidarians

Chairperson: Dr. Paulyn Cartwright

Date accepted: May 29th, 2013

Abstract

Polypodium hydriforme and Myxozoa, represented in this study by *Myxobolus cerebralis*, are both enigmatic, intracellular parasites with very unusual life cycles and body plans, which has long made their phylogenetic placement unclear. It has been suggested that *P. hydriforme* and Myxozoa have an affinity with cnidarians because of the presence of nematocyst-like structures in both organisms. Recently phylogenomic studies have lent support to the hypothesis that Myxozoa is cnidarian. However, the placement of *P. hydriforme* and Myxozoa within cnidarian and in relation to each other remains unknown, and many questions about their evolution and transition to parasitism still remain. To address these questions, we have generated partial transcriptomes of *M. cerebralis* and *P. hydriforme*, and searched within them for important families of developmental regulatory genes and nematocyst-specific genes. The *P. hydriforme* transcriptome contained a much larger complement of both putative Hox/Parahox genes and Wnt-family genes, which may relate increased body plan complexity as compared with *M. cerebralis*. Both transcriptomes contained a number of minicollagen sequences, confirming their placement within Cnidaria.

Acknowledgements

The completion of this work would not have been possible without the experience, support and time of number of individuals. First, I would like to thank my graduate advisor, Paulyn Cartwright, for guiding me through the first years of graduate school and through the creation of my first major scientific product. I owe much of my development as a scientist to her advice and encouragement.

I would also like to thank Kirsten Jensen and Lena Hileman for taking time to serve on my committee, and providing lots of helpful input throughout the process of fulfilling all of my requirements for my master's degree. I also thank members of the Cartwright lab, past and present, for help and support. I would like to thank Annalise Nawrocki, Bastian Bentlage, Steve Saunders and Steve Davis their research advice and companionship. I would particularly like to thank Mariya Shcheglovitova for her computational expertise and for the use of some of the programs she developed, as well as the preparation of samples for sequencing. Thanks also to Amanda Shaver for her help collecting and for the use of her figure. I would also like to thank our collaborator, Dorothee Huchon, for her expertise and advice, especially about nematocust-specific programs.

I give thanks to Ron Hedrick and the University of California, Davis Veterinary School for getting samples of *M. cerebralis*. Special thanks to Brent Gordon and everyone at the Oklahoma Paddlefish Research and Processing center for helping us and putting up with us while we collected samples of *P. hydriforme*.

I would also like to thank my family in Maryland and around the country for being so supportive while I was finishing this thesis, and for constantly pursuing my career in science. I

would also like to thank my good friends in the department, particularly my roommates Logan Luce and Boryana Koseva, for going above and beyond in supporting me. Lastly, I would like to acknowledge the help and support my boyfriend Michael Pope, who had the misfortune of getting to know me right as I started getting super busy with this project.

Completion of this degree was supported by funds from the University of Kansas Department of Ecology and Evolutionary Biology. My research was funded from an NSF grant to Paulyn Cartwright.

Table of Contents

Introduction

The evolutionary transition to parasitism is often accompanied by drastic changes in body plan organization that can include degeneration or loss of prominent morphological features that are present in free-living relatives. The evolution of parasitism can also involve an increase in life cycle complexity due to the requirement of host obtainment and utilization (Phillips et al. 2012; Parker et al. 2003). Increased rates of DNA sequence and genomic evolution can accompany these dramatic changes in morphology and life cycles (Hafner et al. 1994; Hide and Isokpehi 2004; Peyretaillade 2011). The enigmatic, putative cnidarian parasites, *Polypodium hydriforme* and the phylum Myxozoa, are prime examples of extreme morphological and molecular divergences in association with the evolution of parasitism. In this study, we characterize the transcriptomes of *Polypodium hydriforme* and the myxozoan *Myxobolus cerebralis* in order to provide insight into their origins and evolution.

Life cycle of *Polypodium hydriforme*. *Polypodium hydriforme* has been described as a monotypic species from the phylum Polypodiozoa (Raikova 1994). The larvae of *P. hydriforme* spends several years as a binucleate cell-within-a-cell stage within the oocytes of female acipenseriform fishes (paddlefish and sturgeon) (Fig. 1A). Just prior to the host spawning, *P. hydriforme* develops inside the host's oocyte as a planula-type larvae (Fig. 1B). At this point, the germ layers of *Polypodium hydriforme* are reversed with respect to other metazoans, with the ectoderm developing interiorly and the gastroderm exteriorly. (Fig.1B). This larval form grows into an elongate stolon (Raikova 1994, 2008) (Fig. 1C).

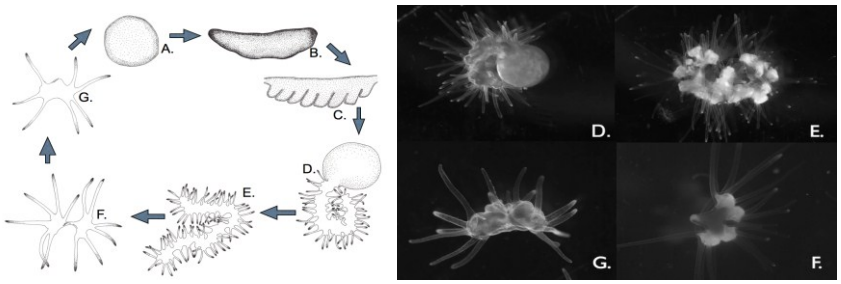


FIGURE 1: The life cycle of *Polypodium hydriforme*. A. *Polypodium* infects the host egg and has embryonic/postembryonic development in the fish oocyte. B. Planula larval form. C. Early stolon stage starting to develop tentacles. D. Stolon emerging from host egg upon spawning. E. Free-living stolon. F. Stolon fragmenting and dividing by fission. G. Completely developed and sexually reproductive *Polypodium.* Photograph labels correspond the labeled life cycle stage in illustrations. Illustration credit: Amanda Shaver.

Upon spawning of the host, *Polypodium hydriforme* everts, and emerges from the oocyte while reversing its germ-layer orientation, revealing ectodermally-derived tentacles along the length of the stolon (Fig. 1D). Once freely living outside the host's oocyte, the long stolon-like body of the *Polypodium hydriforme* begins to divide into many free-living fragments with six tentacles on each side (Raikova 2008), and eventually develops a mouth and begins to feed on small invertebrates (Fig. 1E-F). These twelve-tentacled individuals (Fig. 1F) divide asexually via longitudinal fission, beginning with the development of a new set of tentacles on the aboral side. Eventually, different individuals develop two distinct gonad types, although it appears that the 'female' gonads are not actually involved in reproduction (Raikova 2008). The free-living stage occurs entirely in fresh-water streams or lakes. Adult *Polypodium hydriforme* infects juvenile female fish and the parasites lies dormant within the host's ovaries in the binucleate cell-within-a-cell stage for several years until the host reaches sexual maturity (Raikova 2008).

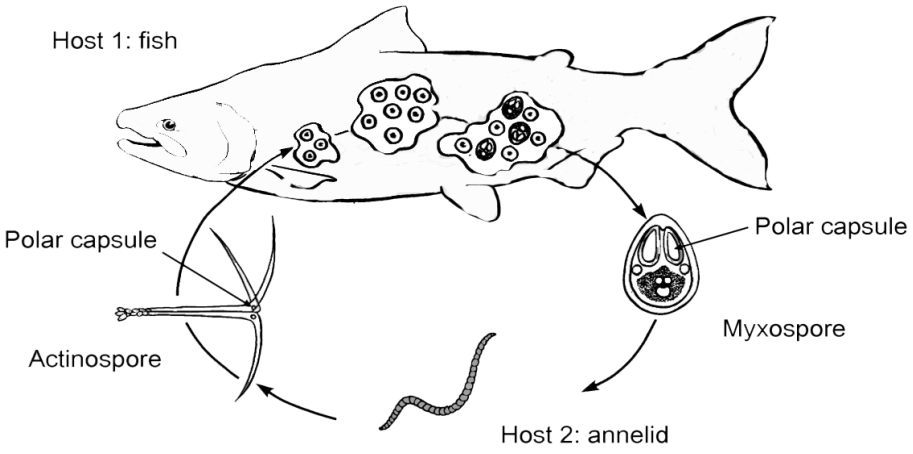
Myxozoan classification, diversity and life cycles. Myxozoa is a parasitic phylum that has been purported to have cnidarian origins. Within the phylum Myxozoa, two main lineages

are recognized, the class Myxosporea and the class Malacosporea (Canning et al. 2000). The malacosporeans contain only four known species, and includes the enigmatic, worm-like *Buddenbrockia plumatellae*, which has been the focus of a number of morphological and developmental studies (i.e. Gruhl and Okamura 2012; Canning et al. 2008). Malacosporeans are known to parasitize freshwater bryozoans for at least part of their life cycle, but no complete life cycle has ever been reported (Tops et al. 2005). However, a recent study has shown the successful experimental transfer of a malacosporean to carp and minnows, indicating that a fish may have a role as a host within this life cycle (Grabner and El-Matbouli 2010).

The myxosporeans are far more diverse, containing over 2,000 species (Lom and Dykova 2006). Members of this clade were found to be the causative agents of a number of diseases affecting economically important salmonid fishes, including whirling disease and proliferative kidney disease (Kent et al. 2001). Molecular phylogenetic studies within Myxozoa have confirmed the monophyly of the myxosporeans and have shown that freshwater and marine species largely fall into separate clades (Kent et al. 2001, Bartošová et al. 2009).

Figure 2: The life cycle and development of the Myxozoa, based on the life cycle of *Myxobolus cerebralis*.

Names of life cycle stages indicated here. Actinospore stage is about 350µm long and the myxospore stage is only about 10 µm long. For most myxozoans the second host is unknown.



The overall life cycle of *Myxobolus cerebralis* is typical of Myxosporea. *M. cerebralis* has two hosts within its life cycle, a polychaete worm and a salmonid fish. Individuals of the triactinomyxon spore stage (actinospore stage; Fig. 2) encounter their host through through dermal contact or consumption of the polychaete worm by the potential host fish (Markiw and Wolf 1983). After initial attachment to the fish, *M. cerebralis* burrows into the host tissue, and undergoes many rounds of replication within host muscular and nervous system tissues (El-Matbouli et al. 1995). This infection is what causes whirling disease in its host (Markiw and Wolf, 1983). Ultimately, individuals are released from the host as myxospores (Fig. 2), which contain only six cells (El-Matbouli et al. 1995), and go on to infect *T. tubifex.* Within the host worm, the *M. cerebralis* propogates and matures into the actinospore stage (Fig. 2), which are released into the water.

Phylogenetic placement of *Polypodium* and Myxozoa The phylogenetic placement of *Polypodium* and Myxozoa has been controversial, especially for the Myxozoa. From their first descriptions in the 1880s until relatively recently, species of myxozoans have been placed as protists, largely due to their reduced body plans and microscopic spores, which range in size from 10-350 micrometers, depending on life cycle stage (Markiw 1992). Unlike Myxozoa, the placement of *Polypodium* as a cnidarian has long been proposed based on morphology. *Polypodium* was reported to be a member of the derived hydrozoans, the Narcomedusae (Berrill, 1950), or amongst the scyphozoans (Lipin, 1925).

With the advent of molecular phylogenetics, it was confirmed that myxozoans were not protists but belonged to Metazoa (Smothers 1994). Using 18S rDNA and morphological data, Siddall et al. (1995) placed the myxozoans within Cnidaria and sister to *Polypodium* and found that Myxozoa and *Polypodium* were sister taxa. This overall relationship was supported by Evans

and colleagues, along with the suggestion that this might be an artifact due to long-branch attraction (LBA) (Evans et al. 2008). Jimenez-Guri et al. (2008), using a phylogenomic approach with a much larger sampling of genes, also found that the myxozoan *Buddenbrockia* fell within Cnidaria. In contrast to these findings, several studies placed Myxozoa as a group of early-diverging bilaterians (Schlegel et al. 1996; Anderson et al. 1998). Evans et al. (2010) found that placement of Myxozoa was highly dependent on character and taxon selection.

Most recently, Nesnidal et al. (2013) conducted a phylogenomic study using two myxozoan taxa; data from the partial genome of *Myxobolus cerebralis* and from the previously published data on *Buddenbrockia*. The authors concluded that this new, larger dataset unequivocally places Myxozoa among the cnidarians and supports the Evans et al. (2010) and Jimenez-Guri et al. (2008) conclusions that myxozoans are sister to the cnidarian clade Medusozoa. However, taxon sampling is limited, particularly among the myxozoa, and *P. hydriforme* is notably absent, making it difficult to draw any definitive conclusions about the relationship of these groups to each other and to the rest of cnidarians.

Nematocyst Structure and Function. In addition to the phylogenetic evidence discussed above, the assertion that *P. hydriforme* and Myxozoa belong within Cnidaria has been supported by the presence of nematocyst-like structures. Nematocysts are intra-cellular organelles unique to cnidarians, used for predation and defense. They are complex structures that are the product of a large post-Golgi vacuole, and consist of a capsule containing a long tubule that is tightly coiled until discharge. This tubule generally bears a number of spines (Reft and Daly 2012). The nematocysts are generally mechanosensitive, and evert their tubule at discharge, often to deliver venom to its target. There are a large variety of structural types among the nematocysts throughout cnidaria (David et al. 2008). These nematocysts types are most complex among the

Medusozoa. Medusozoans also possess some of the largest complements of different types of nematocysts (David et al. 2008).

