ECHINODERM TRANSCRIPTOMICS

by

Gregorio Villaflor Linchangco Junior

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology

Charlotte

2018

Approved by:

_____
Dr. Daniel A. Janies

_____
Dr. Robert R. Reid

_____
Dr. Xinghua Shi

_____
Dr. Jessica Schlueter

_____
Dr. Ian Marriott

ABSTRACT

GREGORIO VILLAFLOR LINCHANGCO JUNIOR. Echinoderm Transcriptomics. (Under the direction of DR. DANIEL A. JANIES)

Tissue regeneration and biomineralization are expressed to a diverse extent across metazoans and their life stages. The potential for repair and regrowth in adult stages varies widely within phyla, class and species. For instance, few adult human tissues can regenerate. In contrast, members of the phylum Echinodermata demonstrate remarkable regenerative capabilities. Holothuroids like the sea cucumber can regenerate vital organs after evisceration, while the echinoid sea urchin lacks this ability. Echinoderms have been model organisms for studies in embryonic developmental biology due to their abundant gametes and often clear embryos. More recently, adult echinoderms have emerged as models in regenerative studies. The ability of echinoderms to repair and regrow body parts as a response to injury or predation is valuable in studies of the basic mechanisms that underpin regeneration. The heterogeneity of regenerative capabilities within echinoderm classes provides insights into how regeneration is gained and lost. From an evolutionary standpoint, echinoderms share a common ancestor with chordates which include humans. Resolving the phylogenetic relationships of echinoderms provides a platform to understand the gain and loss of regeneration and may have future applications in medicine.

Modern echinoderms occur in two major extant lineages, Crinozoa and Eleutherozoa. The evolutionary relationships within Eleutherozoa have remained ambiguous. The motivation for this dissertation is to resolve competing hypotheses in the

evolutionary relationships within Eleutherozoa and examine competing hypotheses explaining the expansion of the *msp130* family. The *msp130* gene family is related to biomineralization - an important process in development and regeneration of echinoderms.

I developed a novel analytical pipeline that produces phylogenetic trees from raw RNA-Seq data of 40 echinoderms. I also performed an analysis surveying the heterogenous regenerative capabilities across echinoderm classes by identifying enriched gene ontology terms that implicate biological processes of interest using class-specific datasets. Using a similar pipeline and the same transcriptome data I exploited to examine taxonomic relationships, my results provide support for an alternative hypothesis regarding the origin of the *msp130* family within echinoderms. The phylogenetic analysis suggests that *msp130* radiated from a deep common ancestral gene set via a complex series of organismal speciation and gene duplication events, rather than multiple independent instances of horizontal gene transfer.

DEDICATION

I dedicate this work to my family and friends who have supported me throughout my academic career.

ACKNOWLEDGEMENTS

First, I express my gratitude to my doctoral advisor Daniel A. Janies, who has served as a great mentor, challenging me to think critically. I would also like to extend my appreciation to the members of my committee, Dr. Robert R. Reid, Dr. Xinghua Shi, Dr. Jessica Schlueter and Dr. Ian Marriott for their valuable counsel during my academic journey. Special thanks to Dr. Cynthia Gibas and the Department of Bioinformatics and Genomics, the College of Computing and Informatics and the Graduate School, UNC Charlotte and the US National Science Foundation who generously funded my endeavors as a doctoral candidate.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SIO-BIC: Scripps Institution of Oceanography, Benthic Invertebrate Collection.

KSORC: Korea South Pacific Ocean Research Center (Chuuk)

FLMNH: Florida Museum of Natural History

Msp130: Mesenchyme specific protein 130 kDa

Msp130rel1: Primary mesenchyme specific protein MSP130-related-1

Msp130rel2: Primary mesenchyme specific protein Msp130-related-2

Msp130rel3: Primary mesenchyme specific protein Msp130-related-3

Msp130rel4: Primary mesenchyme specific protein Msp130-related-4

Msp130rel5: Primary mesenchyme specific protein Msp130-related-5

Msp130rel6: Primary mesenchyme specific protein Msp130-related-6

Msp130rel7: Primary mesenchyme specific protein Msp130-related-7

GO ID: Gene Ontology Term Identifier

CHAPTER 1: INTRODUCTION AND BACKGROUND

Echinoderms are a phylum that demonstrate remarkable regenerative capabilities across a high level of taxonomic diversity. Echinoderms are the second largest clade of deuterostomes after chordates and some species can regenerate vital organs after evisceration [1]. These marine invertebrates have long been used as model organisms for studies in embryonic developmental biology due to their low maintenance in the laboratory and position in the tree of life [2]. For similar reasons, echinoderms have recently emerged as models in regenerative studies. Their ability to repair and regrow body parts makes them a valuable group to study in the largely unknown mechanisms driving regenerative biology [3].

The heterogeneity of regenerative capabilities within echinoderm classes provides insights for therapies that could treat human conditions. However, applied research of echinoderm regeneration using gene expression is rare and undermined by a lack of understanding of their evolutionary relationships [3].

Echinoderms are deuterostome invertebrates that share at least four unique morphological synapomorphies that include: pentaradial symmetry, a water-vascular system, a mesodermal skeleton, and mutable collagenous tissue [4]. The five extant classes of echinoderms include sea stars (Asteroidea), sea urchins (Echinoidea), brittle stars (Ophiuroidea), sea cucumbers (Holothuroidea) and sea lilies (Crinoidea). There are over 7000 described, living echinoderm species making it the second largest group of deuterostomes after the chordates [5]. The close phylogenetic relationship between echinoderms and chordates along with the rich diversity of echinoderms holds

implications across many comparative disciplines including evolutionary biology, systematics, developmental biology and studies of regeneration.

The purple sea urchin, *Strongylocentrotus purpuratus*, is a prime example of the role that echinoderms have played in scientific research. Due to the simplicity of the early development of *Strongylocentrotus purpuratus* and the and low maintenance requirements in the laboratory, *Strongylocentrotus purpuratus* has served as the model organism in embryonic development studies for over a century [2]. Recently, this sea urchin has played a role in biomedical research as it was the first echinoderm for which a completely sequenced and annotated genome was completed. This work has led to the discovery of genes orthologous to those implicated in human disease [6].

The completion of the genome of *Strongylocentrotus purpuratus* highlights the importance of the availability of next generation sequence data to facilitate new discoveries within the diverse echinoderm phylum. In this dissertation I go a step further and use RNA-Seq to profile transcriptomes of 40 exemplars of the phylum Echinodermata [7].

The motivation for this dissertation is to contribute basic research that can be applied to studies in regenerative echinoderm biology through a pipeline based on transcriptome data. This goal is largely driven by resolving competing hypotheses regarding the evolutionary relationships within Eleutherozoa and challenging a proposed idea that the gene related to the regenerative process of biomineralization, *msp130*, was introduced into echinoderms via multiple instances of horizontal gene transfer.

Using RNA-Seq transcriptome data from 40 echinoderms, I developed a novel pipeline that takes raw RNA-Seq data as input and produces phylogenetic trees. Along

the way, I performed an exploratory study considering the heterogenous regenerative capabilities across echinoderm classes. I performed an annotation analysis on the gene content of these 40 transcriptomes to identify enriched gene ontology terms and their associated biological pathways related to regeneration based on class-specific data. Using these techniques on a large-scale comparative transcriptomic dataset, this dissertation addresses three basic research challenges related to regenerative echinoderm biology:

1) Resolving the evolutionary relationships within extant echinoderms.

2) Gaining a better understanding of regenerative heterogeneity amongst echinoderms.

3) Investigating the evolutionary history of *msp130* within echinoderms, a principal member of a gene family implicated in the regenerative process of biomineralization.

1.1 Phylogenetic Reconstruction

Phylogenetic trees are a central organizing framework for comparative studies, yet a debate remains with two competing hypotheses for echinoderm relationships at the class level. This controversy has been further confounded by the discovery of an enigmatic echinoderm, *Xyloplax*, that has been argued to represent a new, sixth taxonomic class. While previous phylogenetic studies have been conducted, most have relied on morphological data or few sequence loci based on Sanger technology [8]. The genome of *Strongylocentrotus purpuratus* has provided many insights into echinoderm biology via studies of gene families [9]. However, taking into consideration the rich morphological and functional diversity of echinoderms there are likely vast amounts of

additional information in the genetic sequences of non-echinoid echinoderms that can be uncovered by from comparative genomic study. Like the genome of *Strongylocentrotus purpuratus*, annotation of the 40 assembled transcriptomes is proving to be a valuable resource [10].

1.2 Annotation of Transcriptomes

Comparative phylotranscriptomic analyses will not only provide an overview of the genetic content within extant echinoderms but also elucidate novel relationships within the historically defined echinoderm classes. The exploration of these new relationships can lead to new insights in the biological processes found within echinoderms. One such process that is important across the animal kingdom is regeneration. Echinoderms possess high regenerative potential and are able to express this to a maximum extent, with species capable of complete regrowth from body fragments alone [1].

1.3 Biomineralization and *msp130*

Echinoderms are capable of two distinct regenerative processes, soft tissue regeneration which includes nerves and muscles associated with tube feet, and biomineralization of the spine. The latter regenerative process, biomineralization, is defined as the biologically controlled formation of mineral deposits resulting in structures that function as support, protection, or feeding anatomy [11]. Biomineralization is wide spread amongst many organisms however, echinoderms and vertebrates are the two phyla

within deuterostomes that form extensive biomineralized structures [12]. All adult echinoderms develop endoskeletal elements which are formed by specialized cells known as primary mesenchyme cells (PMCs) through the process of skeletogenesis. PMCs are heavily involved in this process and express a variety of proteins associated with biomineralization. One such family of proteins is known as the Msp130 family. [11]. The Msp130 family consists of eight members, (Msp130, Msp130-related-1, Msp130-related-2, Msp-130-related-3, Msp130-related-4, Msp-130-related-5, Msp130-related-6, Msp130-related-7) Msp130 was the first PMC-specific protein to be identified in 1987 [13] followed by Msp130-related-1 and Msp130-related-2 in 2002 [14], then Msp130-related-3, Msp130-related-4, Msp-130-related-5, Msp130-related-6 in 2006 [11] and lastly Msp130-related-7 in 2014 [15]. Recently, a worker evaluating the evolutionary relationships of the *msp130* gene within eukaryotes (including echinoderms) and prokaryotes proposed a provocative theory. The worker hypothesized that *msp130* was introduced to echinoderms and molluscs via multiple independent instances of horizontal gene transfer events, facilitated through bacteria or algal intermediates [15]. Horizontal gene transfer events are extremely rare in the animal kingdom and other simpler explanations must exist given enough data.

These challenges in regenerative echinoderm biology will be addressed by the following chapters of this dissertation, resolving several queries. In chapter 2, the following questions are presented: What phylogeny of extant echinoderms is supported by a large transcriptome dataset? Will the resulting phylogeny support the Cryptosyringid or Asterozoa-Echinozoa hypothesis? In chapter 3, I ask the following questions: Can the preferred phylogeny derived from chapter 2 be used to study the

variation of 40 transcriptomes assemblies among echinoderm clades? Will certain functions be enriched based on class-specificity? In chapter 4 I ask a question involving biomineralization. Studies with limited taxonomic sampling have suggested that the biomineralization gene family, *msp130*, was introduced to echinoderms and molluscs via horizontal transfer events either directly from bacteria or algal intermediates [15]. Will a novel large-scale molecular dataset corroborate or refute the findings of this study?

1.4 Objective I – Echinoderm Phylogeny

The shared relationship between echinoderms and chordates within deuterostomes along with their rich diversity in body plans and larval ecologies allow echinoderms to serve as important models in comparative studies across many disciplines including developmental evolution, regulatory regions of genomes, and immune systems. However, the lack of molecular data and unresolved phylogeny undermines their use as a model system in other studies such as regeneration. The objective is to harness the power of next generation sequencing technologies specifically, RNA-Seq, to resolve phylogenetic ambiguities of echinoderms at the class level. Extant echinoderms are considered to consist of two higher level clades, Crinoidea and Eleutherozoa. There are currently two main hypotheses regarding evolutionary relationships within Eleutherozoa. One is the cryptosyringid hypothesis, which posits that the classes Echinoidea, Holothuroidea and Ophiuroidea form a clade known as Cryptosyringida, sister to the Asteroidea. The cryptosyringid hypothesis is based on the putative synapomorphy of enclosed radial elements within the water-vascular system of adults and has been

supported by various morphological studies [16,17] as well as molecular studies [16,18,19]. An alternative hypothesis for the evolutionary relationships of extant Eleutherozoa is the Asterozoa-Echinozoa hypothesis. This hypothesis is based on the synapomorphy of the adult body plan, whereby Asterozoa comprises the stellate forms of echinoderms (asteroids and ophiuroids), while Echinozoa includes the globoid forms (holothuroids and echinoids). The Asterozoa-Echinozoa concept has been supported by morphological studies [20,21] as well as the molecular analyses of Sanger sequencing data [22,23].