The presence of structures similar to nematocysts have long been noted in both *P. hydriforme* and Myxozoa. Ibragimov and Raikova (2004) detected and described the presence of at least two distinct nematocyst types in *P. hydriforme*. Some features of the *P. hydriforme* nematocysts are distinct, however, from the nematocysts of any other cnidarian, most notably the possession double-helical symmetry as well as two helical rows of tiny spines in the discharged state and two helical folds in the undischarged state (Ibragimov and Raikova 2004). The functions of nematocysts in *P. hydriforme* include prey-capture during the free-living stages, tentacle adhesion during locomotion (Lipin 1911), and attachment to host fishes (Ibragimov 2004).

Myxozoans possess structures called polar capsules, which bear resemblance to cnidarian nematocysts and are used for attachment to and penetration of the host organism. As reviewed in Cannon and Wagner (2003), many researchers have noted the similarities between polar capsules and cnidarian nematocysts. Polar capsules are simpler in arrangement than most cnidarian nematocysts. Myxozoan polar capsules possess a thick capsular wall, an eversible polar filament, and a "cap" or "stopper" (Lom and Dykova 1992). A study by Kallert et al. (2005) demonstrated that myxozoan polar capsules may need a combination of both mechanical and chemical stimuli in order to trigger discharge. Reft and Daly (2012) noted that the apical structure was largely identical to the operculum which is restricted to Medusozoa, giving further support to the close relationship between Medusozoa and Myxozoa (Reft and Daly 2012).

Nematocysts are comprised of a number of proteins specific to this organelle and to Cnidaria, including the minicollagen protein family. Minicollagens are very short collagen

molecules that make up a large part of the structure of the nematocyst wall and tubule (Kurz et al, 1991). As shown in the model system *Hydra*, different types of minicollagens may be localized to specific areas of the nematocyst (Adamczyk et al. 2008). The hydrozoans, which have the greatest number of nematocyst types, have also been shown to have the greatest number of minicollagen types, and it has been suggested that the expansion of this protein family has contributed to the complexity and diversity of the nematocysts throughout Cnidaria (David et al. 2008).

Although the molecular components of cnidarian nematocysts are well characterized, little is known about the molecular basis for the nematocysts of *P. hydriforme* and the polar capsules of Myxozoa. Holland et al. (2011) reported a single minicollagen gene from an expressed sequence library from the myxozoan *Tetracapsuloides bryosalmonae*. Discovery and characterization of nematocyst specific genes in *P. hydriforme* and Myxozoa could provide important insight for the homology of *P.hydriforme* nematocysts and myxozoan polar capsules to cnidarian nematocysts, by assessing whether they are composed of the same types of nematocyst-specific proteins. In addition, the DNA sequences of these genes may prove informative for better informing us about the phylogenetic placement of these parasites within Cnidaria.

To further investigate questions regarding the origin and evolution of *P. hydriforme* and Myxozoa, we generated and characterized the transcriptomes of the myxosporean *Myxobolus cerebralis* and for *Polypodium hydriforme*. Transcriptomic data allows for the discovery of expressed genes, which could be important for understanding the evolution of developmental, functional and structural changes that occurred in the transition to a parasitic life cycle in these taxa. In addition, transcriptome sequences can be used in a phylogenetic context to allow for the

characterization of aspects of the genome that may better inform the phylogenetic placement of these groups. We were particularly interested in characterizing nematocyst-specific genes, as these are unique to Cnidaria and appear to be undergoing ongoing evolution within the phylum (Steele et al. 2011).

In addition, we sought to characterize the Wnt and Hox/ParaHox-like developmental regulatory gene families because of their conserved nature and key roles in development. These developmental gene families were found to play major roles in the anterior-posterior body patterning of bilaterians (Gehrig and Hiromi 1986), and have been found to play a number of roles in directing development in cnidarians (Cartwright et al. 2006; Chiori et al. 2005; Duffy et al. 2010; Guder et al. 2006). Previous studies of cnidarian genomes have found a large complement of genes in each gene family in several model cnidarians such as *Hydra* and *Nematostella* (Kusserow et al. 2005; Chiori et al. 2005). Since diversification, loss and changes in expression of these genes has been shown to be involved in creating the diversity of body plans throughout metazoa in general (Schubert and Holland 2006; Finnerty and Martindale 2001). Uncovering the number and types of Wnt and Hox/Parahox genes possessed by *P. hydriforme* and the myxozoans smay provide insight into the evolution of their unique body plans and life cycles.

This research represents a significant step towards a better understanding of the origin and evolution of these enigmatic parasitic organisms. In addition, these transcriptomic resources should contribute significantly to the comparative framework needed for future studies on cnidarian evolution that includes myxozoans and *Polypodium.*

Methods

RNA Extraction and Sequencing. Oocytes infected with *Polypodium hydriforme* were collected from mature, female paddlefish (*Polyodon spathula)* at the Paddlefish Research and Processing center in Twin Bridges State Park, Miami, OK, in March and April of 2011. In order to minimize possible contamination from the host, individual stolons of *P. hydriforme* were allowed to emerge from the host's ooctye and raised in spring water for 24-48 hours to allow for paddlefish yolk to be digested. Specimens were frozen at the elongate stolon stage (Fig. 1E) in liquid nitrogen. Uninfected paddlefish oocytes were also taken immediately from the female paddlefish and frozen in liquid nitrogen. *Myxobolus cerebralis* samples were flash frozen at the actinospore (Fig. 2) life cycle stage by Ron Hedrick at the University of California, Davis School of Veterinary Medicine, as they emerged from their invertebrate host.

In each case, RNA was extracted from multiple individuals of the same developmental stage (actinospore for *M. cerebralis* and elongate stolon stage for *P. hydriforme*) using TriReagent (Life Technologies, Grand Island, NY) following standard protocols, followed by DNAse treatment using the TURBO DNase kit (Ambion). *Myxobolus cerebralis* and *P. hydriforme* samples were prepared for sequencing by Mariya Shcheglovitova (M.S.) using the TruSeq RNA Sample Preparation Kit v2 (Illumina Inc., San Diego, CA). RNA was chemically fragmented, using reagents supplied in the TruSeq kit, resulting in libraries with a median insert size of 155 bp. Both samples were given a single lane and sequenced at the University of Massachusetts Medical School Molecular Biology Core Facility on an Illumina HiSeq 2000

machine. The uninfected paddlefish eggs were prepared for sequencing by Clark Bloomer at the University of Kansas Medical School Sequencing Center, and sequenced at this center.

Illumina sequence assembly. The number of 100-bp reads generated for each of the three libraries is shown in Table 1. Datasets were examined in FastQC v.0.8.0 by Babraham Bioinformatics (2010) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) for overall quality score distributions and to assess quality cutoffs. Reads were quality trimmed with a script created by M.S. (https://github.com/bastodian/shed/blob/master/Python/q-trim.py) using default values of Phred score cutoff of 21, 5bp of contiguous low quality sequence and at least 30bp left of a sequence in order for the read to be retained. Trimmed reads were re-paired and separated into files for forwards and backwards reads using both.py (https://github.com/bastodian/shed/blob/master/Python/both.py), also written by M.S. The trimmed reads were assembled using Trinity (Grabherr et al., 2011) using the 2012-10-05 release for the *M. cerebralis* and *P. hydriforme* assemblies and the paddlefish egg assembly. Each run was implemented with the following settings: (--SS_lib_type RF --CPU 10 --bflyJavaVM64bit --bflyHeapSpace 20G --bflyMinHeapSpace 20G --bflyHeapNursery 20G --bflyJavaGCParallel --bflyJavaGCThreads 16 --repeat 5 --bflyJavaCmdLifespan_min 5 --bflyJavaCmdLifespan_max 1800 --bfly_opts "-V 10 --stderr") on the Bioinformatics Cluster at the Information and Telecommunications Cluster at the University of Kansas.

Contamination Filtering. Since *P. hydriforme* is an obligate endoparasites, we were concerned about contamination by host sequences. The assembled *P. hydriforme* sequences were initially used as a BLAST (Altschul et al. 1997) query against databases of potential contaminant sequences. The first database was created from every expressed sequence tag (EST) and genomic sequence from acipenseriform fishes found in the database of the National Center for

Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov). Given that there are only 7,669 acipenseriform sequences in NCBI databases, an additional database that included ESTs from the model fish *Danio rerio* was generated from these databases. A third BLAST database was created out of the newly generated Illumina sequences from assembled paddlefish oocyte RNA. The acipenseriform, acipenseriform plus *D. rerio* sequences, and our newly generated paddlefish oocyte transcriptome were made into nucleotide BLAST databases using the makeblastdb function. The *P. hydriforme* transcriptome sequences were queried against the databases from NCBI and a database created from our newly-assembled paddlefish oocyte sequences in order to identify *P. hydriforme* sequences that have significant BLAST hits to fish sequences at a variety of e-value levels and using both tblastx and tblastn algorithms.

Transcriptome Annotation. All assembled sequences for both *P. hydriforme* and *M. cerebralis* were initially annotated using the program BLAST2GO (Conesa et al. 2005). A tblastx search against the non-redundant protein database (NR) at an e-value of 1e-3, mapping, and Gene Ontology (GO annotation) steps were implemented. Graphs of GO annotation were also created with BLAST2GO.

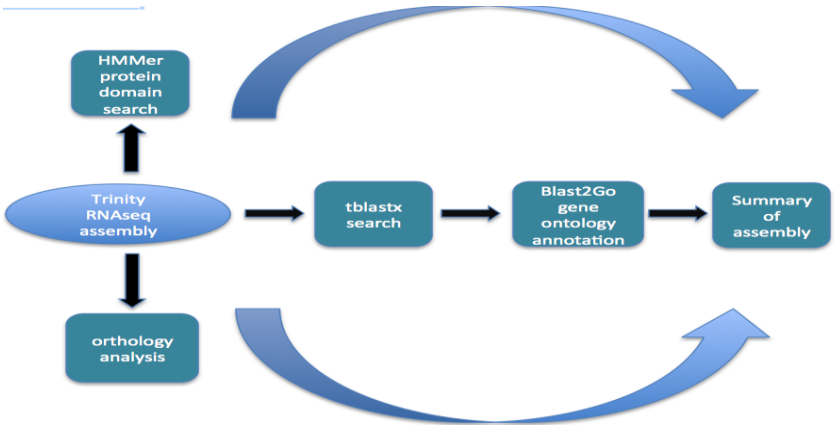


Figure 3. Flowchart of assembly and annotation methods

To assign protein domains to the greatest possible number of sequences, the *M. cerebralis* and *P. hydriforme* assembled transcriptome sequences were run through HMMer (Finn et al. 2011) as implemented on the University of Indiana Integrated Server for Genomic Analysis (ISGA). HMMer uses profile Hidden Markov Models to query the assembled sequences against the PFam (Sonnhammer et al. 1997) and TIGRFAMs (Haft et al. 2001) protein domain databases.

Both data sets of assembled sequences were also run through the gene orthology module on ISGA in order to perform both orthology and paralogy analysis. Through this pipeline, the sequences are first BLAST searched against the OrthoMCL database (Chen et al. 2006), as well as queried against themselves, and the OrthoMCL algorithm is used to form orthology and paralogy groups (Li et al. 2003). The data sets are also BLAST queried against the OrthoDB dataset to assign OrthoDB Group ID. Candidate Hox-like and Wnt pathway developmental regulatory genes and nematocyst-specific genes were identified from the transcriptomes using the above orthology queries, as well as targeted BLAST searches using cnidarian sequences from these gene families taken from NCBI.

Phylogenetic Methods. Sequences from the transcriptomes were identified as belonging to the Hox/Parahox, Wnt, and minicollagen gene families by BLAST searching the entire transcriptomes against known cnidarian sequences for these gene families and against the entire NR database. Sequences identified through these searches were initially imported into Geneious Basic 5.6.3 (Drummond et al. 2012) to check for duplicates and translate into all six reading frames. Correct reading frames were chosen for further analysis by seeing translating sequences in all six reading frame and seeing which of the translations returned the appropriate conserved domain in a BLAST search against NR. For Hox/ParaHox genes, translated sequences were then

aligned using the MAFFT E-ins-I algorithm along with many sequences from the data set of Chiori et al. (2009), which contains Hox/ParaHox-extended genes from a variety of cnidarians and bilaterians, and trimmed down to the 60aa homeodomain region for further analysis using BioEdit v.7.1.10 (Hall 1999). The Wnt and minicollagen genes were aligned with cnidarian sequences from NCBI from their respective gene families using the E-ins-I algorithm and trimmed in BioEdit. A single minicollagen sequence from another myxozoan was included in our analysis. In a 2010 study, Holland et al. isolated a minicollagen sequence from the myxozoan *Tetracapsuloides bryosalmonae,* which found on GenBank and added to our database of cnidarian minicollagen sequences. Trimmed alignments were run through the default settings on the ProtTest 2.4 (Abascal et al. 2005) server to get an initial suggestion of model choice for the phylogenetic analysis of these sequences.

Maximum likelihood inference was implemented in RaxML 2.7.2 (Statamakis 2006), starting with 30 maximum likelihood trees and 200 bootstrap replicates for each analysis. The PROTGAMMA+LG model was used the analysis of the Wnt-like sequences, the PROTGAMMA+BLOSUM62 model was used for the minicollagen sequences and PROTCAT+I+LG was used for final analysis of the Hox/ParaHox sequences. Bayesian analysis for the Wnt-like and minicollagen data sets was implemented in MrBayes3.2.1 (Ronquist and Huelsenbeck 2003) on Cyberinfrastructure for Phylogenetics Research Portal (CIPRES) (Miller et. al, 2010) with a GAMMA model with a mixed substitution rates model (the aamodelpr=mixed setting). Two runs of 3 hot chains and 1 cold chain were run for 40 million generations with a 10 percent burn-in for each of these analyses, and trees were summarized using sumt in MrBayes3.2.1.