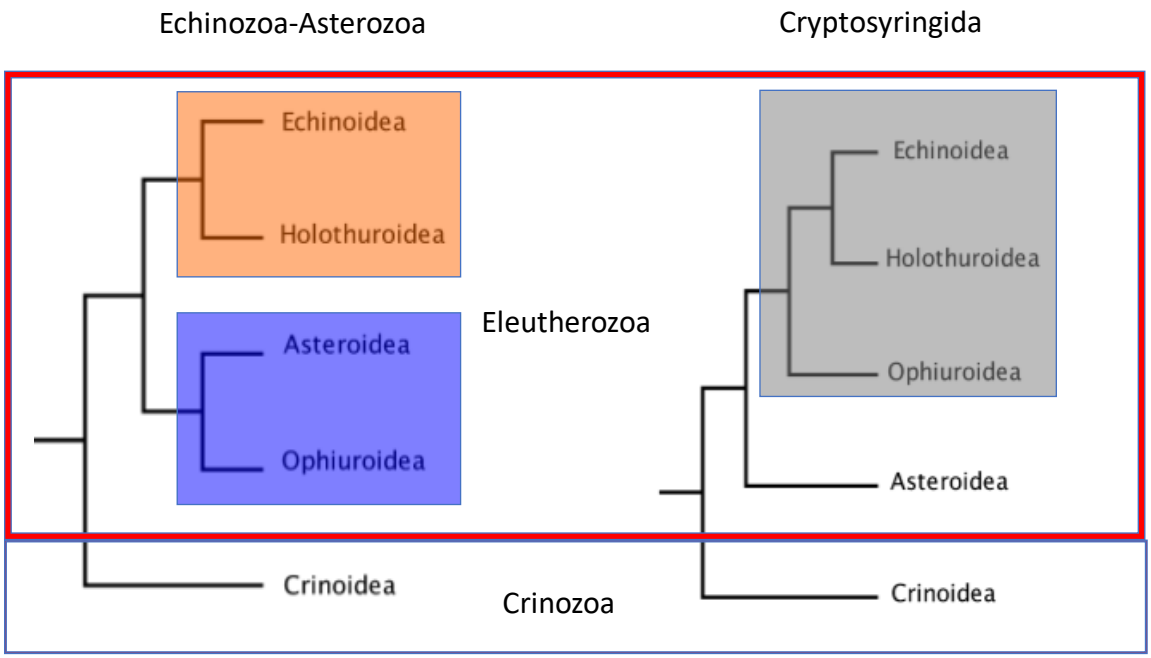

Figure 1.1 The two competing hypotheses regarding relationships within Eleutherozoa. Within the red rectangle, the tree on the left-hand side depicts the Echinozoa-Asterozoa hypothesis and the tree on the right-hand side displays the Cryptosyringid hypothesis. This former hypothesis is based on the synapomorphy of the adult body plan, whereby Asterozoa comprises the stellate forms of echinoderms (asteroids and ophiuroids), while Echinozoa includes the globoid forms (holothuroids and echinoids). The cryptosyringid hypothesis is based on the putative synapomorphy of enclosed radial elements within the water-vascular system of adults.

Objective one is designed to use a large-scale transcriptomic dataset to corroborate or refute these hypotheses and to resolve the placement of the species *Xyloplax*. Objective one has been separated into three sub-objectives below.

Objective 1.1 Reconstruct a diverse sample set of echinoderms spanning deep divergences with consistent RNA-Seq methods and assemble their transcriptomes.

Objective 1.2 Enhance the taxonomic and molecular resolution of relationships within Eleutherozoa, and test for the Echinozoa-Asterozoa vs Cryptosyringid hypotheses.

Objective 1.3 Resolve the placement of the enigmatic *Xyloplax* within echinoderms.

1.5 Objective II –Annotation of transcriptomes of extant echinoderms

Annotation of transcriptomes is a crucial step in adding to the knowledgebase of echinoderm biology as it evaluates the content of the 40 newly assembled transcriptomes. The characterization of gene function based on expressed sequence data will contribute to the field of developmental and evolutionary echinoderm biology. Just as the genome of the purple sea urchin ushered in new discoveries in echinoderm biology, the wealth of data present in 40 novel transcriptomes can provide new insights into echinoderms at the molecular level. The objective is to profile extant echinoderm transcriptomes using existing bioinformatics techniques. The results of this work allow for future comparative transcriptomics studies across echinoderm classes. One outcome can be the identification of gene regulatory networks that allow for echinoderm regeneration.

Objective 2.1 Assign annotations to the 40 assembled echinoderm transcriptomes.

Objective 2.2 Visualize an overview of patterns and relationships between the transcriptomes based on Gene Ontology annotations.

Objective 2.3 Using a phylogenetic approach to map areas of commonality and difference to identify and visualize candidate expressed genes involved in echinoderm diversification.

1.6 Objective III – Biomineralization in echinoderms

The understanding of echinoderm phylogeny and the classification of its accompanying transcript variants is important in the study of the developmental biological processes. One such biological process that echinoderms possess is their great potential for regeneration. Regeneration is a characteristic type of developmental process involving tissue repair, cell turnover, reconstruction of external and internal organs. Echinoderms exhibit extraordinary regenerative capabilities and can regenerate full adult forms from detached body elements [24]. The lack of the genomic information regarding regeneration has limited the understanding of the cellular pathways and mechanisms involved in its process. A biological process closely related with regeneration in echinoderms is biomineralization. Fossilized biomineralized animal remains are the principal source of evidence on extinct lineages [25,26]. Though some studies in this field have been performed, a vast resource of novel transcriptomes such as this work presents has not been studied as a dataset. Profiling the gene expression data of 40 new transcriptomes will add knowledge of the pathways involved in the process of biomineralization. The objective is to identify expressed *msp130* genes associated with

the regenerative process of biomineralization based on sequence homology. Msp130 stands for mesenchyme-specific protein, 130 kDa. Understanding the relationships of the *msp130* gene family orthologs within echinoderm can provide insights into the biochemical and cellular pathways that control the regenerative abilities of echinoderms. Fundamental research of the phylogenetic relationships within the *msp130* gene family will lay down the groundwork for studies in biomineralization and regeneration within echinoderms. This work fuels the possibility for regenerative therapies downstream. Basic research on genes of interest similar to green fluorescent proteins (GFPs) which was initially discovered in cnidarians, and later copepods and cephalochordates have become valuable reagents for measuring molecular and cellular properties [27].

In objective three I will develop a better understanding of the *msp130* gene family and biomineralization in echinoderms.

Objective 3.1 Identify biomineralization-related proteins within 40 echinoderm transcriptomes using known proteins for sequence similarity searches. This includes the genes mesenchyme-specific protein, 130 kDa, mesenchyme-specific protein, 130 kDa related 1, mesenchyme-specific protein, 130 kDa related 2, mesenchyme-specific protein, 130 kDa related 3, mesenchyme-specific protein, 130 kDa related 4, mesenchyme-specific protein, 130 kDa related 5, mesenchyme-specific protein, 130 kDa related 6 and mesenchyme-specific protein, 130 kDa related 7. (Msp130rel1-7).

Objective 3.2 Perform phylogenetic analyses of the *msp130* family of genes.

1.7 Phylogenetic methods and orthology

Several key evolutionary concepts involved in the phylogenetic methods must be defined to gain a better understanding of this dissertation. One of these core concepts is biological sequence homology. Sequence homology between two different species can be defined in terms of shared ancestry. Biological sequences can have shared ancestry in one of two ways 1) orthology or 2) paralogy. Sequences that are inferred to have descended from the same ancestral sequence separated by a speciation event are said to be orthologous. In other words, when a species diverges into two separate species the instances of each gene in each of the new species are orthologs to one another [28]. In contrast, paralogous genes share ancestry stemming from gene duplication events that occurred within one or both of the two species being compared [29]. These concepts are important in both phylogenetic reconstruction of an entire phylum (Chapter 2) as well as the phylogenetic gene tree construction of the *msp130* gene family (Chapter 4). Speciation events largely direct the methods used in Chapter 2 while both organismal speciation and gene duplication events help explain the methods and results derived in Chapter 4.

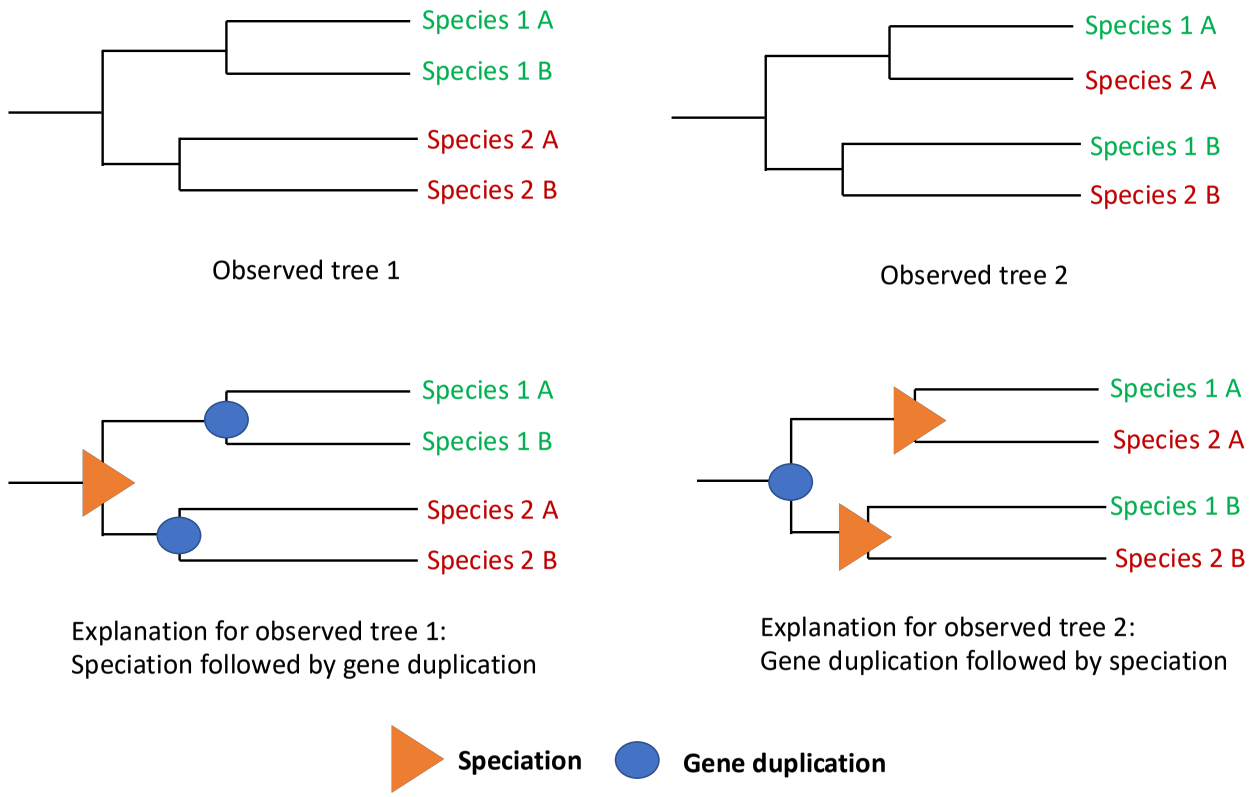

Figure 1.2: Speciation and Gene Duplication example. Letters A and B denote genes, gene A and its copy gene B.

CHAPTER 2: ECHINODERM PHYLOGENY

2.1 Introduction to the phylum

The evolution of this phylum has been well documented by a rich fossil record, placing their emergence around 530-524 MYA during the Cambrian Period [30]. Echinoderms are members of the superphylum, Deuterostomia, a grouping that consists of two other phyla, Chordata and Hemichordata. Echinodermata is most closely related to its sister group Hemichordata and the two form a clade of invertebrates known as Ambulacraria [31]. Although the five extant echinoderm classes share features that separate them from other organisms including the four synapomorphies mentioned above, each class exhibits diverse larval and adult forms. The adult body plans of echinoderms cover a wide spectrum, from the ambulatory and stellate Asteroidea and Ophiuroidea (starfish and brittle stars), soft bodied Holothuroidea (sea cucumbers), spiky armored Echinoidea (sea urchins), to the stalked and un-stalked Crinoidea (sea lilies). These five classes are grouped into two clades, the Crinoidea and Eleutherozoa (classes Asteroidea, Ophiuroidea, Echinoidea, and Holothuroidea). Crinoidea are thought to have diverged from the four other classes around 485-515 MYA [32]. Eleutherozoa is thought to have rapidly separated into four classes around 480 MYA within a 5 Myr window [19,33]. This swift and ancient radiation of Eleutherozoa has been problematic in inferring the evolutionary relationships among these classes.

The ambiguity of internal Eleutherozoan relationships undermines the use of other echinoderms as model systems in biology, where sea urchins have become experimental models. Sea urchins are widely studied by developmental biologists because of the ability of the animals to produce massive amounts of eggs and sperm. Moreover, the

transparency of the embryo allows for the study of skeletogenesis and organogenesis in living embryos. Many other echinoderm species exhibit these advantageous characteristics and along with the rich diversity in echinoderm larval and adult body plans. As a result these organisms have been used for comparative study of the evolution of body plans across deep time divergences [34]. Defining clear relationships within Eleutherozoa using novel transcriptomes will allow for new model systems to emerge.

Evolutionary relationships of extant echinoderms have long been debated, more specifically relationships within Eleutherozoa. The ancient rapid radiation of this subphylum has raised questions regarding its internal relationships bringing upon many studies that have often employed molecular methods (See Table 2.1). A phylogenetic study using larval and adult morphologies as well as molecular data from two ribosomal DNA proposed three incongruent topologies [18]. Several other studies employed molecular approaches reconstructing echinoderm phylogenies using mitochondrial protein coding genes [35–41]. Additional molecular based studies have used nuclear and mitochondrial gene sequences to estimate echinoderm phylogeny. Most of these studies have resulted in the monophyly of the extant classes however, the topologies observed remain dependent on the methods used to construct them (Janies, 2001;Janies et al., 2011). Thus I have taken a sensitivity analysis approach that considers increasingly permissive missing data [42].

The rapid radiation of Eleutherozoa will greatly benefit from data produced by next-generation sequencing technologies like RNA-Seq. The most recent studies have taken advantage of these technologies enabling phylogenetic reconstruction using transcriptome data as shown in Table 2.1[33,43–45].

Table 2.1: From left to right, this table depicts recent studies of echinoderm phylogeny with the total numbers of associated genes, unique sampling representatives of each extant echinoderm class, outgroups added, and echinoderms.

Author	Number of Loci	Asteroidea	Holothuroidea	Echinoidea	Ophiuroidea	Crinoidea	Outgroup Added	Total Echinoderms
(D. Janies, 2001)	2	20	5	10	7	6	4	48
(Daniel A. Janies et al., 2011)	7	35	9	22	12	3	5	81
(Telford et al., 2014)	219	2	2	4	N/A	1	4	9
(O'Hara et al., 2014)	425	4	1	N/A	52	1	2	60
(Cannon et al., 2015)	185	4	5	3	3	2	19	17
(Reich et al., 2015)	219	14	4	5	4	1	6	28
This work	1-1256	15	8	4	4	9	2	40

These studies have begun to converge on the Asterozoa-Echinozoa hypothesis but have included limited loci and limited taxonomic representatives of the five extant echinoderm classes (e.g. the relatively small numbers illustrated in white cells of table 2.1). Furthermore, transcriptomic studies prior to Linchangco et al. (2017), lack the enigmatic *Xyloplax*. In the 1980s, two species of *Xyloplax*, a small disc-shaped echinoderm were discovered on sunken wood in the deep sea off the Bahamas and New Zealand [46]. *Xyloplax* species were initially described as having a circular water vascular system [46,47] but later was re-described as a derived pentaradial system with ambulacra splayed out to the periphery of the body [8,22,48]. The small size and odd morphology fueled the already fierce debate over echinoderm phylogeny [22,30,49]. Some workers erected a new sixth class of extant echinoderms containing only *Xyloplax* [46]. Subsequently, [50] proposed a working hypothesis uniting *Xyloplax* and valvatid asteroids but in doing so expressed a desire to keep *Xyloplax* as a class distinct from the asteroids, which is self-contradictory. In contrast, several researchers have argued that *Xyloplax* is an aberrant asteroid and that erecting a new class is not warranted [8,22,23,30].

Resolving the evolutionary relationships in echinoderms is essential to advance their use as regenerative models. Ambiguous nodes at the subphylum and class levels can be resolved using taxonomic sampling that spans deep time divergences and includes the morphologically unusual *Xyloplax*. These topologies produced by this dataset provides support for the placement *Xyloplax* within the context of Eleutherozoa and Asteroidea.

2.2 Research Design and Methods

To better understand the phylogeny of extant echinoderms and the placement of *Xyloplax*, I used RNA-Seq to retrieve large numbers of orthologous loci from each transcriptome of 40 unique echinoderm species spanning the deepest divergences within the five extant classes. Using these loci, I discovered orthologous sequences that are well represented in asteroids, holothuroids, echinoids, ophiuroids and crinoids for the construction of phylogenetic trees explaining the topology of relationships within Eleutherozoa and the placement of *Xyloplax*. To investigate the effects of locus selection and alignment quality in the phylogenetic reconstruction, a sensitivity analysis of distinct subsets of orthologous loci based on alignment occupancy and gap percentage and was applied. To reconstruct a representative phylogenetic tree of a phylum that contains approximately 7000 described extant species requires several considerations [5]. To accomplish Objective 1, sampling included members of clades that span the largest evolutionary divergence. This includes 40 species from 24 orders and 37 families collected from deep and polar seas. There are 16 asteroids, 4 ophiuroids, 9 holothuroids, 4 echinoids, and 9 crinoids included in the dataset. RNA-Seq was used to profile these 40 organisms using Illumina Hi-Seq 2000, 100 bp paired-end reads.

Table 2.2: Detailed descriptions of sampled echinoderms in this study.

Class	Order	Family	Genus	Species	NCBI BioProject Accession
Asteroidea	Concentricycloidea(infraclass?)	Xyloplacidae	*Xyloplax*	sp.	PRJNA299326
Asteroidea	Spinulosida	Echinasteridae	*Echinaster*	*spinulosus*	PRJNA299548
Asteroidea	Velatida	Pterasteridae	*Pteraster*	*tesselatus*	PRJNA299398
Asteroidea	Forcipulatida	Asteriidae	*Pisaster*	*ochraceus*	PRJNA299406
Asteroidea	Velatida	Korethrasteridae	*Peribolaster*	*folliculatus*	PRJNA299409
Asteroidea	Paxillosida	Astropectinidae	*Psilaster*	*charcoti*	PRJNA299410
Asteroidea	Forcipulatida	Labidiasteridae	*Labidiaster*	*annulatus*	PRJNA299414
Asteroidea	Velatida	Korethrasteridae	*Remaster*	*gourdoni*	PRJNA299412
Asteroidea	Paxillosida	Luidiidae	*Luidia*	*clathrata*	PRJNA299463
Asteroidea	Spinulosida	Echinasteridae	*Henricia*	*leviuscula*	PRJNA299471
Asteroidea	Paxillosida	Astropectinidae	*Astropecten*	*duplicatus*	PRJNA299417
Asteroidea	Valvatida	Poraniidae	*Glabraster*	*antarctica*	PRJNA299415
Asteroidea	Valvatida	Asteropseidae	*Asteropsis*	*carinifera*	PRJNA299419
Asteroidea	Notomyotida	Benthopectinidae	*Cheiraster*	*antarcticus*	PRJNA299420
Asteroidea	Brisingida	Brisingidae	*Odindella*	*nutrix*	PRJNA299897
Crinoidea	Comatulid	Colobometridae	*Oligometra*	*serripinna*	PRJNA299887
Crinoidea	Comatulida	Bourgueticrinidae	*Democrinus*	*brevis*	PRJNA299465
Crinoidea	Hyocrinida	Hyocrinidae	*Gephyrocrinus*	*messingi*	PRJNA300546
Crinoidea	Comatulida	Ptilometridae	*Ptilometra*	*australis*	PRJNA299466
Crinoidea	Comatulida	Comasteridae	*Cenolia*	*trichoptera*	PRJNA299468
Crinoidea	Comatulida	Antedonidae	*Phrixometra*	*nutrix*	PRJNA299469
Crinoidea	Comatulida	Antedonidae	*Isometra*	*vivipara*	PRJNA299411
Crinoidea	Comatulida	Antedonidae	*Promachocrinus*	*kerguelensis*	PRJNA299478
Crinoidea	Comatulida	Zenometridae	*Psathryometra*	*fragilis*	PRJNA299480
Echinoidea	Echinoida	Strongylocentrotidae	*Strongylocentrotus*	*purpuratus*	PRJNA299888
Echinoidea	Arbacioida	Arbaciidae	*Arbacia*	*punctulata*	PRJNA299547
Echinoidea	Cidaroida	Cidaridae	*Eucidaris*	*tribuloides*	PRJNA299418
Echinoidea	Clypeasteroida	Dendrasteridae	*Dendraster*	*excentricus*	PRJNA299549
Holothuroidea	Apodida	Synaptidae	*Synapta*	*maculata*	PRJNA299890
Holothuroidea	Dendrochirotacea	Psolidae	*Psolus*	sp.	PRJNA299550
Holothuroidea	Aspidochirotida	Stichopodidae	*Stichopus*	*chloronatus*	PRJNA299896
Holothuroidea	Aspidochirotida	Synallactidae	*Paelopatides*	*confundens*	PRJNA299551
Holothuroidea	Dendrochirotacea	Cucumariidae	*Abyssocucumis*	*cf. albatrossi*	PRJNA299552
Holothuroidea	Aspidochirotida	Synallactidae	*Pseudostichopus*	sp.	PRJNA299883

Phylogenetic reconstruction was performed using a pipeline developed for starting with raw sequence data (See Figure 2.1). After sequencing from Duke Center for Genomic and Computational Biology, raw reads from Illumina Hi-Seq 2000 were assessed for quality using FASTX toolkit [51]. *De novo* assembly of contigs was performed using Trinity [52]. The resulting contigs from each sample were translated into protein space using Transdecoder (http://transdecoder.github.io/) and the PFAM-B protein family database with a cutoff of 100 amino acids [53]. The resulting protein sequences were compared against one another via BLASTP to discover "orthoclusters" in OrthoMCL [54]. Orthoclusters are defined as groups of similar sequences that are orthologous or paralogous. The orthoclusters were then filtered to include 75% of the (30/40) total taxa. The longest sequences in each of these orthoclusters was identified and used as the query in a BLASTP search against a two taxa hemichordate and cephalochordate outgroup; *Saccoglossus kowalevskii* (NCBI taxon 10224), *Branchiostoma floridae* (NCBI taxon 7739).