Results

Sequencing and Assembly Results

Number of Illumina generated sequences and overall assembly statistics are shown in Table 1. Overall number of raw reads and number of assembled sequences are similar for both *P. hydriforme* and *M. cerebralis*, with *P. hydriforme* having N50 and N75 values nearly twice as long as *M. cerebralis*. By contrast, the *P. spathula* oocyte assembly produced over three times as many sequences as the others. Looking at the sequence-size distribution of this assembly, over 70,000 of these sequences are between relatively short (200-299bp, not shown). The large number of short sequences may be due to the failure to assemble full transcripts due to degradation.

Table 1. Statistics from the Trinity RNA-seq assemblies for *P. hydriforme, M. cerebralis* and *P. spathula*

	P. hydriforme	***M. cerebralis***	***P. spathula* oocyte**
# of Raw Reads	467 million	312 million	128 million
# of Assembled Sequences	76,434	83,398	235,055
N25	2864	1660	2108
N50	1660	822	1081
N75	744	412	493
Longest Contig	22,163	9634	16,480

Contamination filtering results. Using nucleotide BLAST (blastn) against a database of only acipenseriform sequences identified a very low number of contaminant sequences. The translated query-translated query algorithm (tblastx) identified a larger number of potential contaminant sequences, although proved particularly poor at specifically identifying contaminant

sequences to the exclusion of potentially cnidarian sequences. Surprisingly, BLAST-searching the transcriptome of *P. hydriforme* against a database of sequences from the actual host tissue itself (the paddlefish egg transcriptome) did not appear to improve filtering specificity, with regards to whether potential cnidarian sequences were also identified as being contaminant.

Table 2. Results of BLAST searches of the *P. hydriforme* transcriptome against databases of fish sequences

Database	Algorithm	E-value Cutoff	# of significant hits in *P. hydriforme* transcriptome	Overlap with potential cnidarian sequences
Acipenseriformes	Blastn	1e-3	56	0
Acipenseriformes	Tblastx	1e-30	394	280
AP+ Danio rerio	Blastn	1e-3	323	210
AP + Danio rerio	Tblastx	1e-30	1023	1022
Paddlefish egg transcriptome	Tblastn	1e-3	295	218

Annotation Results. Several different approaches were utilized to annotate and functionally categorize the assemble contigs from the transcriptomes. About 37% of assembled sequences from each of the *M. cerebralis* and *P. hydriforme* transcriptomes were given blastx hits at an e-value cutoff of 1e-03 (Table 3). Notably, in each of these cases, the top two BLAST top-hit species were cnidarians, the model organisms *Hydra vulgaris* and *Nematostella vectensis.* Of the sequences with significant BLAST hits, 12,167 and 10, 418, respectively for the *P. hydriforme* and *M. cerebralis* transcriptomes were identified as 'predicted proteins' and could not be annotated further using the gene ontology (GO) pipeline. Sequences identified as cnidarian by their tblastx top-hit make up about 25% (of the 37% that had significant blast hits) of each of these transcriptomes (Fig. 4). GO annotation assigned categories to 18% and 19% of the *P. hydriforme* and *M. cerebralis* transcriptomes (Table 2) respectively. Of those GO annotations, the overall distribution of top molecular function, biological process and cellular component

categories were almost identical between the two transcriptomes (see figures 5, 6 and 7).

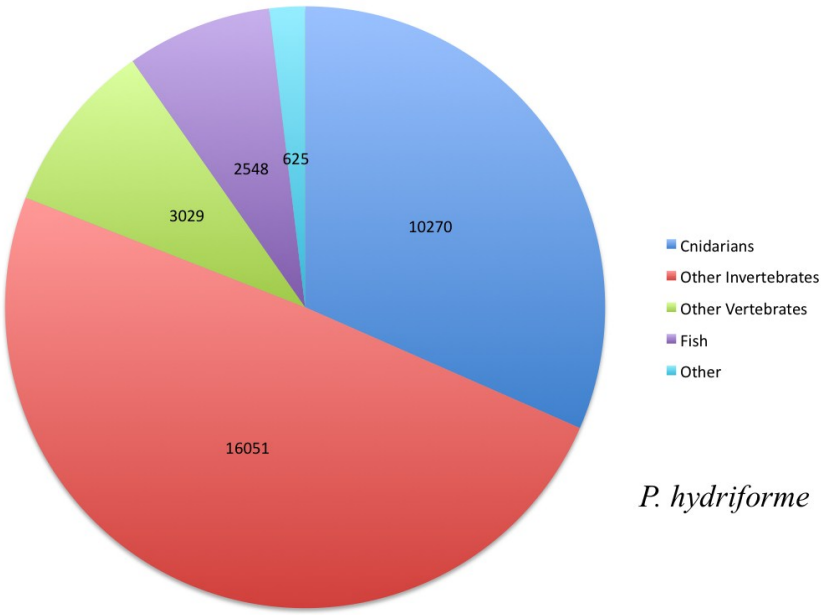
HMMer searching uses Hidden Markov models to identify short stretches of similarity from sequences to protein domains from the PFAM and TIGRFAM databases. This search was able to assign protein family identifications, or both *M. cerebralis* and *P. hydriforme*, to a number of sequences that had no significant BLAST hits or ones that GO annotation assigned the identification 'Predicted Protein' or 'Hypothetical Protein' or other non-specific titles (Table 2).This allowed for the identification of several more candidate genes (as noted in tables 3, 4 and 5).

Table 2. Annotation for *P. hydriforme* and *M. cerebralis* transcriptomes.

Sequences in each category	*Polypodium hydriforme*	*Myxobolus cerebralis*
Total Assembled Sequences	76,434	83,398
With BLAST nr hits	28,343	30,846
With GO category	14,006	16,020
Total 'predicted proteins'	12,167	10,418
Assigned HMM families	22,153	22,918
Assigned OrthoMCL ortholog groups	6503	7319
# of paralog within OrthoMCL ortholog groups	22,869	27,250
Total # (or percent) of assembled sequences with a search term	28,930 (37.8%)	33,981 (40.7%)

The OrthoMCL pipeline, which identifies sequences to paralog and ortholog groups, was informative for grouping and categorizing unidentified sequences. From the *P. hydriforme* and *M. cerebralis* assemblies, respectively, 6503 and 7319 sequences were assigned ortholog groups, and a very large number from each were assigned as paralog groups within the ortholog group

(Table 2). Many of these “paralog” groups are likely incompletely assembled contigs and isotigs.



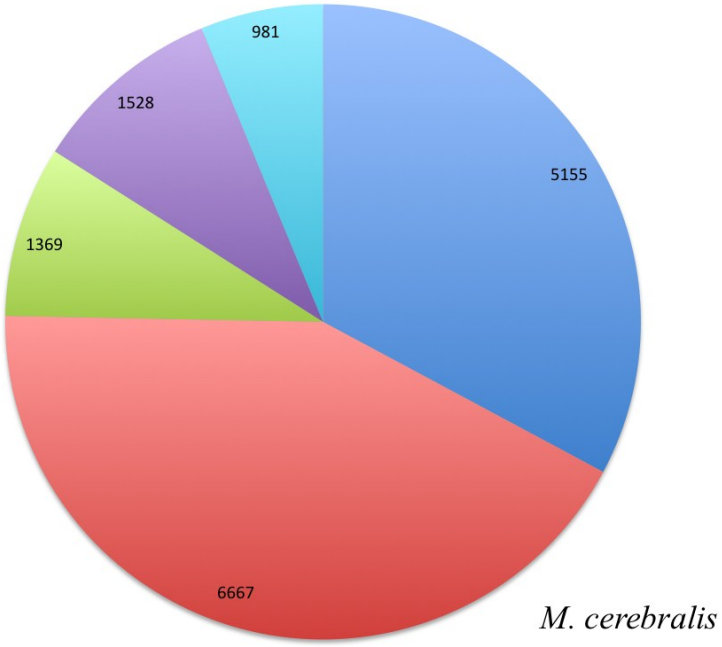


Figure 4. Distribution of taxonomic categories of BLAST top hits for the top 20 species for each transcriptome.

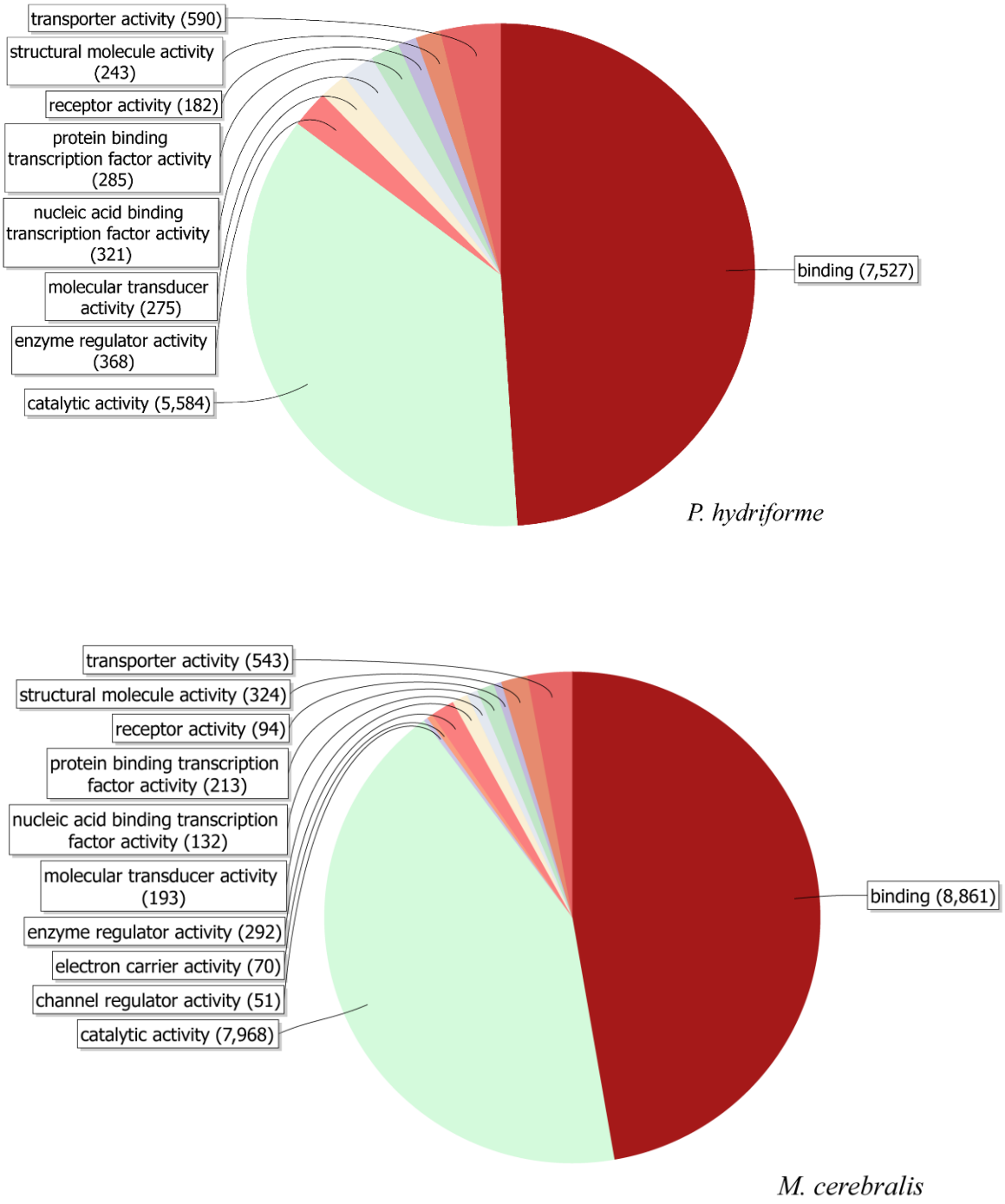


Figure 5. Distribution of Level 2 molecular function categories produced by GO annotation of the *P. hydriforme* and *M. cerebralis* transcriptomes