After the creation of this dataset, each of the orthoclusters was aligned using MAFFT [55], a multiple sequence alignment program. MAFFT drastically reduces the CPU time required to perform multiple sequence alignment by rapidly identifying homologous regions using a fast Fourier transform that is applied to the amino acid sequence. This is especially important when large numbers of taxa and sequences need to be aligned. After the orthoclusters were aligned, RAxML [56] was used to construct putative gene trees using the command PTHREADS-SSE3 -T 16 -f a -p $RANDOM -x $RANDOM -#100 -m PROTGAMMAAUTO -s ${FILE}.phy -n $, which created a gene tree dataset. The trees produced by this analysis are then further analyzed for paralogous

sequences with the use of PhyloTreePruner [57]. Although methods have been developed for orthology prediction, paralogous sequences can still be erroneously grouped together with orthologous sequences, affecting the results of phylogenetic reconstruction with supermatrix methods [58]. PhyloTreePruner is a software utility that uses a phylogenetic approach to refine orthology inferences made using phenetic methods. This software will check single gene trees for evidence of paralogy and generates a new alignment for each group containing only sequences inferred to be orthologous. One of the important features of PhyloTreePruner is its ability to collapse poorly supported nodes into polytomies, avoiding unnecessarily discarding sequences in cases where a weakly supported tree topology incorrectly supports paralogy. In a study that includes entire transcriptomes yielding thousands of putative genes, incongruence among gene trees are sure to arise. To address this issue of incongruence, a combined analysis approach was used. Accurate and efficient and alignments are crucial in phylogenetic analysis. Thus, I performed a quality check on alignments after paralogy assessment. The check was a two-step process using the program TrimAL [59] and the custom python script that I developed called "BOXER" (See Figure 2.2). First, TrimAL removed difficult to align sequences via an automated command line interface employing alignment statistics. BOXER then selected from aligned orthoclusters produced by TrimAL, preferring those aligned orthoclusters with the user-defined number of unique taxa and allowed percentage of gaps in the alignment. The difficult to align sequences were identified via the two parameters from TrimAL as described in the manual [60] (1) "residue overlap" which is a percentage of residues in an aligned orthocluster column that must be occupied with other residues (not gaps or missing data), and 2) "sequence

overlap" which is a percentage of positions with observed residues (not gaps or missing data) that a row in an aligned orthocluster must have in order to be kept in the alignment. If a sequence did not fulfill both thresholds for these parameters, it was removed from the orthocluster alignment.

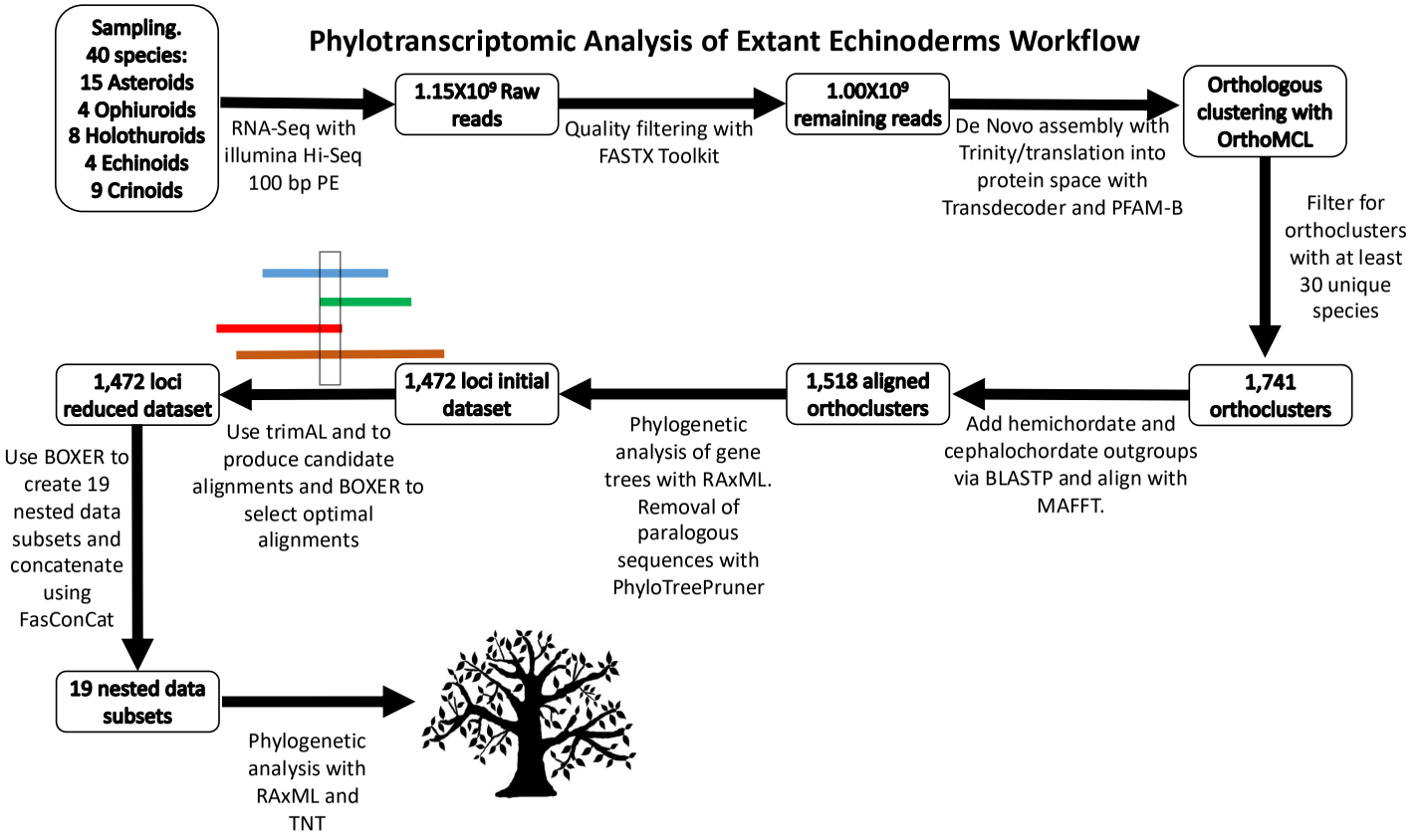

Figure 2.1: Starting in the upper left-hand corner with sampling, this graphic illustrates a high-level overview of the echinoderm phylogeny reconstruction workflow using transcriptome data.

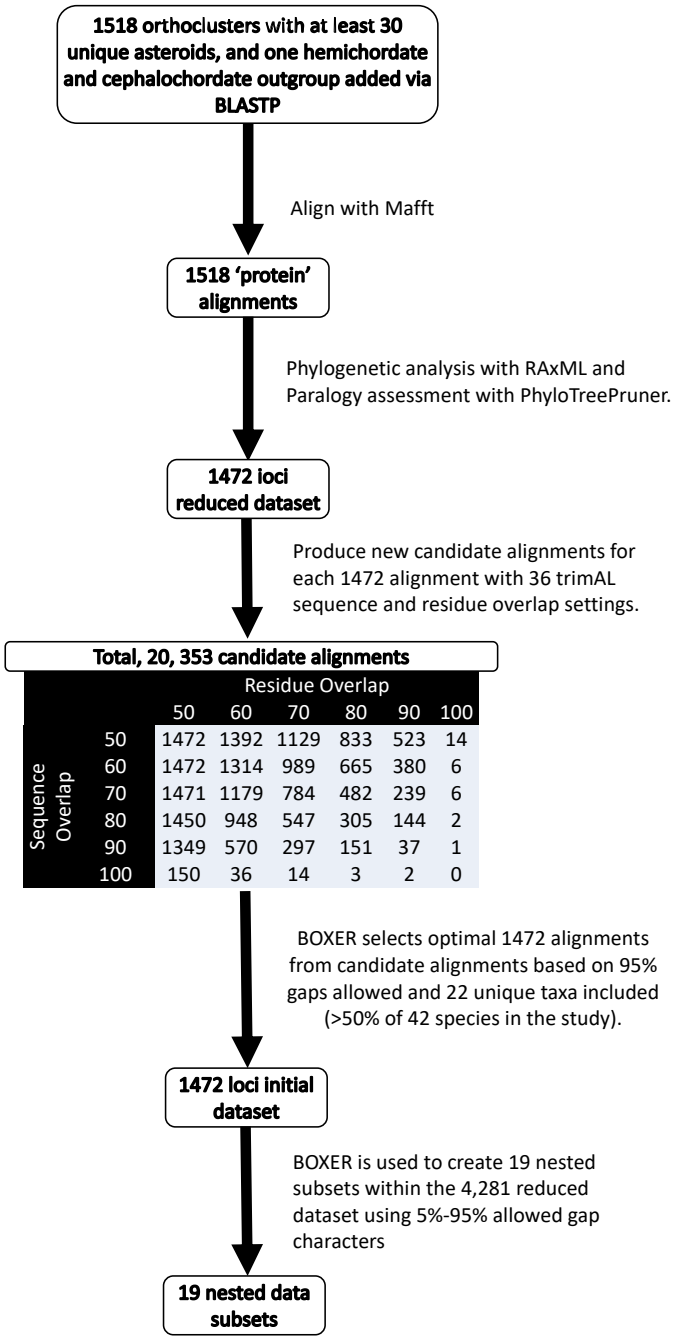

Figure 2.2 The process of selecting alignments using the TrimAL and BOXER programs.

Table 2.3: This table describes the differing methods used to produce orthologs and super matrices across four studies.

Author	Orthology Assessment Pre-supermatrix	Post supermatrix "cleansing"
(Telford et al., 2014)	blastp, tblastn	Phylocriteria filter with 4 out of 5 echinoderm classes represented, Manual visual inspection and removal of poorly aligned regions
(O'Hara et al., 2014)	blastx, Manual empirical assessment Phylocriteria filtering using the inclusion of at least 66% total taxa.	Aliscore (Misof & Katharina, 2009)
(Cannon et al., 2015)	HAMSTER (Ebersberger, Strauss, & von Haeseler, 2009), Phylocriteria filtering for clusters that contain at least 15 ambulacrarian taxa, PTP (Kocot et al., 2013), MARE	Aliscore and Alicut (Mlsof & Katharina, 2009), Manual removal of mistranslated sequences
(Linchangco Jr et al., 2017), This work	Markov Cluster algorithm grouping putative orthologs and paralogs via OrthoMCL (Li, Stoeckert, & Roos, 2003), Phylocriteria filtering via BOXER and TrimAL (Capella-Gutiérrez, Silla-Martínez, & Gabaldón, 2009) for alignment selection, PhyloTreePruner (Kocot et al., 2013) for paralogy assessment.	None

To retain as much data as possible, the most inclusive alignments with a high percentage of gap characters were selected. This was performed using thresholds that allow for 95% gap characters and at least 50% of the original species retained within an

alignment. Once paralogous sequences have been removed, the resulting alignments are concatenated into a "supermatrix" using the software FasconCAT [61]. FasconCAT creates a supermatrix by extracting each taxon associated amino acid sequences out of a given set of sequence alignments and links them together into one string. The resulting supermatrix allows for the simultaneous analysis of combined data from all alignments in the dataset and is used as input for phylogenetic maximum likelihood analysis in RAxML. Phylogenetic analyses on the supermatrix was then performed using the RAxML with a CAT model of rate heterogeneity.

2.3 Results

The procurement of echinoderm samples was facilitated through the NSF funded echinoTOL project (National Science Foundation Grant No. 1322141). This project recognized the challenges in assembling the echinoderm tree of life and built a team of experts specializing in paleontology, genomics, informatics, developmental biology, anatomy and phylogenetics. Samples collected by this team were expertly curated and provided a diverse sampling of extant echinoderms that led to a venerable set of novel transcriptome data. This data was used in deciphering the ambiguous evolutionary relationships within the echinoderm phylum at an unprecedented scale. The large sampling of echinoderms allowed for the resolution of evolutionary relationships within Eleutherozoa and provided support for the Echinozoa-Asterozoa hypothesis. The phylogenetic tree produced by this work provides insights on the placement of *Xyloplax* within the five extant classes of echinoderms.

For the ingroup species, RNA-Seq produced a total of 2,360,841,332 raw reads. Following trimming and adapter removal, 2,101,192,636 reads remained, a reduction of 259,648,696 reads or approximately 11%. The sample from the asteroid *Pisaster ochraceus* had the most reads at 88,987,394, while the asteroid *Cheiraster* sp. had the least amount of reads at 30,190,658. The sample from the featherstar crinoid *Promachocrinus kerguelensis* had the most reads removed, with a decrease of nearly 19%. In contrast, the sample from the stalked crinoid *Gephyrocrinus messingi* had the least reads removed at a reduction of only 3.64%.

Table 2.4: This table describes each sequenced species read counts. Vouchers are abbreviated as following: SIO-BIC Scripps Institution of Oceanography, Benthic Invertebrate Collection. KSORC Korea South Pacific Ocean Research Center (Chuuk), FLMNH Florida Museum of Natural History.

Genus	Species	Voucher	RAW Reads	After Quality filter and adapter removal	Percent Reads remaining	Percentage Reads removed	NCBI BioProject Accession
Abyssocucumis	*cf. albatrossi*	SIO-BIC E6816	47221102	39642599	84%	16%	PRJNA299552
Arbacia	*punctulata*	SIO-BIC E6740	37971266	31511066	83%	17%	PRJNA299547
Asteropsis	*carinifera*	KSORC-000276	55750932	46359435	83%	17%	PRJNA299419
Astropecten	*duplicatus*	SIO-BIC E6745	58536934	47996898	82%	18%	PRJNA299417
Astrophyton	*muricatum*	SIO-BIC E6741	52762224	43885881	83%	17%	PRJNA299886
Cenolia	*trichoptera*	SIO-BIC E6767	41085524	33986308	83%	17%	PRJNA299468
Cheiraster	sp.	SIO-BIC E4758	30190658	25463134	84%	16%	PRJNA299420
Democrinus	*brevis*	SIO-BIC E4759	51874816	49975324	96%	4%	PRJNA299465
Dendraster	*excentricus*	SIO-BIC E5640	59537456	50012042	84%	16%	PRJNA299549
Echinaster	*spinulosus*	specimen consumed in study	58483838	55409121	95%	5%	PRJNA299548
Eucidaris	*tribuloides*	SIO-BIC E6742	38230910	31760367	83%	17%	PRJNA299418
Gephyrocrinus	*messingi*	SIO-BIC E4427	47932756	46184392	96%	4%	PRJNA300546
Glabraster	*antarctica*	SIO-BIC E5520	54906620	45521873	83%	17%	PRJNA299415
Henricia	*leviuscula*	SIO-BIC E5597	61618432	50637189	82%	18%	PRJNA299471
Isometra	*vivipara*	SIO-BIC E5696	63688704	52768945	83%	17%	PRJNA299411
Labidiaster	*annulatus*	SIO-BIC E4756	88022282	84184087	96%	4%	PRJNA299414
Luidia	*clathrata*	SIO-BIC E6744	60248000	50001997	83%	17%	PRJNA299463
Molpadia	*granulata*	SIO-BIC E5790	46380306	39577018	85%	15%	PRJNA299464
Odinella	*nutrix*	SIO-BIC E4241	40650604	33577253	83%	17%	PRJNA299897
Oligometra	*serripinna*	SIO-BIC E4711	87318994	84008407	96%	4%	PRJNA299887
Ophiocoma	*wendtii*	specimen consumed in study	39062818	36460981	93%	7%	PRJNA299897
Ophioderma	*brevispina*	SIO-BIC E6743	55865966	47238004	85%	15%	PRJNA299887
Ophiothrix	*spiculata*	specimen consumed in study	47255758	44667610	95%	5%	PRJNA299898
Paelopatides	sp.	SIO-BIC E5609	62013488	52263682	84%	16%	PRJNA299551
Peribolaster	*folliculatus*	SIO-BIC E4754	77608578	74006052	95%	5%	PRJNA299409
Phrixometra	*nutrix*	SIO-BIC E6829	56526242	47416953	84%	16%	PRJNA299469
Pisaster	*ochraceus*	SIO-BIC E6726	88987394	85025177	96%	4%	PRJNA299406
Promachocrinus	*kerguelensis*	SIO-BIC E4881	42324150	34284514	81%	19%	PRJNA299478
Psathryometra	*fragilis*	SIO-BIC E4567	51808244	47195245	91%	9%	PRJNA299480
Pseudostichopus	sp.	SIO-BIC E6363	38720764	32877879	85%	15%	PRJNA299883
Psilaster	*charcoti*	SIO-BIC E4754	62076466	59370692	96%	4%	PRJNA299410
Psolus	sp.	SIO-BIC E6832	69815772	66794601	96%	4%	PRJNA299550
Pteraster	*tesselatus*	SIO-BIC E6724	63162964	59909679	95%	5%	PRJNA299398
Ptilometra	*australis*	SIO-BIC E4720	52898308	43969637	83%	17%	PRJNA299466
Remaster	*gourdoni*	SIO-BIC E4757	55504406	53433302	96%	4%	PRJNA299412
Stichopus	*chloronotus*	FLMNH 11289	80227376	76995813	96%	4%	PRJNA299896
Strongylocentrotus	*purpuratus*	specimen consumed in study	38738340	35769091	92%	8%	PRJNA299888
Synallactes	sp.	SIO-BIC E5607	55160390	46385880	84%	16%	PRJNA299885
Synapta	*maculata*	FLMNH 11293	67842766	64670226	95%	5%	PRJNA299890
Xyloplax	sp.	SIO-BIC E6809	61076078	57240185	94%	6%	PRJNA299326

During ortholog identification and alignment creation, 93,908 orthoclusters that contained four or more taxa were detected. Taxonomic filters yielded a set of 1518 putative orthoclusters for extant echinoderm classes. Following the removal of highly divergent and paralogous sequences, a dataset of 1472 orthoclusters was created and a sensitivity analysis based on alignment occupancy was performed. The sensitivity analysis tested the independent variable values of gap permissiveness across alignments would impact on the dependent variable, which is the resulting tree topology for each gap-variable alignment. Sensitivity analysis was processed by the "BOXER" program, dividing the 1472 orthoclusters into 19 nested data subsets based on alignment occupancy (percentage gaps allowed in matrix). Phylogenetic analysis performed on these 19 matrices recovered topologies that support the Asterozoa-Echinozoa hypothesis in all but two of the smallest datasets. Of the remaining 17 datasets, we consistently recover: 1) class-level monophyly, 2) monophyly of Eleutherozoa and 3) Crinozoa as sister taxon to Eleutherozoa. For the 10% through 20% allowed gaps datasets, the observed topologies placed *Xyloplax* as sister to *Remaster gourdoni*, *Peribolaster folliculatus* and *Pteraster tesselatus*. Topologies from the 25%-35% allowable gaps datasets placed *Xyloplax* as sister to *Pteraster tesselatus*. The 40% allowable gaps dataset placed *Xyloplax* as sister to *Remaster gourdoni* and *Peribolaster folliculatus*. The remaining datasets from 45% to 95% allowable gaps placed *Xyloplax* as sister to all other asteroids. The five largest datasets with 22 or greater unique taxa and allowable gap character percentages at 75%, 80%, 85, 90%, 95% have no less than 73% bootstrap support at major nodes defining the clades: Asteroidea, Ophiuroidea, Asterozoa, Echinoidea, Holothuroidea, Echinozoa, Eleutherozoa, and Crinoidea. I present the tree of the 90% allowable gaps dataset based

on maximum bootstrap support value for maximum loci included. The dataset producing this tree had no bootstrap support values lower than 81 on all nodes and consisted of 1256 loci.

Table 2.5: This table provides an overview of the results for each nested data subset.

Gap % allowed	Loci	Lowest bootstrap value	Groups recovered	Placement of *Xyloplax*
5	1	11	Class monophyly	N/A
10	17	28	Asterozoa-Echinozoa	sister to *Remaster gourdoni* , *Peribolaster folliculatus* , *Pteraster tesselatus*
15	46	48	Cryptosyringida	sister to *Remaster gourdoni* , *Peribolaster folliculatus* , *Pteraster tesselatus*
20	104	10	Asterozoa-Echinozoa	sister to *Remaster gourdoni* , *Peribolaster folliculatus* , *Pteraster tesselatus*
25	177	22	Asterozoa-Echinozoa	sister to *Pteraster tesselatus*
30	246	49	Asterozoa-Echinozoa	sister to *Pteraster tesselatus*
35	324	21	Asterozoa-Echinozoa	sister to *Pteraster tesselatus*
40	414	21	Asterozoa-Echinozoa	sister to *Remaster gourdoni* , *Peribolaster folliculatus*
45	501	15	Asterozoa-Echinozoa	sister to Asteroids
50	593	20	Asterozoa-Echinozoa	sister to Asteroids
55	702	40	Asterozoa-Echinozoa	sister to Asteroids
60	806	18	Asterozoa-Echinozoa	sister to Asteroids
65	911	0	Asterozoa-Echinozoa	sister to Asteroids
70	1008	0	Asterozoa-Echinozoa	sister to Asteroids
75	1079	78	Asterozoa-Echinozoa	sister to Asteroids
80	1172	89	Asterozoa-Echinozoa	sister to Asteroids
85	1239	86	Asterozoa-Echinozoa	sister to Asteroids
90	1256	81	Asterozoa-Echinozoa	sister to Asteroids
95	1256	73	Asterozoa-Echinozoa	sister to Asteroids

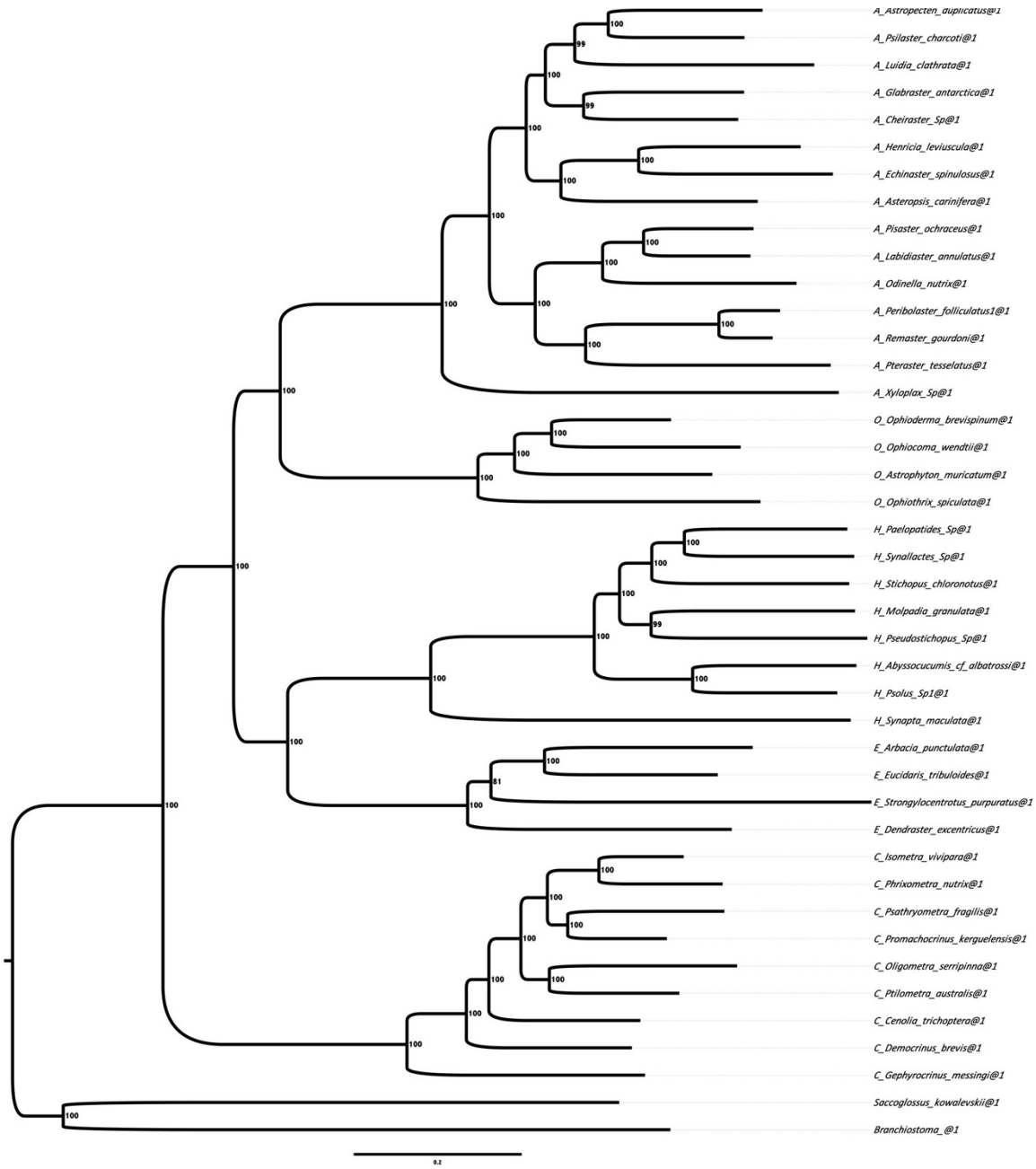

Figure 2.3: Phylogeny of extant echinoderms using a multi-locus transcriptomic dataset of 1256 loci. This allows for 90% indels and at least 22 unique taxa per orthocluster (1256 loci). Here we observe class level monophyly and support for the Asterozoa-Echinozoa hypothesis with *Xyloplax* placed as sister to all asteroids. This topology was selected based on bootstrap support (no nodes lower than 81) and the most inclusive dataset (1256 loci).