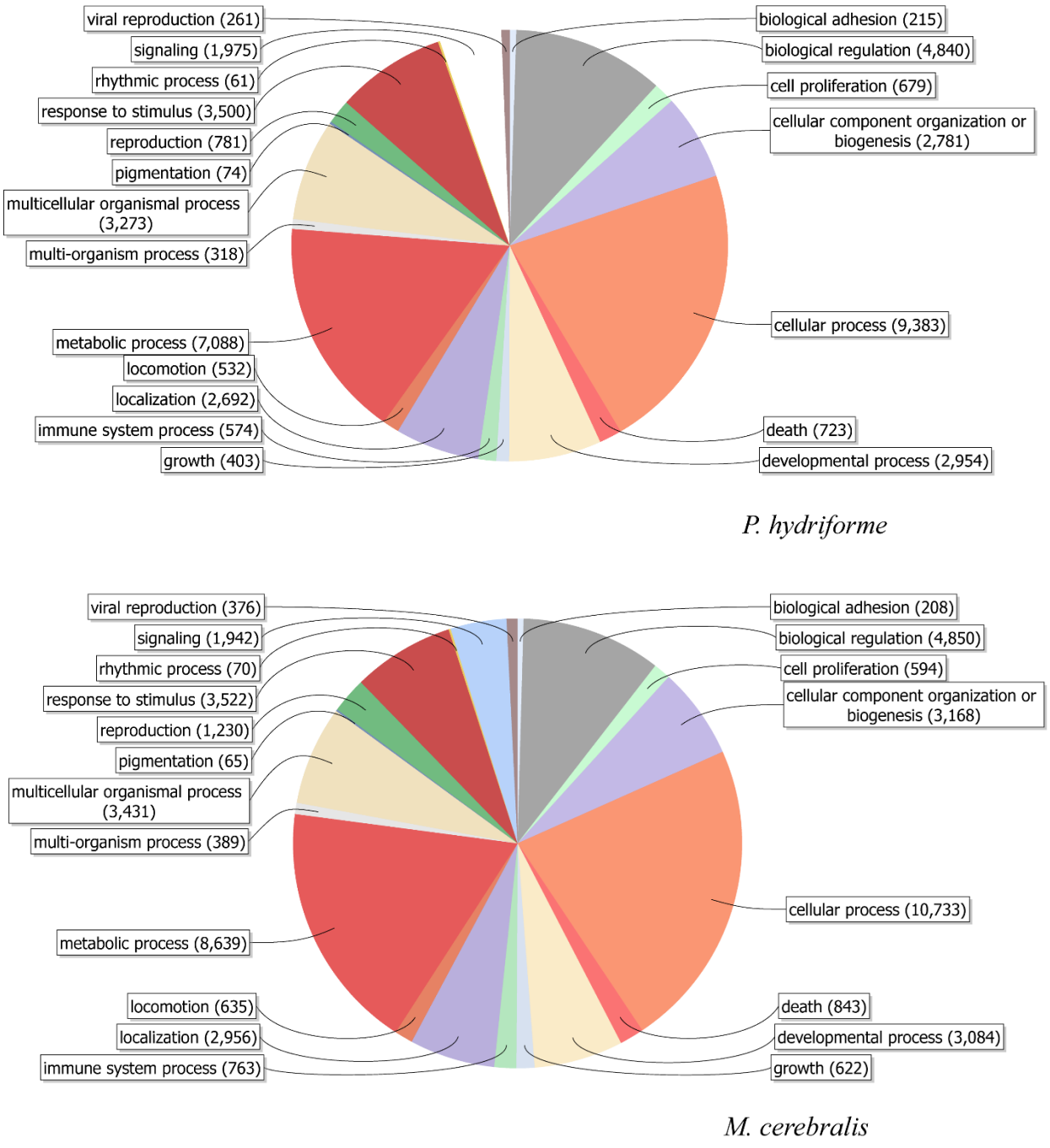
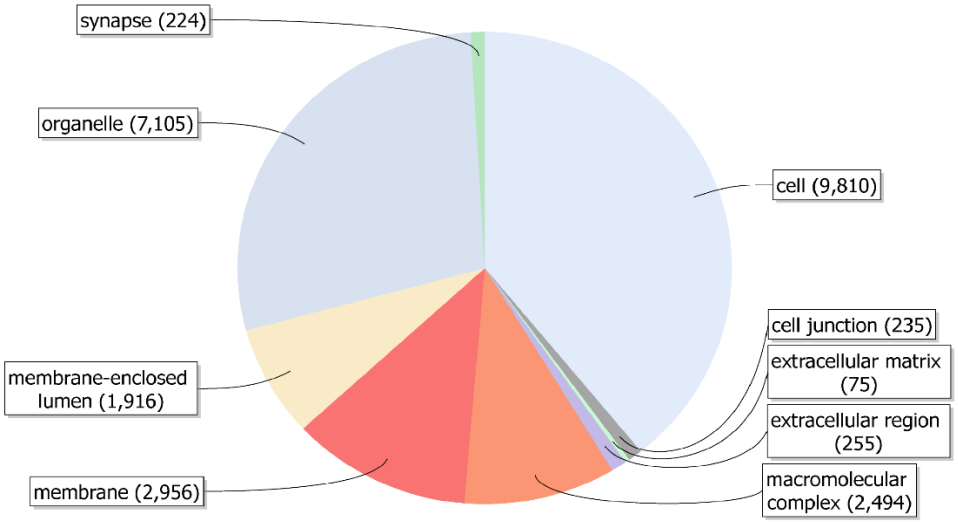
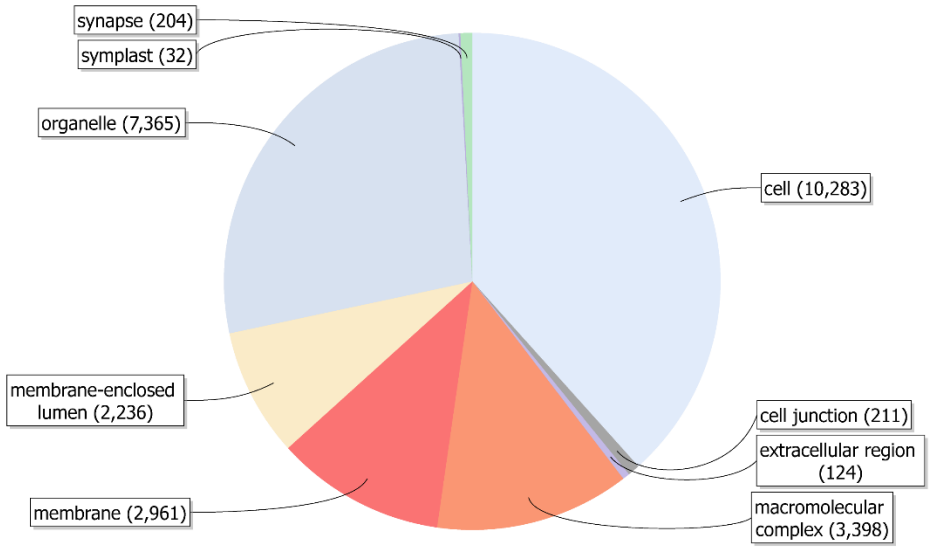


Figure 6. Distribution of Level 2 biological process categories produced by GO annotation of the *P. hydriforme* and *M. cerebralis* assemblies**.**



P. hydriforme



M. cerebralis

Figure 7. Distribution of Level 2 cellular component categories produced by GO annotation of the *M. cerebralis* and *P. hydriforme* assemblies.

Hox and ParaHox genes. Using BLAST searching, the *P. hydriforme* transcriptome yielded five unique homeodomain-containing sequences from the Hox/Parahox-extended gene family (Table 3). An additional sequence was found through an HMMer search against the PFAM database. In the transcriptome of *Myxobolus cerebralis* only a single sequence from the Hox-extended family was retrieved through both tblastx against NR and through a protein-domain search. This sequence was identified through BLAST as a putative Cdx sequence.

Table 3. **Homeodomain-containing candidate genes extracted from assembled transcripts of *M. cerebralis* and *P. hydriforme*.** Sequences from *P. hydriforme* begin with Phy, and the single sequence beginning *Mce* is from *M. cerebralis*. Cnidarian species are in bold.

Sequence ID	Sequence Length	BLAST top-hit ID	BLAST top-hit species	Top-hit e-value	Top-hit %sequence identity
Phy9787	697	Cnox3	***H. viridissima***	1.00e-26	52.4
Phy16739	947	Cnox4	***E. dichotoma***	4.00e-18	64.7
Phy52035	323	Cnox2	***A. millepora***	8.00e-25	71.6
Phy19578	879	MoxD	***N. vectensis***	3.00e-29	59.4
Phy11495	1130	Msx	***N. vectensis***	8.00e-16	35.8
Phy16555	1570	Pdx/Xlox	***T. dohrnii***	1.00e-14	51.8
Mce16424	632	Cdx	*P. excavatus*	4.00e-6	55.3

Phylogenetic analyses revealed six *P. hydriforme* genes as orthologs of Hox/Parahox groups (Fig. 9). Placement of these sequences and overall topology of the tree were consistent among maximum likelihood (ML) analyses and was not dependent on whether or not a proportion of variant sites was included. In general, bootstrap support was low throughout the

tree (Fig. 9), particularly for the relationships deeper in the trees between the major Hox/Parahox groups. Two sequences, Phy7987 and Phy11495 were recovered as putative Msx genes (a homeobox-containing gene that is not classified as part of the Hox family). A single putative Mox gene, Phy19578, was also recovered. Following the naming scheme of Chiori et al. 2009, a single "anterior" Hox (HOX1/HOX2/GSX) gene sequence was identified, Phy52035, as well as a several "posterior" Hox sequence (Hox9-14/Cdx) sequence, Phy16739, Phy16555 and Mce16424 (Fig. 9).

Notably, the branch representing the *M. cerebralis* sequence Mce16424 is extremely long. The sequence from *M. cerebralis* possesses only 34% sequence identity to the gene it grouped most closely with in the phylogenetic analysis, which was the Hox9 gene of the lancelet, *Branchiostoma floridae* (Fig. 9). For comparison, *B. floridae* Hox9 and Hox10 genes have 79% shared sequence identity. Examining the alignment of the sequence from *M. cerebralis* with other "median" Hox genes confirms the divergent nature of the homeobox region of the sequence Mce16424 and at this point it should be considered a putative Hox gene (Fig. 8).

```
HvuCnox2     KSKRIRTAYTSIQLLELEKEFQNNRYLSRLRRIQIAAILDLTEKQVKIWFQNRRVKWKKD
Phy52035     ******-----R---------VS-------------TF-N-S-----------------
Phy7987      -CRKP--VFSDM--MV--R--LEH----TPQ-SKL-ER-R-DPI---T------M-L--E
Phy16739     PTM-T-PCF-AH-TN---R--LKSQ-ITKA--MELSSL-V-S-Q-I----------ER-T
Phy16555     SR--C--T-SRL-Q----R-YRLS--I--AK-VEL-GS-L----HI-V------M-D-RV
Phy19578     --RKE---FSKL-IR---M--VR-N--T----YE--VA-G---R---V------M---RI
Phy11495     SQFQP-RPFSRK--DV--AF-KRRQ--DLVD--YL-DK-NM--R----------A-S-RQ
Myxo16424    EE-TRGNRHDKRAIRL-----LKSPFI-AA-KRSLGGLIG-SDE--SV-YK-K-******
```

Figure 8. Alignment diagram comparing Hox/ParaHox-like sequences to the *Hydra vulgaris* Cnox2 gene. Missing amino acids are marked with an asterisk.

Wnt Sequences. A total of eight Wnt-like genes were isolated from the transcriptome of *P. hydriforme* via a tblastx search against NR and a targeted search of known cnidarian Wnt

sequences from GenBank against the transcriptome. These were included in gene tree analyses along with a comprehensive sampling of cnidarian sequences from GenBank. The eight Wnt-like sequences from *P. hydriforme* span some of the diversity of known Wnt gene subfamilies, as established in Kusserow et al. (2005). *P. hydriforme* sequences were placed, both in the ML and Bayesian analysis, within the Wnt7/8 (Phy18670, Phy10224, Phy7444, Phy 23685), Wnt5 (Phy 16339), Wnt3 (Phy4183 and Phy5694) and Wnt 6 (Phy4979) orthology groups, although with low nodal support. Phy10871 has variable placement, grouping with Wnt9/10 in the ML analysis (Fig.10) and with Wnt5 in the Bayesian analysis (Fig. 11). The presence of three unique Wnt7/8-like sequences for *P. hydriforme* appears to represent a lineage-specific duplication, as only up to two Wnt8 sequences have been found in any other cnidarian species (see *Hydra vulgaris* sequences in Fig. 10 and 11). In contrast to *P. hydriforme*, a search of the *M. cerebralis* transcriptome did not recover any Wnt-like sequences using either BLAST and HMMer searches.

Table 4. Wnt-like candidate genes extracted from assembled transcriptome of *P. hydriforme*. Sequences from *P. hydriforme* begin Phy, and the sequence beginning *Mce* is from *M. cerebralis*. Blast hits from cnidarians are in bold. Sequences found through HMMer searching are marked with an asterisk.

Sequence ID	Sequence Length	BLAST top-hit ID	BLAST top-hit species	Top-hit e-value	Top-hit percent sequence identity
Phy4979*	1611	predicted protein	***N. vectensis***	1E-54	33
Phy5183	1707	wingless-type member 3 precursor	*S. kowalevskii*	1E-79	39.1
Phy7444	1491	WNT8 protein variant 2	*D. pulex*	1E-79	39.1
Phy10224	1589	WNT16	*D. pulex*	2E-27	33.3
Phy10871	1486	Wnt5a	***H. vulgaris***	0.00002	51.5
Phy16339	1456	Wnt-5b-like	*N. vitripennis*	3E-33	29.1
Phy18982	906	WNT10	*P. excavatus*	3E-65	32.1
Phy23685	448	Wnt7-like protein	***N. vectensis***	4.00e-16	44.9

Minicollagen sequences. Minicollagen sequences were recovered from both *P. hydriforme* and *M. cerebralis.* Notably, nine distinct minicollagens were recovered from the *P. hydriforme* transcriptome while only two unique sequences were found from *M. cerebralis*. Phylogenetic analyses of the new sequences with the rest of the cnidarian sequences recovered different topologies using Bayesian and maximum likelihood inferences (Fig. 12 and 13). Notably, in the ML analysis, Phy1038, Phy272, Phy394 and Phy268 form a clade with Mce21132 and the *T. bryosalmonae* gene from Holland (2010), referred to as Tbr1 during the analysis (Fig. 12).