2.4 Discussion

The sensitivity analysis indicated that increased gap characters within the alignments allowed for longer sequences and stabilized the relationships within Eleutherozoa including the placement of *Xyloplax* as sister to all remaining asteroids. This notion is supported by increased bootstrap values at all nodes of alignments that allowed for the most gaps (75% through 95% allowed gaps). This observation highlights the importance of including orthoclusters that are less stringently controlled for indels but provide more data via longer sequences. The five most inclusive datasets in this study unequivocally support the Asterozoa-Echinozoa hypothesis and have done so using a much larger (1256 loci) and better sampled (40 ingroup species and two outgroups) dataset than previously achieved (See Table 2.1). The recovery of Echinozoa and Asterozoa in the most inclusive phylogenies indicate the lack of support for the alternative Cryptosyringida hypothesis which was only supported by one small dataset with only 46 loci and low bootstrap support values.

In my analyses, I investigated the five-class echinoderm dataset across 19 distinct data subsets of alignment occupancy. This method stratified our datasets into two groups with regards to the placement of *Xyloplax.* Concatenated alignments composed of more stringent and therefore less loci (17-414) placed *Xyloplax* as a member of velatid asteroids. There are three topologies that describe *Xyloplax* as a velatid asteroid 1) as sister to *Remaster gourdoni, Peribolaster folliculatus* and *Pteraster tesselatus* (10-20% allowable gaps), 2) as sister to *Pteraster tesselatus* (25-35% allowable gaps) and 3) as sister to *Remaster gourdoni* and *Peribolaster folliculatus* (40% allowable gaps). These

results are consistent with those from Sanger sequencing efforts using seven loci, which also placed *Xyloplax* as a velatid asteroid (Janies et al., 2011).

In contrast, results from datasets with large numbers of loci and gaps allowed (i.e. >500 loci and 45-95% allowable gaps) *Xyloplax* is sister to asteroids. The placement of *Xyloplax* as sister to the remaining asteroids has been discussed by others [46]. Mah used Infraclass Concentricycloidea for the placement of *Xyloplax* as sister to infraclass Neoasteroidea [62] (which includes all post Paleozoic asteroids as defined by Gale (1987).

CHAPTER 3: ANNOTATION OF ECHINODERM TRANSCRIPTOMES

3.1 Introduction

The group who published the genome of the purple sea urchin, *Strongylocentrotus purpuratus*, estimates 23,300 total genes [6]. Many of the *Strongylocentrotus purpuratus* genes are similar to well described vertebrate gene families including families those previously thought to be specific to vertebrates [6]. The genome of *Strongylocentrotus purpuratus* also tells us that some gene families occurred independently in echinoderms and vertebrates. Furthermore, molecular data from *Strongylocentrotus purpuratus* revealed the complexity of its revealed its refined immune system including a wide range of pathogen recognition proteins [6]. These data and findings provide bases for the annotation of genes across all extant echinoderms through sequence similarity. Annotation and comparative analyses of the echinoderm transcriptome variation have potential for addressing a wide range of questions. The annotation and comparison of the transcriptomes in this study can answer questions regarding the conserved and unique genes in some adult tissues across the five classes of echinoderms. For example, as some adult echinoderms (Ophiuroids, Asteroids, and Holothuroids) can regenerate tissues and others have limited regenerative ability (Echinoids), this work provides the basic science needed for understanding novel processes in echinoderms such as tissue regeneration.

3.2 Research Design and Methods

To create an annotated dataset, orthoclusters based on phylogenetic criteria were generated using OrthoMCL [54]. Class specific taxonomic filters were applied to obtain five groupings of orthoclusters. To provide an overview of the all five classes, the sequences from the clusters derived from class-specific OrthoMCL were then visualized in OrthoVenn [63]. This program provides an interactive Venn diagram that provides summarizes counts, and the functional intersections and reverse complements of clusters shared between clades includes in-depth views of the clusters using various sequence analysis tools [63]. OrthoVenn also allows for a hypergeometric test for Gene Ontology (GO) enrichment. The Gene Ontology database is among the most widely used gene description databases used for the detection of enriched genes. Gene Ontology terms (GO terms) consists of biological processes, cellular components, and molecular functions that are organized in a directed acyclic graph of parent-child relationships. This method profiled molecular processes or functions that are expressed in a certain phenotype, in this case an echinoderm class [64]. GO enrichment analysis finds the most differentially expressed genes based on GO term annotations for each echinoderm class. Hypergeometric testing for GO enrichment uses a discrete probability distribution that describes the probability of a gene with a specific GO term is selected by random in n number of draws without replacement.

3.3 Results

Using a 1e-5 expectation value cutoff in OrthoVenn analysis, I detected the largest number of total orthoclusters and sequences for asteroids followed by crinoids, holothurians, echinoids and ophiuroids. Asteroids also contained the most singletons (protein sequences that could not be clustered) followed by holothuroids, crinoids, echinoids and ophiuroids. These results are expected as the rank of order of the sequences and clusters reflect the number of taxonomic samples per class (See Table 3.1 below).

Table 3.1 Results of OrthoVenn Clustering with e-value cutoff of 1e-5.

Classes	Sequences	Clusters	Singletons
Asteroidea	170508	30732	31673
Crinoidea	68043	14839	12185
Holothuroidea	56678	13593	12785
Echinoidea	22023	7406	6531
Ophiuroidea	19675	6501	6351

The Venn diagram produced by OrthoVenn provided reverse complements and intersections among orthoclusters of the data described in Table 3.1. Figure 3.1 illustrates an intersection of 865 orthoclusters that contained at least one representative from each echinoderm class. 19,512 orthoclusters were found that only included

asteroids, 2,458 that only included echinoids 1,937 for ophiuroids, 6,378 for crinoids and 6,314 for holothurians.

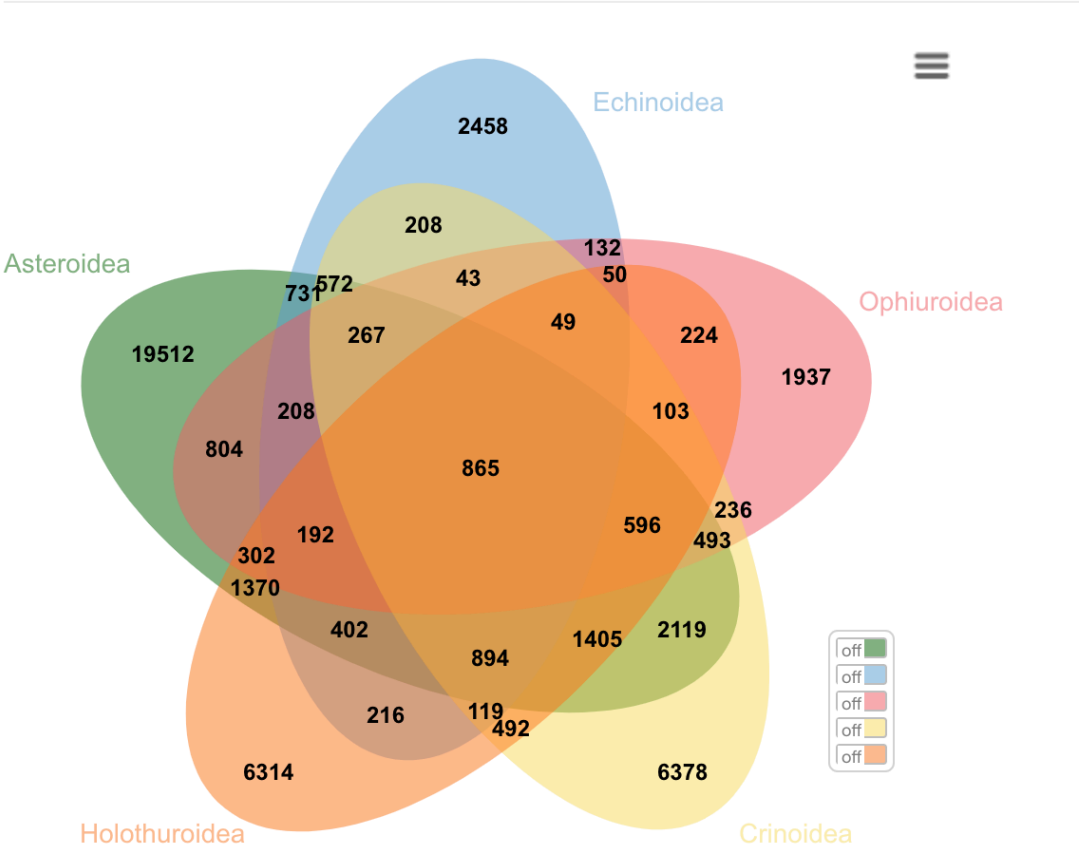

Figure 3.1 OrthoVenn Diagram [63] depicting the intersections and reverse complements of 336,927 echinoderm protein sequences. The values within the diagram indicate the number of orthoclusters found to form distinct groups. The work done here focuses periphery and center of the diagram. The peripheral values show reverse complements, in other words, the orthoclusters unique to each of the five classes (Asteroidea:19,512, Holothuroidea:6,314, Crinoidea:6,378, Ophiuroidea:1,937, Echinoidea: 2,458). The central number (865) is the junction of all classes, or all orthoclusters containing a representative sequence from each class.

The hypergeometric test GO enrichment detected three significantly enriched GO terms with a p-value of <0.05. All enriched GO terms were from the biological processes namespace and are in crinoids and ophiuroids (See Table 3.2).

Table 3.2: The hypergeometric test results of GO enrichment for 336,927 echinoderm protein sequences as produced by OrthoVenn.

Hypergeometric test result of GO enrichment (p-value < 0.05)				
Clade	GO ID	Name	Namespace	p-value
Crinoidea	GO:0050774	negative regulation of dendrite morphogenesis	biological process	0.00286
Crinoidea	GO:0048846	axon extension involved in axon guidance	biological process	0.03793
Ophiuroidea	GO:0034472	snRNA 3'-end processing	biological process	0.018

Within crinoids, the two statistically significant enriched GO IDs GO:0050774 and GO:0048846 are closely related. GO:0050774 is defined by any process that stops, prevents, or reduces the frequency, rate or extent of dendrite morphogenesis [65]. GO:0048846 is defined by the long-distance growth of a single cell process that is involved in the migration of an axon growth cone, where the migration is directed to a specific target site by a combination of attractive and repulsive cues [65]. Dendrites and axons are components of nerve cells. Dendrites are branched extensions of a nerve cell that transfer electrical impulses from other cells at synapses. Axons are a thread-like part of a nerve cell that conduct action potentials away from the nerve cell body. The protein sequences associated with both of these GO IDs were classified as cadherins via Swiss-

Prot [66]. Cadherins are a superfamily of adhesion molecules that facilitate cell to cell adhesion in both vertebrates and invertebrates [67]. Cadherins are named for their calcium dependent adhesion and play a major role in maintaining cell and tissue structure through the formation of adherens junctions [68]. At these junctions, cadherins bind to an actin cytoskeleton with catenin binding partners, forming a system that is important in morphogenesis and functions of vertebrate and invertebrate nervous systems [69]. Within echinoderms, this biological pathway is directly linked to the biological process of regeneration. Previous studies have provided evidence that a driving mechanism responsible for regeneration in four of the five classes of echinoderms are nerve-dependent, including crinoids [1]. In echinoderms, regeneration is widely used to reconstruct external parts such as arms, spines, and pedicellariae, and internal organs. This type of repair-based regeneration is often observed in crinoids which have long and fragile arms that are commonly subject to predation. Regeneration due to predation is so prevalent among crinoids that most specimens collected for analysis often exhibit two or more arms at various stages of regrowth [1]. This provides a possible explanation as to why these two nervous system related GO IDs have been enriched within crinoids and absent in other classes. Ancestor charts generated by QuickGO illustrate GO IDs and their relationships in Figure 3.2 and Figure 3.3.

The single GO ID enriched in ophiuroids was GO:0034472 and named as snRNA 3'-end processing. This term is defined as any process involved in forming the mature 3' end of an snRNA molecule [65]. Small nuclear RNA (snRNA) is a subtype of small RNA that are never translated and remain in the nucleus where they form part of the spliceosome [70]. The protein sequences associated with this GO ID were annotated as

Cyclin-dependent kinase 8 (CDK8) via Swiss-Prot. The relationships of this GO ID can be seen in Figure 3.4 where the GO term snRNA 3'-end processing is in a parental relationship with the five boxes below it. This includes U1 snRNAs, U2 snRNA, U4 snRNA, U5 snRNA, U6 snRNA, which are the five Uridine rich snRNAs that form the major spliceosome [70].

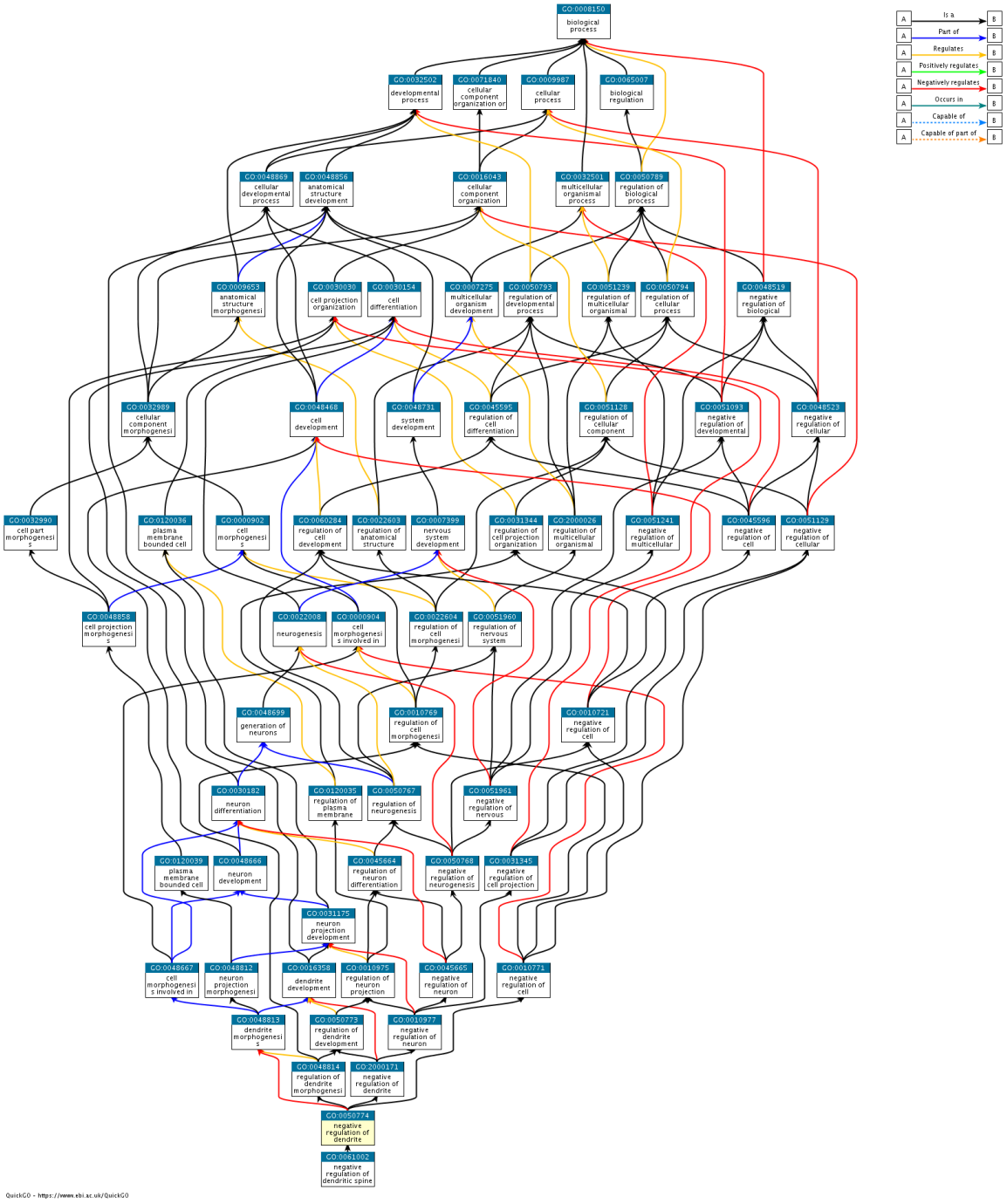

Figure 3.2: Above is an ancestor chart for the crinoid enriched GO:0050774 generated by QuickGO [71]. At the very top of the figure, the overarching GO description is biological process. The second to last box from the bottom highlighted in yellow describes this GO ID as the negative regulation of dendrite, the bottom box is a child of this GO ID, while the preceding boxes form a parental relationship.

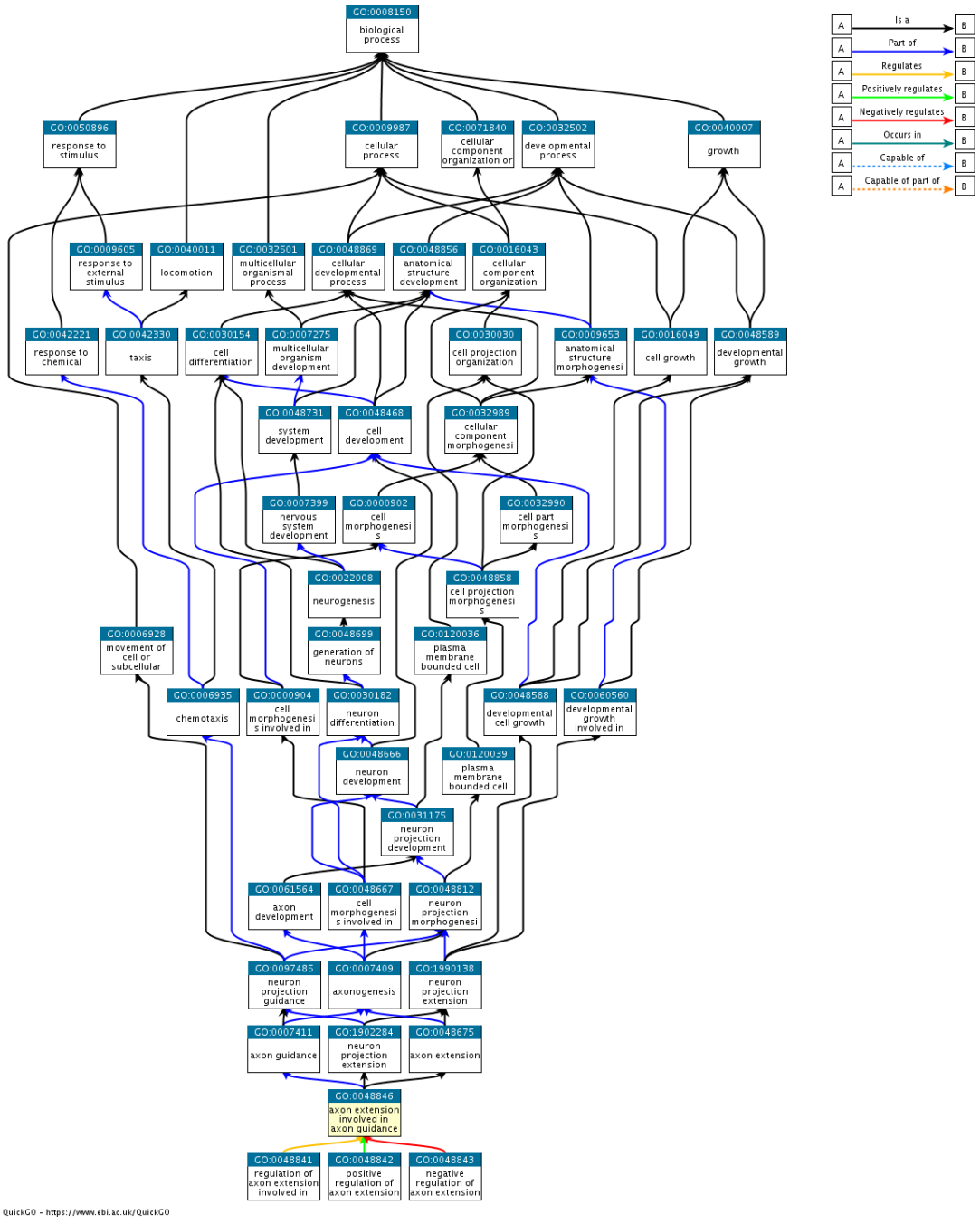

Figure 3.3: Above is an ancestor chart for the crinoid enriched GO:0048846 generated by QuickGO [71]. At the very top of the figure, the overarching GO description is biological process. The second to last box from the bottom highlighted in yellow describes this GO ID as axon extension involved in axon guidance, the bottom boxes are children of this GO ID, while the preceding boxes form a parental relationship.

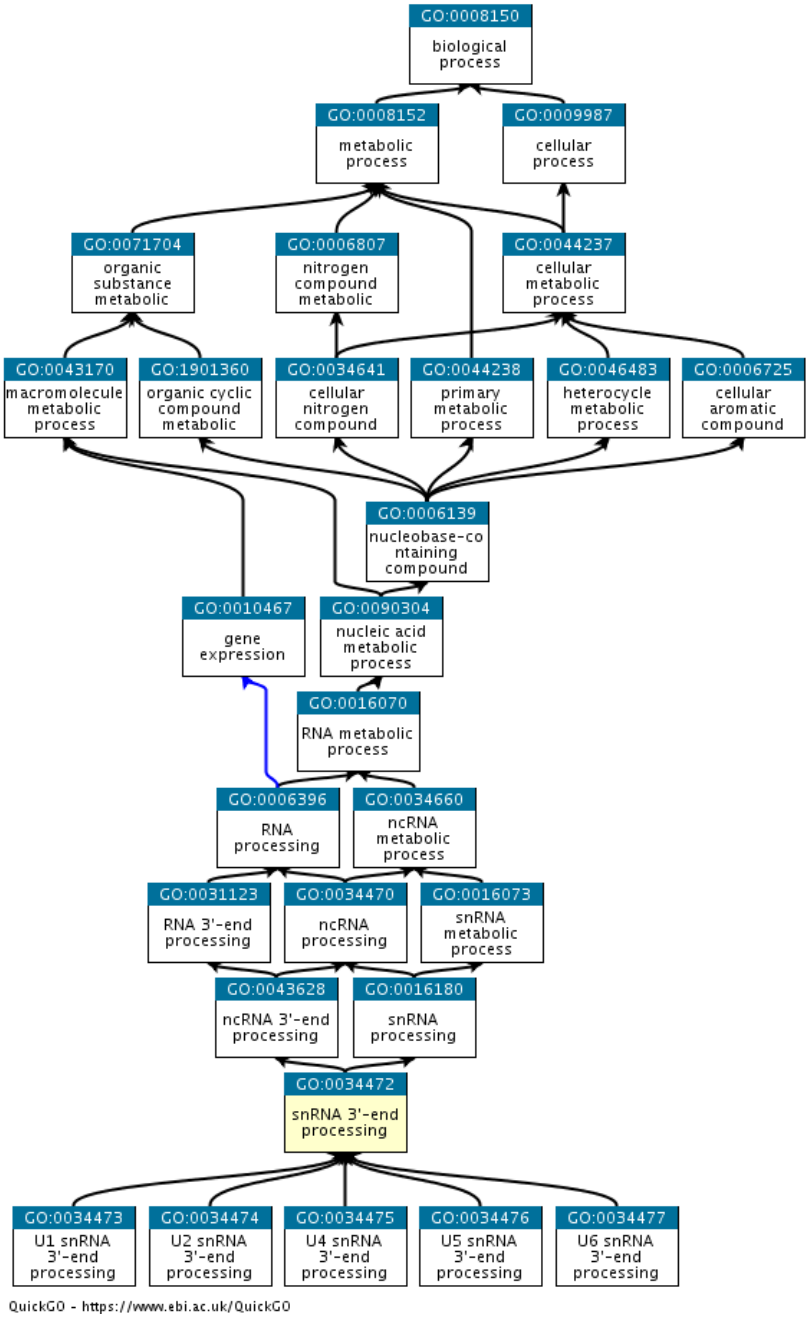

Figure 3.4: Above is an ancestor chart for the ophiuroid enriched GO:0034472 generated by QuickGO [71]. At the very top of the figure, the overarching GO description is biological process. The second to last box from the bottom highlighted in yellow describes this GO ID as snRNA 3'-end processing, the bottom boxes are children of this GO ID, while the preceding boxes form a parental relationship.