Table 5. **Minicollagen-like candidate genes extracted from assembled transcripts of *M. cerebralis* and *P. hydriforme.*** Sequences from *P. hydriforme* begin Phy, and the sequence beginning *Mce* is from *M. cerebralis*. Cnidarian sequences are in bold. Sequences found through HMMer searching are marked with an asterisk.

Sequence ID	Sequence Length	BLAST top-hit ID	BLAST top-hit species	Top-hit e-value	Top-hit percent sequence identity
Phy1083	838	Nematocyte-specific minicollagen	***Hydra magnipapillata***	1e-09	54.5
Phy1422	1261	Minicollagen 15	***Hydra magnipapillata***	5e-04	36.7
Phy13939	788	Minicollagen 1	***Clytia hemisphaerica***	1e-04	33.1
Phy1507	1116	Minicollagen 10	***Hydra magnipapillata***	0.001	61.4
Phy1565	702	Minicollagen	***Acropora donei***	8e-05	48.7
Phy1567*	933	Predicted protein	***Nematostella vectensis***	.12	60.0
Phy268	999	Nb001	***Hydra magnipapillata***	7e-06	58.1
Phy272	716	Minicollagen 5	***Nematostella vectensis***	2e-12	59.2
Phy3727	898	Minicollagen 5	***Nematostella vectensis***	8e-05	38.6
Phy394	838	Minicollagen 5	***Hydra magnipapillata***	8e-07	65.5
MCe17071	616	Minicollagen 15	***Hydra magnipapillata***	3e-08	23.9
MCe21132*	404	N/A			

The other genes from *P. hydriforme* are recovered in groups with cnidarian genes, with the exception of Phy13939, which falls outside of any clades of minicollagens. *M. cerebralis* sequence Mce17071 falls out in a clade of minicollagen-1 genes. In the Bayesian inference analysis, Mce21132 was not recovered as an ortholog of the sequence from the myxozoan *T. bryosalmonae* (TbrMC1), and Phy268 is recovered in the same group as this *M. cerebralis* gene (Fig. 13). In this tree, Phy272, Phy1038, Phy397, Phy 394 and Phy3727 form a clade, similar to the large grouping of *P. hydriforme* sequences in the maximum likelihood tree.

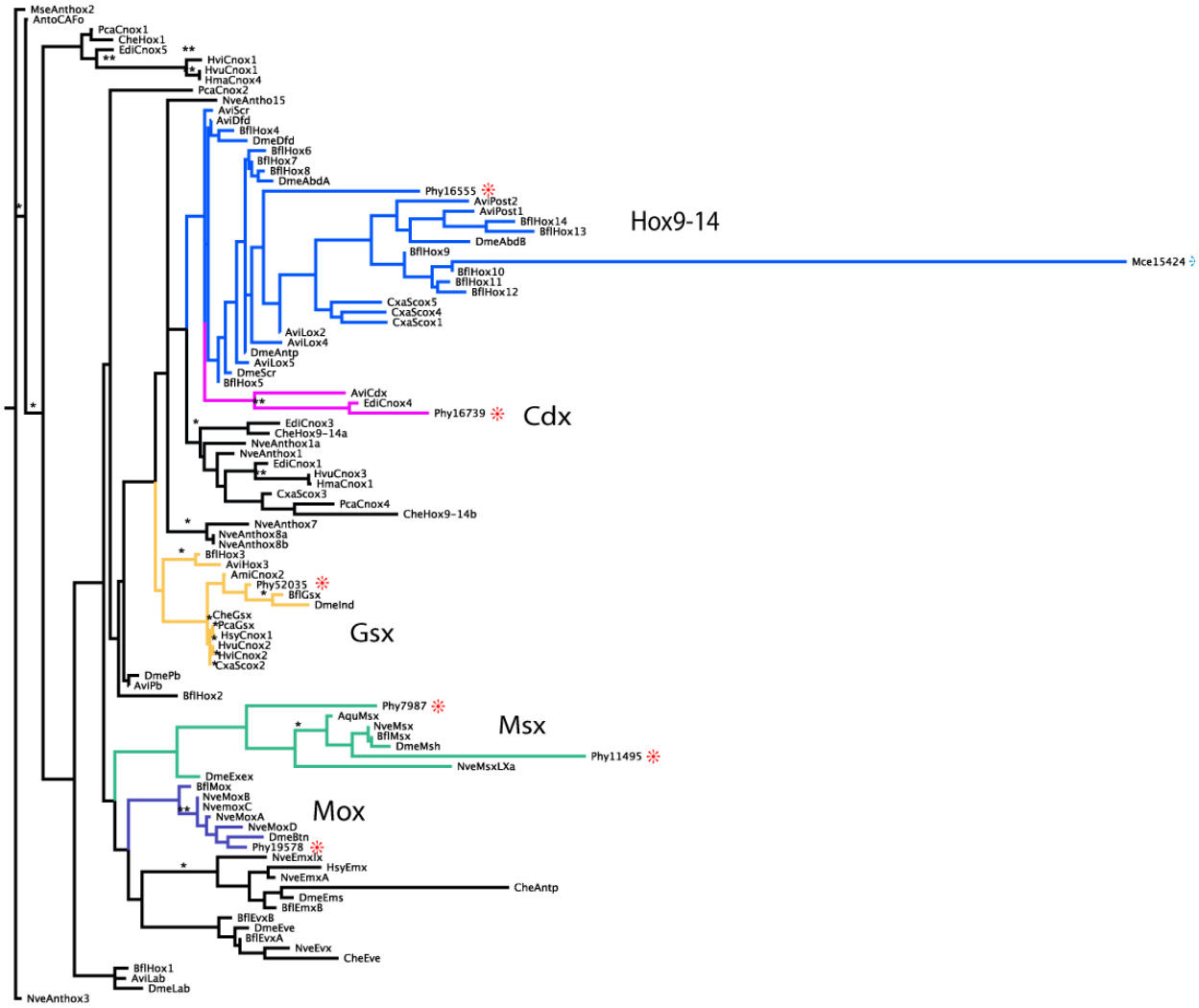


Figure 9. Phylogenetic relationships between cnidarian and bilaterian Hox/Parahox homeodomain sequences, including new sequences from *P. hydriforme* and *M. cerebralis* , inferred through ML analysis. Species names: Ami=*Acropora millepora*; Aqu=*Amphimedon queenslandica*; Avi =*Alitta virens*; Bfe=*Branchiostoma floridae*; Che=*Clytia hemisphaerica*; Cxa=*Cassiopea xamachana*; Dme=*Drosophila melanogaster*; Edi=*Eleutheria dichotoma*; Hma=*Hydra magnipapillata*; Hsy=*Hydractinia symbiolongicarpus*; Hvi=*Hydra viridissima*; Hvu=*Hydra vulgaris*; Mse=*Metridium senile*; Mce=*Myxobolus cerebralis*; Nve=*Nematostella vectensis*; Pca=*Podocoryne carnea*; Phy=*Polypodium hydriforme*. * = > 70bootstrap support, **=>90 bootstrap support. Sequences from *P. hydriforme* are marked with a red asterisk, and those from *M. cerebralis* are marked with a blue asterisk.

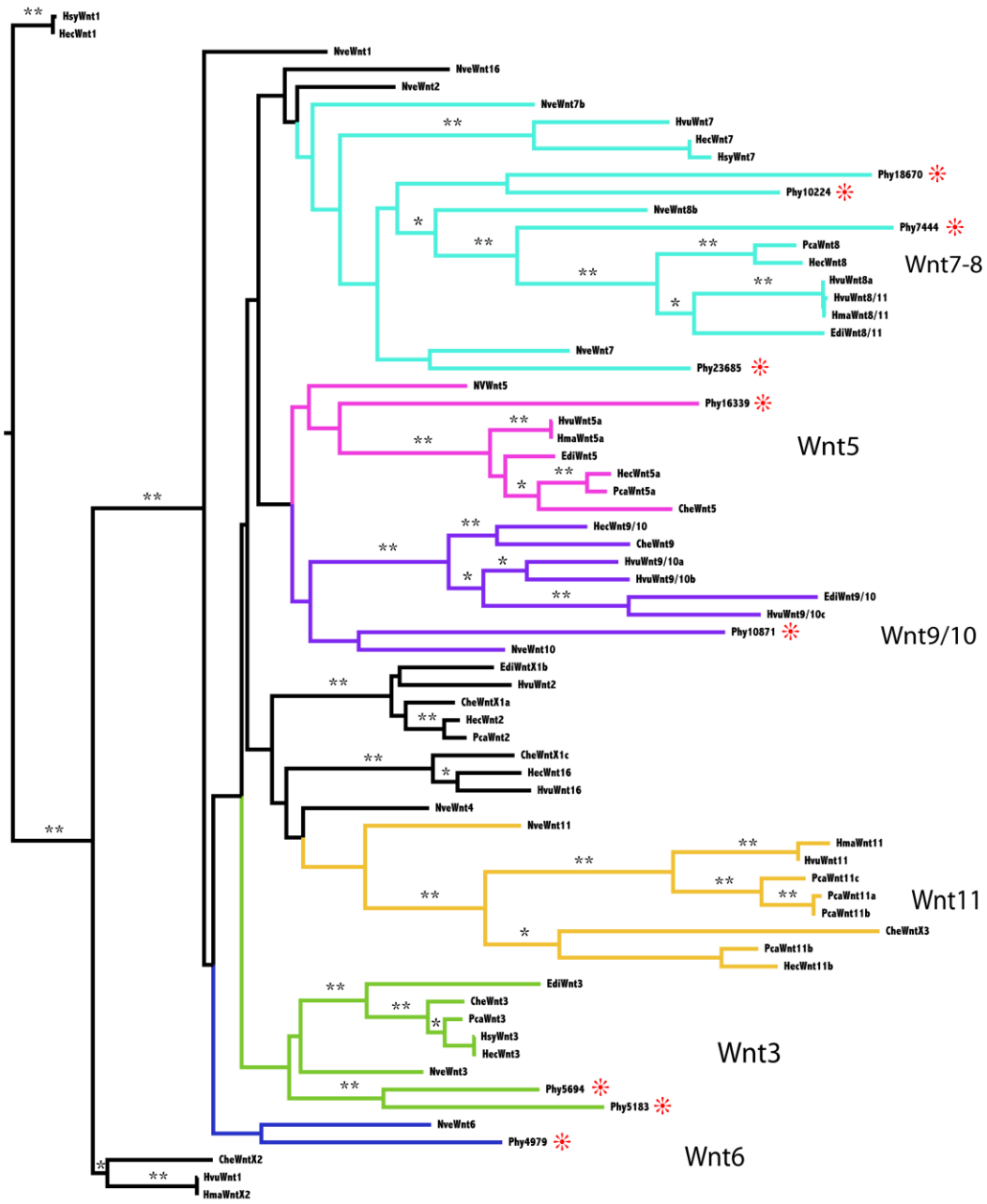


Figure 10. Phylogenetic relationships between cnidarian and bilaterian Wnt-like sequences, including new sequences from *P. hydriforme* and *M. cerebralis*, inferred by **ML analysis.** Taxonomic abbreviations: Che=*Clytia hemisphaerica*; Edi=*Eleutheria dichotoma*; Hec=*Hydractinia echinata*. Hma=*Hydra magnipapillata*; Hsy=*Hydractinia symbiolongicarpus*; Hvu=*Hydra vulgaris*; Mse,=*Metridium senile*; Mce=*Myxobolus cerebralis*; Nve=*Nematostella vectensis*; Pca=*Podocoryne carnea*; Phy=*Polypodium hydriforme.* * = > 70bootstrap support, **=>90 bootstrap support. Sequences from *P. hydriforme* are marked with a red asterisk.

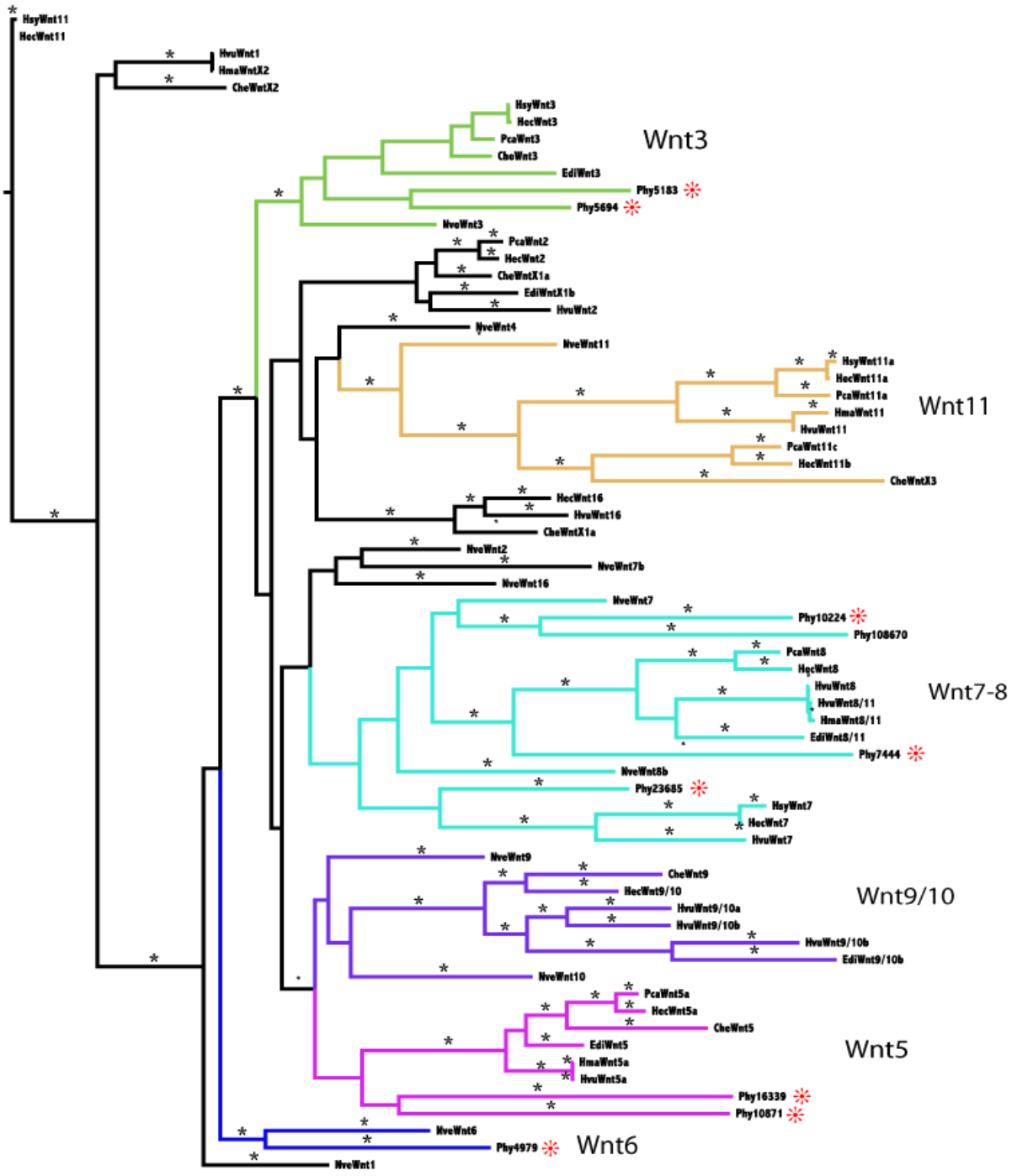


Figure 11. Phylogenetic relationships between cnidarian and bilaterian Wnt-like sequences, including new sequences from *P. hydriforme* and *M. cerebralis*, inferred through Bayesian analysis. Che=*Clytia hemisphaerica*; Edi=*Eleutheria dichotoma*; Hec=*Hydractinia echinata*. Hma=*Hydra magnipapillata*; Hsy=*Hydractinia symbiolongicarpus*; Hvu=*Hydra vulgaris*; Mse,=*Metridium senile*; Mce=*Myxobolus cerebralis*; Nve=*Nematostella vectensis*; Pca=*Podocoryne carnea*; Phy=*Polypodium hydriforme.* * = > 95% posterior probability. Sequences from *P. hydriforme* are marked with a red asterisk.

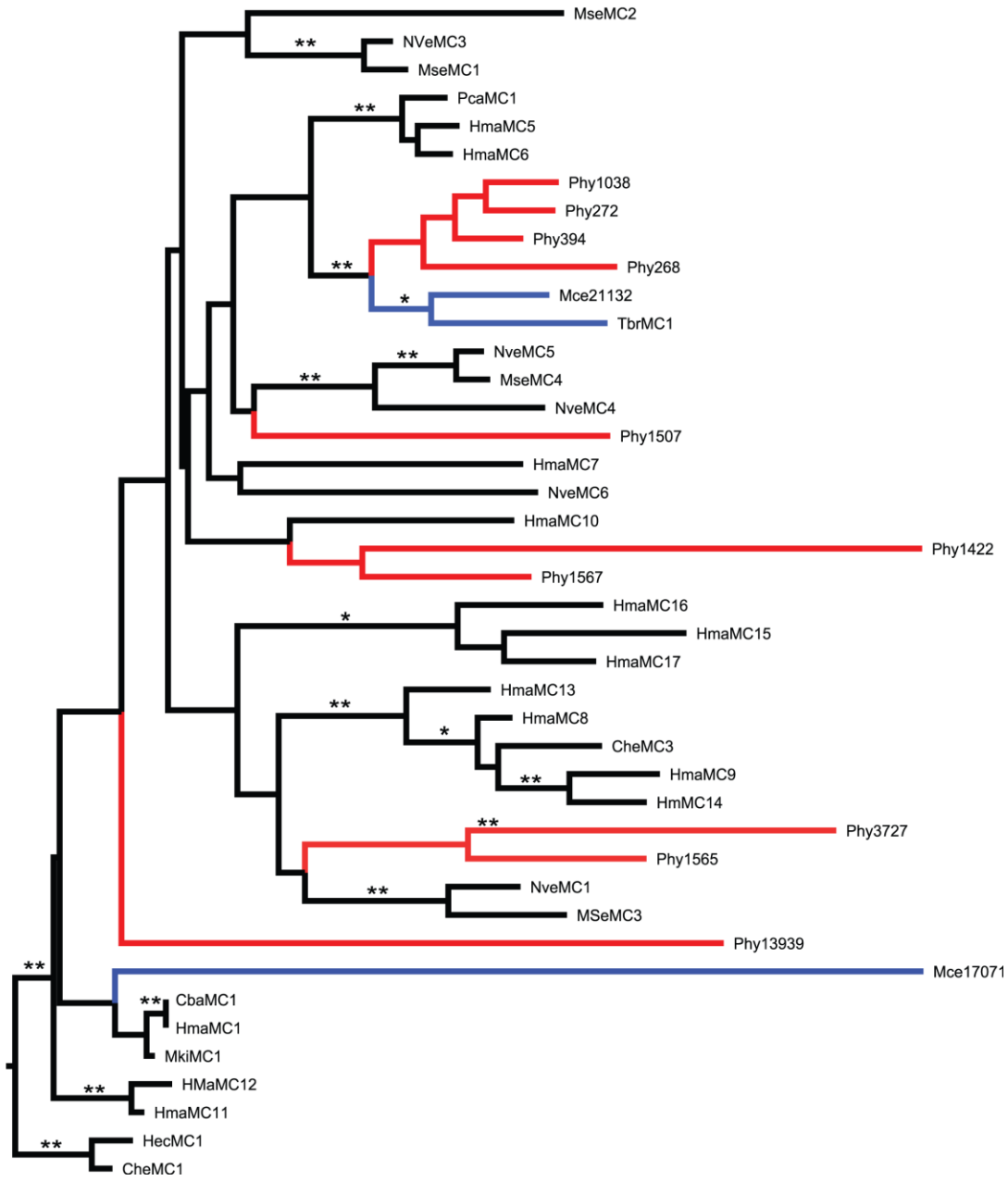


Figure 12. Phylogenetic relationships of minicollagen sequences including new sequences from *P. hydriforme* and *M. cerebralis* using maximum likelihood. Taxonomic abbreviations: Ami=*Acropora millepora*; Aqu,=*Amphimedon queenslandica*; Cba=*Carukia barnesi*; Che=*Clytia hemisphaerica*; Hec,=*Hydractinia echinata*; Hma=*Hydra magnipapillata*; Hsy=*Hydractinia echinata;* Mce=*Myxobolus cerebralis*; Mse=*Metridium senile*; Nve=*Nematostella vectensis*; Pc=*Podocoryne carnea*; Phy,=*Polypodium hydriforme*; Tbr=*Tetracapsuloides bryosalmonae*. * = > 70 bootstrap support, **=>90 bootstrap support

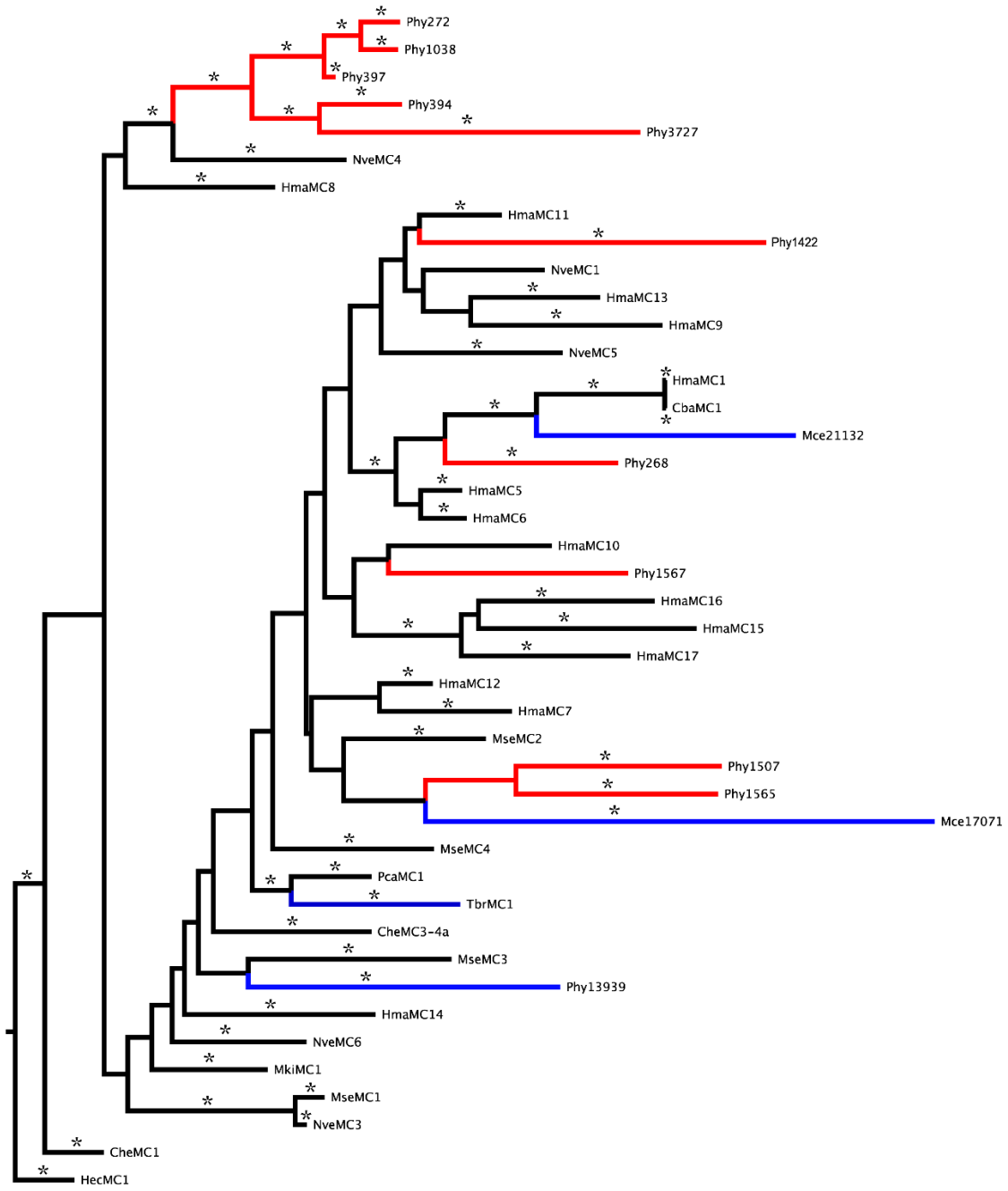


Figure 13. Phylogenetic relationships of minicollagen sequences including new sequences from *P. hydriforme* and *M. cerebralis* inferred by Bayesian analysis. Taxonomic abbreviations: Taxonomic abbreviations: Ami=*Acropora millepora*; Aqu=*Amphimedon queenslandica*; Cba=*Carukia barnesi*; Che=*Clytia hemisphaerica*; Hec,=*Hydractinia echinata*; Hma=*Hydra magnipapillata*; Hsy=*Hydractinia echinata;* Mce=*Myxobolus cerebralis*; Mse=*Metridium senile*; Nve=*Nematostella vectensis*; Pc=*Podocoryne carnea*; Phy,=*Polypodium hydriforme*; Tbr=*Tetracapsuloides bryosalmonae*. Posterior probabilities > 0.95 are indicated by *.

Discussion

Recent phylogenomic studies have made great progress towards solidifying the phylogenetic placement of *P. hydriforme* and Myxozoa (Jimenez-Guri et al. 2008; Evans et al.2010; Nesnidal et al. 2013), but there are still many unanswered questions about the origin and evolution of these organisms. According to Nesnidal et al. (2013), "More detailed analyses of myxozoan genomes in the future will enhance our understanding of metazoan evolution by revealing the genetic underpinnings that drive these profound changes during myxozoan evolution." Here, we characterize the transcriptomes of *Polypodium hydriforme* and the myxozoan, *Myxobolus cerebralis*, to generate data that may help inform us about their phylogenetic placement and evolutionary history.

Characterization of the transcriptomes allows us to uncover genes associated with major changes in morphology and life cycle which can be used for further developmental studies. In addition, transcriptome sequences can be used in future phylogenetic studies, which, given the highly divergent nature of their DNA sequences, is difficult to obtain using traditional PCR approaches. In characterizing these transcriptomes, we had a few major goals in mind: identify genes that would inform the placement of these two enigmatic groups among cnidarians, and characterize genes that may help us understand their evolutionary origins and the evolution of their unusual life cycles and body plans. This study provides an overall description of these transcriptomes. In addition, it provides valuable preliminary data for genes that appear promising for use in future studies with more taxon sampling, which might shed more light on where specifically in Cnidaria myxozoans and *P. hydriforme* should be placed. Lastly, our characterization of developmental regulatory genes should help in understanding the genetic

pathways that may be involved in the transition to a parasitic life style and morphological changes that accompanied this transition.

Particular methodological issues arise that are associated with working with such divergent, parasitic organisms. Perhaps the most pressing was identification of potential contaminating sequences from the host organism. Due to the sensitivity of the sequencing technology, contamination is a serious and issue in next-generation sequencing (Hoy et al. 2013; Oyola et al. 2013), which may be particularly important for researchers who study parasites or endosymbiotic organisms. For example, Jiminez-Guri et al. (2010), found that putative Hox genes from the myxozoan *Buddenbrockia plumatellae*, which led Anderson et al. (1998) to conclude that this was evidence of a bilaterian origin of Myxozoa, were contaminant sequences from their bryozoan host.

Ultimately, for the *P. hydriforme* transcriptome, we found that strategies based on BLAST-searching, even using the actual transcriptome of the target host organism, could not specifically select contaminant sequences without also selecting potential cnidarian sequences. The set of sequences captured by BLAST against contaminant databases was very dependent on BLAST algorithm choice and stringency setting, and could not be considered a reliable estimate of the amount of contamination present in the transcriptome. Potentially, a strategy based on the elimination of contaminants from the *P. hydriforme* transcriptome at the raw read stage, rather than post-assembly, may prove more effective. Additionally, further analysis will be carried out to test these methods on the *M. cerebralis* transcriptome in order to detect contamination.

In addition to the issue of contamination, functional annotation and gene orthology analyses are particularly difficult with non-model and highly divergent organisms. Because such

large percentages of *P. hydriforme* and *M. cerebralis* sequences with BLAST hits returned a generalized annotation of 'predicted protein', a comprehensive means of classifying and assigning functions to sequences produced by assemblies is needed. If the end goal of a study is to fully characterize a family of genes or completely understand the gene expression profile of a particular life-cycle stage, a way of identifying functional regions or general categories for these genes is important. Other means of providing more detailed description of the genes, outside of the traditional BLAST and GO annotation pipeline, is necessary to more thoroughly describe their transcriptome. Fortunately, a number of searching processes have been developed since the advent of next generation sequencing. These search methods identify shorter, conserved sequences or group similar proteins into ortholog/paralog groups. The protein-domain based methods, such as HMMer which searches against protein-family databases, appear particularly promising.

Using HMMer, thousands of sequences from the *P. hydriforme* and *M. cerebralis* transcriptomes were assigned to protein domain families that failed with recover tblastx hits or only had tblastx hits with no functional information (i.e. 'Predicted proteins' or 'Hypothetical Proteins'). Using this methodology, a new candidate Hox/ParaHox sequence and a new minicollagen sequence was discovered for use in further phylogenetic analysis. Overall, a multi-faceted approach, which includes blast searches, protein domain searches, orthology predictions and GO annotation, shows promise for expanding our understanding of a given transcriptome, and is useful for characterizing transcriptomes in non-model systems which do not have full genomic resources available. In particular, these techniques may prove important for organisms too diverged from their nearest model-organism for proper identification of homologous genes

using BLAST searching alone, or for genes that have conserved domains but a wide amount of variation in the rest of the sequence.

Characterization and of these transcriptomes provided important data that will allow us, in future studies, to answer long-standing questions about the evolution of Myxozoa and *P. hydriforme*. An initial step towards understanding the complexities of these body plans and life cycle stages is to investigate known gene families of important developmental regulatory genes, such as the Hox/ParaHox family and the Wnt gene family. Diversification, loss and changes in the regulation of these genes has been shown to be involved in the diversity of body plans throughout Metazoa in general (Schubert and Holland 2000; Garcia-Fernandez 2005) and cnidarians in particular (Guder et al. 2006; Finnerty and Martindale 2001). Thus it is likely that these genes also played a key role in body plan modifications in these parasites.

Originally discovered in mice and *Drosophila*, homeobox-containing genes were found to play a major role in antero-posterior body patterning (Gehrig and Hiromi 1986). Hox genes, and their sister cluster, the ParaHox genes, are thought to be part of the ancestral developmental toolkit of bilaterians. Initial studies of complete cnidarian genomes, showed an initially surprising amount of diversity for these gene families, such as 8 genes in *Hydra* and 15 in *Nematostella* (Chiori et al. 2005). In a single life-cycle stage of *P. hydriforme,* we recovered a large number of homeodomain-containing genes, and six putatively from the Hox/Parahox extended gene family that are being expressed. From the transcriptome of *M. cerebralis*, only one highly divergent putative Hox/Parahox-extended sequence was recovered (Table 4), indicating that *M. cerebralis* may have a reduced complement of this gene family.

Wnt genes in bilaterians largely play a role in body axis formation and the guidance of gastrulation (Sigfried and Perrimon 1994). In cnidarians, the Wnt gene family plays a diverse set of roles in development and regeneration (Guder et al. 2006). Analysis of Wnt gene families in *Nematostella* indicates that it has all of the Wnt gene groups found in bilaterians with the exception of Wnt9 (Kusserow et al. 2005). In addition, there have been several cnidarian-specific duplications within the Wnt7 and Wnt8 groups in *Nematostella* (Kusserow 2005). We found that *P. hydriforme* also has a diversity of Wnt-like sequences. No Wnt- like sequences were recovered from *M. cerebralis.* The eight Wnt-like genes recovered from the *P. hydriforme* may be orthologs to the cnidarian Wnt7/8, Wnt5, Wnt3 and Wnt6 and include a lineage-specific additional duplication within the Wnt7/8 group (Figs. 10 and 11).

There are several possible explanations of our unexpected result that *P. hydriforme* possesses nearly the full complement of cnidarian genes whereas no Wnt genes and only a single putative Hox-/ParaHox-like sequence were found in the *M. cerebralis* assembly. A simple explanation is that that none of these genes are expressed in *M. cerebralis* during the life cycle stage from which the RNA was extracted (triactinomyxon). This is also a possible explanation for the lack of Wnt and Hox/ParaHox genes from certain families in the *P. hydriforme* assembly. Although a transcriptomic study aids the discovery of candidate functional genes by focusing on expressed genes, this may also be a weakness as the expression of genes may be is usually temporally or tissue-specific life-stage or tissue-specific.

Another potential explanation as to why we did not recover more of these gene families for *M. cerebralis* is that the sequences were present but too divergent to identify by methods that compare the sequences directly to full genes from other species (i.e. BLAST techniques). The extremely divergent sequences of the putative Hox/ParaHox-like gene from *M. cerebralis* as

compared to the rest of the sequences suggests that the degree of sequence divergence may indeed be high for this organism. However the protein-domain searches we conducted, provide a way of only look for short conserved domains and thus these searches would be expected to identify these genes if they were present in the transcriptome. HMMer protein-domain searching did not recover any Wnt and only the one divergent putative Hox/ParaHox-like sequences from the *M. cerebralis* assembly, making it unlikely that additional sequences were missed due to sequence divergence.

The third, most intriguing hypothesis about the lack of developmental gene diversity in *M. cerebralis*, is that these genes were lost because of the extreme reductions in body plan or genome size due to the evolution of a parasitic life cycle. The actinospore stage, from which the RNA was extracted, only reaches sizes of about 150 micrometers, and contains three polar capsules and 64 germ cells contained in a sporoplasm (El-Matbouli 1996, Markiw 1992, El-Matbouli et al. 2002). This is in contrast with *P. hydriforme*, which appears to have retained a number of complex, cnidarian traits including the presence of tentacles, several nematocyst types and a free-living stage with epithelia, a mouth and gut (Raikova 2004). These complex structures require developmental patterning mechanisms in *P. hydriforme* that are perhaps not needed to form the simple body plan of *M. cerebralis*.

There have been a number of recent studies regarding the genome size and content of parasites, and the results appear mixed as to whether genome size reduction or gene loss is widespread throughout parasitic taxa. The genomes of the microsporidia, which are fungus-like obligate intracellular parasites, are a striking example of genome compaction and reduction (Peyretaillade et al. 2011). Human parasites *Encephalitozoon cuniculi* and *Encephalitozoon*

intestinalis have two of the smallest eukaryotic genomes known, 2.9 and 2.3 Mbp respectively (Peyretaillade et al. 2011). Part of this major reduction in genome size appears to be due to loss of major gene families, as well as compaction of introns and intergenic spacers. For example, the Tor pathway, which is used for nutrient-sensing and is essential to many processes in other eukaryotes, is completely absent in Microsporidia. (Peyretaillade et al. 2011). This is thought to be due to the microsporidian parasite's complete reliance on its host. *Myxobolus cerebralis* may have undergone similar loss of the whole Wnt gene family, which was no longer used due to the relative simplicity of the myxozoan body plan.

In contrast, a study on parasitic nematodes reveals that its genome is actually larger than closely related model nematode species, and it in fact contains more genes in certain categories related to metabolism (Dieterich 2008). Likewise, a recent study of parasitic nematodes by Liu et al. (2012) found a large number of genes encoding excretory-secretory proteins shown to be important in host/parasite interactions. However, a comparative study of prokaryotic obligate parasites of humans found that obligatory intracellular parasites had reduced genome size and found that reductive evolution was common (Sakharkar et al. 2004), although it is unclear whether these findings can be extrapolated to eukaryotic endocellular parasites. In a genome size study of hymenopterans, researchers found that parasitism alone might not be enough to explain genome size fluctuations among the group (Ardila-Garcia et al. 2010). These mixed results are consistent with our findings that *M. cerebralis*, although lacking expressed key developmental regulatory genes, did not have a reduction in total number of assembled transcripts produced as compared with *P. hydriforme;* 76,434 sequences for *P. hydriforme* and 83,398 for *M. cerebralis.*

Adaptations to parasitism may come in the form of adjustments to the expression or existence of very particular gene families, rather than genome-scale losses. The above studies,

particularly those of Dieterich et al. (2008) and Ardila-Garcia et al. (2010), indicate that certain categories of genes may be overexpressed in parasitic members of particular group. However, it appears that more studies of gene expression in divergent parasitic taxa may be needed to further confirm this assumption. It is possible that *M. cerebralis* and *P. hydriforme* may represent different evolutionary approaches to the problems associated with parasitism, with *Myxobolus* being much more reduced in body structure complexity and body-plan related gene families.

The transcriptomes characterized in this study were used to investigate the molecular components of the nematocysts, the complex intracellular stinging structures characteristic of Cnidaria. Recent phylogenomic evidence for the placement of a myxozoan within cnidaria (Nesnidal 2013) suggests that myxozoan polar capsules are homologous to cnidarian polar capsules. Previously, one minicollagen gene was reported from a myxozoan (Holland et al. 2008). We recovered multiple minicollagen sequences from both *M. cerebralis* and *P. hydriforme,* confirming their position within Cnidaria, and providing important data for future studies on the evolution of the minicollagen gene family and nematocysts. Similar to the developmental regulatory genes, we have identified a greater number and variety of minicollagen sequences from *P. hydriforme* than *M. cerebralis* (Table 6). The number of minicollagen proteins varies between cnidarian taxa, and appears to correlate with the number and complexity of nematocyst types in a given group (David et al. 2008), with Anthozoa having fewer minicollagens and less diversity of nematocysts than Medusozoa. For comparison, the medusozoan (hydrozoan) *Hydra vulgaris* has 17 types of minicollagens, and the anthozoans *Nematostella* and *Metridium* have five and four minicollagens, respectively (David et al., 2008). The *M. cerebralis* complement of minicollagens is even less than that of *Metridium senile*, which has the smallest complement found so far, with four unique minicollagen

sequences (David et al. 2008). This reduction in gene family size may be due reduced body plan and genome complexity as an adaptation to parasitism, and may relate to the difference in life cycles between *M. cerebralis* and *P. hydriforme*.

Although there are no clear functional groupings within the minicollagen sequences, it appears that minicollagens sequences may have some phylogenetic signal and that a more comprehensive sampling from cnidarians may help resolve deep nodes in cnidarian phylogeny. Throughout the tree, there are multiple occurrences of minicollagens from the same species grouped together, although this signal is less consistent in the Bayesian inference (Fig. 13). It also appears that using minicollagens for phylogenetic analysis can separate sequences from Anthozoa and Medusozoa throughout the tree, particularly in ML analysis (Fig. 12). In both reconstructions, there is at least one instance of a *M. cerebralis* sequence grouping with *P. hydriforme* sequences, which may lend some support to the idea that Myxozoa and *P. hydriforme* are sister taxa. The presence and number of minicollagens found in these assemblies does provide further support for the hypothesis that *P. hydriforme* and the myxozoans are derived, parasitic cnidarians.

Conclusions

We have generated and characterized transcriptomes for two enigmatic, cnidarian parasites, *P. hydriforme* and *M. cerebralis*. In this process we implement and integrated diverse methods for contamination filtering and annotation that should serve as a guide for future transcriptome characterizations of non-model organisms. Our finding of the presence of multiple nematocyst-specific genes provide compelling support the placement of these taxa in the phylum Cnidaria, confirming recent phylogenomic studies supporting this placement (Jimenez-Guri

2008, Nesnidal 2013). Preliminary data supports the sister-group relationship of these taxa, but more taxon sampling is needed. Confirming the placement of *Polypodium* and Myxozoa within Cnidaria and determining their exact phylogenetic position will enable us to better understand the origin of these groups and the evolutionary innovations necessary for their transition to parasitism. There has been no published study to date that utilizes multiple markers for the placement of *Polypodium*. Preliminary phylogenetic studies using 454 EST data did not prove informative (C. Dunn, unpublished data). Future studies using Illumina generated transcriptomes and whole genomes in a comprehensive sampling of cnidarians should prove informative in placing *Polypodium* and Myxozoa precisely within Cnidaria. The diverse complement of expressed developmental regulatory genes and minicollagens in *P. hydriforme* and the almost complete lack in *M. cerebralis* may mirror the greater complexity in morphology and life history in *P. hydriforme* compared to *M. cerebralis*. This work provides data for a comparative framework for future studies investigating the evolution of these enigmatic parasites from free-living cnidarians.

References

Abascal F, Zardoya R, Posada D. (2005) ProtTest: Selection of best-fit models of protein evolution. Bioinformatics: 21(9):2104-2105.

Adamczyk, P., et al. (2008). "Minicollagen-15, a novel minicollagen isolated from *Hydra*, forms tubule structures in nematocysts." Journal of Molecular Biology 376(4): 1008-1020.

Altschul, S. F., et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Research 25(17): 3389-3402.

Ardila-Garcia, A. M., et al. (2010). "An expansion of the genome size dataset for the insect order Hymenoptera, with a first test of parasitism and eusociality as possible constraints." Insect Molecular Biology 19(3): 337-346.

Berrill, N. J. (1950). "Development and Medusa-Bud Formation in the Hydromedusae." The Quarterly Review of Biology 25(3): 292-316.

Bartosova, P., et al. (2009). "Concatenated SSU and LSU rDNA data confirm the main evolutionary trends within myxosporeans (Myxozoa: Myxosporea) and provide an effective tool for their molecular phylogenetics." Molecular Phylogenetics and Evolution 53(1): 81-93.

Canning, E. U., et al. (2000). "A new class and order of myxozoans to accommodate parasites of bryozoans with ultrastructural observations on *Tetracapsula bryosalmonae* (PKX organism)." Journal of Eukaryotic Microbiology 47(5): 456-468.

Canning, E. U., et al. (2008). "Early development of the myxozoan *Buddenbrockia plumatellae* in the bryozoans *Hyalinella punctata* and *Plumatella fungosa*, with comments on taxonomy and systematics of the Myxozoa." Folia Parasitologica 55(4): 241-255.

Cannon, Q. and E. Wagner (2003). "Comparison of discharge mechanisms of Cnidarian cnidae and Myxozoan polar capsules." Reviews in Fisheries Science 11(3): 185-219.

Cartwright, P., et al. (2006). "Expression of a Gsx parahox gene, Cnox-2, in colony ontogeny in Hydractinia (Cnidaria : Hydrozoa)." Journal of Experimental Zoology Part B-Molecular and Developmental Evolution 306B(5): 460-469.

Chen, F., et al. (2006). "OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups." Nucleic Acids Research 34: D363-D368.

Chiori, R., et al. (2009). "Are Hox Genes Ancestrally Involved in Axial Patterning? Evidence from the Hydrozoan *Clytia hemisphaerica* (Cnidaria)." Plos One 4(1): e4231.

Chourrout, D., et al. (2006). "Minimal ProtoHox cluster inferred from bilaterian and cnidarian Hox complements." Nature 442(7103): 684-687.

Conesa, A., et al. (2005). "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research." Bioinformatics 21(18): 3674-3676.

David, C. N., et al. (2008). "Evolution of complex structures: minicollagens shape the cnidarian nematocyst." Trends in Genetics 24(9): 431-438.

Dieterich, C., et al. (2008). "The Pristionchus pacificus genome provides a unique perspective on nematode lifestyle and parasitism." Nature Genetics 40(10): 1193-1198.

Drummond AJ, Ashton B, Buxton S, Cheung M, Cooper A, Duran C, Heled J, Kearse M, Markowitz S, Moir R, Stones-Havas S, Sturrock S, Swidan F, Thierer T, Wilson A (2012) Geneious v5.6, Available from http://www.geneious.com

Duffy, D. J., et al. (2010). "Wnt signaling promotes oral but suppresses aboral structures in Hydractinia metamorphosis and regeneration." Development 137(18): 3057-3066.

Elmatbouli, M., et al. (1995). "Light and electron-microscopic observations on the route of the triactinomyxon-sporoplasm of *Myxobolus cerebralis* from epidermis into rainbow-trout cartilage." Journal of Fish Biology 46(6): 919-935.

Evans, N. M., et al. (2008). "Phylogenetic placement of the enigmatic parasite, Polypodium hydriforme, within the Phylum Cnidaria." Bmc Evolutionary Biology 8.

Evans, N. M., et al. (2008). "Phylogenetic placement of the enigmatic parasite, Polypodium hydriforme, within the Phylum Cnidaria." Bmc Evolutionary Biology 8.

Finn, R. D., et al. (2011). "HMMER web server: interactive sequence similarity searching." Nucleic Acids Research 39: W29-W37.

Finnerty, J. R. and M. Q. Martindale (1999). "Ancient origins of axial patterning genes: Hox genes and ParaHox genes in the Cnidaria." Evolution & Development 1(1): 16-23.

Garcia-Fernandez, J. (2005). "The genesis and evolution of homeobox gene clusters." Nature Reviews Genetics 6(12): 881-892.

Gehring, W. J. and Y. Hiromi (1986). "Homeotic Genes and the Homeobox." Annual Review of Genetics 20(1): 147-173.

Grabherr, M. G., et al. (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome." Nature Biotechnology 29(7): 644-U130.

Grabner, D. S. and M. El-Matbouli (2010). "Experimental transmission of malacosporean parasites from bryozoans to common carp (Cyprinus carpio) and minnow (Phoxinus phoxinus)." Parasitology 137(4): 629-639.

Gruhl, A. and B. Okamura (2012). "Development and myogenesis of the vermiform Buddenbrockia (Myxozoa) and implications for cnidarian body plan evolution." Evodevo 3.

Guder, C., et al. (2006). "The Wnt code: cnidarians signal the way." Oncogene 25(57): 7450-7460.

Hafner, M. S., et al. (1994). "Disparate Rates of Molecular Evolution in Cospeciating Host and Parasites." Science 265(5175): 1087-1090.

Haft, D. H., et al. (2001). "TIGRFAMs: a protein family resource for the functional identification of proteins." Nucleic Acids Research 29(1): 41-43.

Hall, T. A. (1999). "BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT." Nucleic Acids Symposium Series 41: 95-98.

Hedrick, R. P. and M. El-Matbouli (2002). "Recent advances with taxonomy, life cycle, and development of Myxobolus cerebralis in the fish and oligochaete hosts." American Fisheries Society Symposium 29: 45-53.

Hide, W. A. and R. D. Isokpehi (2004). "Positive selection scanning of parasite DNA sequencing." Methods in Molecular Biology 270: 127-150.

Holland, J. W., et al. (2011). "A novel minicollagen gene links cnidarians and myxozoans." Proceedings of the Royal Society B-Biological Sciences 278(1705): 546-553.

Holstein, T. (2009). Wnt Signaling in Cnidarians. Wnt Signaling. E. Vincan, Humana Press. 469: 47-54.

Hoy, M. A., et al. (2013). "Transcriptome sequencing and annotation of the predatory mite Metaseiulus occidentalis (Acari: Phytoseiidae): a cautionary tale about possible contamination by prey sequences." Experimental and Applied Acarology 59(3): 283-296.

Ibragimov, A. and E. Raikova (2004). "Nematocysts of *Polypodium hydriforme*, a cnidarian parasite of acipenseriform fishes." Hydrobiologia 530: 165-171.

Jimenez-Guri, E., et al. (2008). "Buddenbrockia is a Cnidarian worm." Science 317(5834): 116-118.

Kallert, D. M., et al. (2005). "Polar filament discharge of Myxobolus cerebralis actinospores is triggered by combined non-specific mechanical and chemical cues." Parasitology 131: 609-616.

Kent, M. L., et al. (2001). "Recent advances in our knowledge of the Myxozoa." J Eukaryot Microbiol 48(4): 395-413.

Kurz, E. M., et al. (1991). "Mini-collagens in *Hydra* nematocytes." Journal of Cell Biology 115(4): 1159-1169.

Kusserow, A., et al. (2005). "Unexpected complexity of the Wnt gene family in a sea anemone." Nature 433(7022): 156-160.

Li, L., et al. (2003). "OrthoMCL: Identification of ortholog groups for eukaryotic genomes." Genome Research 13(9): 2178-2189.

Lipin, A. (1911). "Die Morphologie und Biologie von Polypodium hydriforme Uss." Zoologische Jahrbuecher Jena Abteilungen fur Anatomie 31: (317-420).

Lipin, A. (1925). “Geschlechtliche Form, Phylogenie und systematische Stellung von Polypodium hydriforme Ussov.” Zool Jahrb Anat 1925, 47:541-635

Miller, MA; Holder, MT; Vos, R; Midford, PE; Liebowitz,T; Chan, L.; Hoover, P; Warnow, T. The CIPRES Portals. CIPRES. 2009-08-04. URL:http://www.phylo.org/sub_sections/portal. Accessed: 2013-2-20. (Archived by WebCite(r) at http://www.webcitation.org/5imQlJeQa)

Liu, X., et al. (2012). "Global gene expression analysis of the zoonotic parasite *Trichinella spiralis*. Revealed Novel Genes in Host Parasite Interaction." PLoS Negl Trop Dis 6(8): e1794.

Lom, J., Dyková, I., (1992). Protozoan parasites of fishes. Developments in Aquaculture and Fisheries Science, vol.26, Elsevier, 315 pp.

Lom, J. and I. Dykova (2006). "Myxozoan genera: definition and notes on taxonomy, life-cycle terminology and pathogenic species." Folia Parasitologica **53**(1): 1-36.

Markiw, M. E. and K. Wolf (1983). "*Myxosoma cerebralis* (myxozoa, myxosporea) etiologic agent of salmonid whirling disease requires tubificid worm (annelida, oligochaeta) in its life-cycle." Journal of Protozoology 30(3): 561-564.

Nei, M. and R. K. Chesser (1983). "Estimation of fixation indexes and gene diversities." Annals of Human Genetics 47(JUL): 253-259.

Nesnidal, M. P., et al. (2013). "Agent of Whirling Disease Meets Orphan Worm: Phylogenomic Analyses Firmly Place Myxozoa in Cnidaria." Plos One 8(1).

Oyola, S. O., et al. (2013). "Efficient Depletion of Host DNA Contamination in Malaria Clinical Sequencing." Journal of Clinical Microbiology 51(3): 745-751.

Parker, G. A., et al. (2003). "Evolution of complex life cycles in helminth parasites." Nature 425(6957): 480-484.

Peyretaillade, E., et al. (2011). "Extreme reduction and compaction of microsporidian genomes." Research in Microbiology 162(6): 598-606.

Philipp, I., et al. (2009). "Wnt/beta-Catenin and noncanonical Wnt signaling interact in tissue evagination in the simple eumetazoan Hydra." Proceedings of the National Academy of Sciences of the United States of America 106(11): 4290-4295.

Phillips, S.R. (2001). Parasitism: Life Cycle and Host Defences against Parasites. eLS, John Wiley & Sons, Ltd.

Raikova, E. V. (1994). "Life cycle, cytology, and morphology of *Polypodium hydriforme*, a coelenterate parasite of the eggs of acipenseriform fishes." The Journal of Parasitology 80(1): 1-22.

Raikova, E. V. (2008). "Cytomorphological peculiarities of Polypodium hydriforme (Cnidaria)." Journal of the Marine Biological Association of the United Kingdom 88(8): 1695-1702.

Reft, A. J. and M. Daly (2012). "Morphology, distribution, and evolution of apical structure of nematocysts in hexacorallia." Journal of Morphology 273(2): 121-136.

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

Sakharkar, K. R., et al. (2004). "Genome reduction in prokaryotic obligatory intracellular parasites of humans: a comparative analysis." International Journal of Systematic and Evolutionary Microbiology 54: 1937-1941.

Schlegel, M., et al. (1996). "Phylogenetic analysis of complete small subunit ribosomal RNA coding region of Myxidium lieberkuehni: Evidence that Myxozoa are Metazoa and related to the Bilateria." Archiv Fur Protistenkunde 147(1): 1-9.

Schubert M, Holland LZ. The Wnt Gene Family and the Evolutionary Conservation of Wnt Expression. In: Madame Curie Bioscience Database [Internet]. Austin (TX): Landes Bioscience; 2000-. Available from: http://www.ncbi.nlm.nih.gov/books/NBK6212/

Siegfried, E. and N. Perrimon (1994). "*Drosophila* wingless - a paradigm for the function and mechanism of wnt signaling." Bioessays 16(6): 395-404.

Smothers, J., et al. (1994). "Molecular evidence that the myxozoan protists are metazoans." Science 265(5179): 1719-1721.

Sonnhammer, E. L. L., et al. (1997). "Pfam: A comprehensive database of protein domain families based on seed alignments." Proteins-Structure Function and Genetics 28(3): 405-420.

Stamatakis, A. (2006). "RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models." Bioinformatics 22(21): 2688-2690.

Steele, R. E., et al. (2011). "A genomic view of 500 million years of cnidarian evolution." Trends in Genetics 27(1): 7-13.

Tops, S., et al. (2005). "Diversity and systematics of the Malacosporea (Myxozoa)." Invertebrate Biology 124(4): 285-295.