3.4 Discussion

Among the five extant classes, results indicate that asteroids have the most unique orthoclusters followed by crinoids, holothuroids, echinoids and ophiuroids. This is an expected result as the numbers of clusters detected are relative to the number of sampled individuals per echinoderm class. However, in hypergeometric test for GO enrichment performed on OrthoVenn, GO terms were not enriched in asteroids, the echinoderm class with the most orthocluster representatives. This was also the case for holothuroids and echinoids. Surprisingly, representatives from only two classes (crinoids and ophiuroids) were detected by the hypergeometric test of GO enrichment on OrthoVenn with default settings. A possible explanation of the absence of enriched GO terms for asteroids, echinoids and holothuroids is simply the lack of GO annotations contributing to a certain predetermined biological process, molecular function or cellular component that could be considered statistically significant at a p-value of <0.05 in this given dataset. This may be symptom of a wider diversity of GO terms spread across more genes in classes that have a larger quantity of orthoclusters rather than an accumulation of GO terms implicated in one of the three main categories of Gene Ontology on less orthoclusters. It is also important to note that although GO terms strive to be species-neutral, many of the annotations are derived from a few model organisms, none of which directly include any echinoderms.

Despite this, based on this analysis of this set of gene expression data the results show the heterogenous regenerative capabilities of echinoderms. Of the five classes within this dataset, only crinoids had enriched GO terms involved regeneration, though it

is known that regeneration is present amongst all classes [1]. Crinoids are well known for their extensive regeneration potential often replacing arms lost to predation or autotomy [1]. The species recovered in orthoclusters represented by the enriched GO terms include *Oligometra serripinna*, *Isometra vivipara* and *Ptilometra australis*. This set includes 33% of the all crinoids included in this study. In the case of *Oligometra serripinna*, research indicates that its regenerative capabilities has been previously studied [72]. The results of the enrichment analysis indicate that putative cadherin orthologs from these three species of crinoids are over-represented and implicated in pathways regulating nerve cell development, a biological process found to be a mechanism important in crinoid regeneration [1]. This alludes to the possible use of these three crinoid species as candidates for the expansion of transcriptomic nervous system regeneration studies that have already been performed on the holothuroid *Holothuria glaberrima* [3].

The only other class with an over-represented GO ID was within ophiuroids where *Ophiothrix spiculata* paralogs were detected in two orthoclusters. *Ophiothrix spiculata* represents 25% of the total ophiuroids included in this dataset and provides protein sequences that are members of the Cyclin-dependent kinase (CDK) family which are important regulators in the progression of the cell cycle. CDK8 was described to have an activating or inhibitory effect on transcription factor functions via binding or phosphorylation [73]. Preliminary studies regarding the complete set of kinases within the echinoderm genome (kinome) has provided evidence for the echinoderm kinome as being closer in total number kinases to the *Drosophila* kinome than the human kinome. Despite this variation in total number of kinases, the diversity of the echinoderm kinome

is more similar to that of human kinome, lacking only 2.1% of the total human subfamilies in comparison to *Drosophila*'s 12.9% [74]. Interestingly in humans, CDK8 was found to be a colorectal oncogene that regulates beta-catenin activity [75]. However, due to both the inhibitory and activating effect of CDK8 as mentioned above, the protein encoded by this gene may also act as a tumor suppressor [76]. Although these are preliminary results, the enrichment of the GO ID associated with this gene within ophiuroids suggests that they may be instrumental in basic research in defining the mechanisms that govern the oncogenesis of colorectal cancer in humans.

CHAPTER 4: EVOLUTION OF BIOMINERALIZATION IN ECHINODERMS

4.1 Introduction

Biomineralization is the biologically controlled formation of mineralized structures that function as support, protection or feeding. Biomineralization occurs across several metazoan lineages. The seemingly concerted emergence of biomineralization among metazoan lineages during the Cambrian explosion is a poorly understood evolutionary event that has progressed into the diverse biomineralized structures observed in modern metazoans, including echinoderms [77,78].

The central question in the evolution of biomineralization concerns the degree at which a common biomineralization toolkit of ancestral biochemical pathways was used that was then eventually independently co-opted for biomineralization across diverse taxa [15]. Echinoderm structure is supported by a rigid endoskeleton comprised of calcite and an organic matrix with the main skeletal structures including the test, spines, pedicellariae, tube feet and teeth. Due to the abundance of echinoderm skeletal elements in the fossil record, their skeletons have been of a major area of focus for paleontology [79].

Modern echinoderms have been important organisms for understanding the mechanisms of the regenerative process of biomineralization [80]. For example, the formation of the sea urchin embryonic endoskeleton and the cells (primary mesenchyme cells, or PMCs) that produce it are extensively studied at the gene expression level. The first PMC gene to be identified was the Mesenchyme-Specific-Protein, 130 KD (*msp130*) [11]. Including *msp130*, the entire protein family consists of eight members, (Msp130,

Msp130-related-1, Msp130-related-2, Msp-130-related-3, Msp130-related-4, Msp-130-related-5, Msp130-related-6, Msp130-related-7) Msp130-related-1 and Msp130-related-2 were later discovered in 2002 [14], followed by Msp130-related-3, Msp130-related-4, Msp-130-related-5, Msp130-related-6 in 2006 [11] and the most modern, Msp130-related-7 was described in 2014 [15]. The transcript levels of all the *msp130* genes are at their highest levels during embryogenesis while Msp130 and Msp130 rel1-3 are the most abundant in biomineralized tissues of the adult sea urchin [15].

Other workers have performed phylogenetic analyses of *msp130* across eukaryotes and prokaryotes but only studied only one member of the Msp130 protein family [15]. This worker argues for the hypothesis that there were several independent horizontal gene transfer events of ancestral *msp130* gene into early metazoans via symbiotic relationships with microbial communities [15,81].

In this study, I used methods modified from chapter 2 to survey 40 transcriptomes of extant echinoderms for proteins involved in biomineralization. Moreover, the transcriptome dataset at hand significantly expands the collection of known biomineralization-related proteins both in terms of the Echinoderm taxa sampled and in terms of the members of the gene family studied. With these data, I investigate the competing hypotheses of horizontal gene transfer versus radiation of the gene family from a common ancestral *msp130* gene family in echinoderms and metazoans (e.g. molluscs). Previously, studies of the evolution of the Msp130 protein family in echinoderms has only considered the three species; *Heliocardis erythrogamma, Heliocardis tuberculata* and *Strongylocentrotus purpuratus* [15].

4.2 Research Design and Methods

To identify biomineralization genes within the 40 transcriptomes, the eight protein sequences from the *msp130* gene family (Msp130 and Msp130-related-1, Msp130-related-2, Msp130-related-3. Msp130-related-4. Msp130-related-5. Msp130-related-6, Msp130-related-7) [11] were used in a BLASTP sequence homology search. The Msp130 family query sequences were sourced from www.echinobase.org [82,83]. Command line ncbi-blast-2.2.30+ was used to create a database of 1,198,706 protein sequences [84]. BLASTP was used with an expectation value cutoff of 1e-5. The resulting hits were first filtered for duplicate sequences which were removed. The hits were then filtered for a sequence length minimum of 200 residues.

Sequences that passed the length filter were then matched with outgroup sequences from NCBI. Two matrices were created for *msp130*, one was matched with only the cephalochordate *Branchiostoma floridae* and another was matched with using bacteria, algae, molluscs, cephalochordates and hemichordates [15]. Data from non-echinoderm taxa such as algae, bacteria, molluscs, cephalochordates and hemichordates sequences were collected from NCBI BLASTP [15]. The filtered sequence hits from non-echinoderm taxa were then aligned with Msp130 echinoderm sequences using MAFFT. The remaining Msp130 family members were matched with the corresponding Msp130 proteins from the cephalochordate *Branchiostoma floridae*. The resulting datasets were then aligned using MAFFT producing a total of nine alignments for the Msp130 family.

Following multiple sequence alignment, the BOXER program (as initially described in chapter 2 but this time without paralogy control) was used to select optimal alignments based on two criteria 1) number of unique taxa of n-1 and 2) 90% of gap characters allowed in each alignment. The resulting multiple sequence alignments are then used as input for RAxML tree search analyses under maximum-likelihood criterion using a CAT model of rate heterogeneity, chosen for its computational efficiency [85]. RAxML then produced a set of gene trees reflecting the radiation of each members of the *msp130* gene family. The results produced by these methods are presented in Figures 4.1- 4.2, 4.3 - 4.9 for the gene tree topologies of *msp130* with different outgroups, *msp130rel1* through *msp130rel7*, respectively.

4.3 Results

To search 40 transcriptomes for Msp130 sequences, query sequences first needed to be identified. To identify protein sequences from the Msp130 family within echinoderms, initial searches began on NCBI protein database resulting in only three members including Msp130, Msp130rel1 and Msp130rel2 from echinoid echinoderms. To gather the complete set of query sequences, www.echinobase.org was queried for Msp130 sequences. Query sequences were sourced from Echinobase as it contains the most recent assembly of the purple sea urchin genome [82]. These query sequences were derived from a combination of protein-coding RNAs of *Strongylocentrotus purpuratus* in different life stages including 10 different embryonic stages, six feeding larval and metamorphosed juvenile stages, and six adult tissues [86]. This search yielded the eight

sequences previously described members of the Msp130 protein family. The creation of the blast database of 40 echinoderm transcriptomes was performed using the makeblastdb command from ncbi-blast-2.2.30+. Input used for the database were 1,198,706 protein sequences produced from Transdecoder [52] in Chapter 2. The custom blast database was then searched using BLASTP for each query sequence representing the Msp130 family with a e-value cutoff of 1-e5. This process produced 1202 blast hits across the Msp130 family. At an e-value cutoff of 1-e5, Msp130 had the most hits while Msp130rel3 had the least. After the removal of duplicates, Msp130rel6 contained the most unique sequences with 116 while Msp130rel3 had the least at 43. The remaining unique sequences were then filtered for a minimum length of 200 residues where all but two (Msp130rel3 and Msp130rel7) of the Msp130 family proteins had sequence hits containing at least 200 residues or greater. These results are summarized in the following table (See Table 4.1).

Table 4.1: This table describes the protein sequence query used, its description, source, and number of hits statistics.

Protein Sequence Query used	**Msp130 family member**	**Source**	**Number of hits**	**After duplicates removed**	**Hits with length ≥ 200**
SPU_002088.3a	Msp130	Echinobase	213	108	50
SPU_013822.3a	Msp130rel1	Echinobase	149	101	50
SPU_016506.3a	Msp130rel2	Echinobase	150	104	50
SPU_013823.3a	Msp130rel3	Echinobase	81	43	26
SPU_014496.3a	Msp130rel4	Echinobase	166	115	50
SPU_015763.3a	Msp130rel5	Echinobase	162	111	50
SPU_015326.1	Msp130rel6	Echinobase	164	116	66
SPU_021242.3a	Msp130rel7	Echinobase	117	75	40

Table 4.2: This table indicates the BOXER settings used the creation of the *msp130* gene family trees.

Msp130 family member	Number of aligned taxa	Number of unique taxa	BOXER unique taxa cutoff	BOXER aligned taxa	BOXER Gap % permitted in alignment
Msp130 bacteria outgroup	93	42	41	84	90
Msp130	54	19	18	43	90
Msp130rel1	54	18	17	39	90
Msp130rel2	54	18	17	39	90
Msp130rel3	30	13	12	21	90
Msp130rel4	54	18	17	41	90
Msp130rel5	54	18	17	43	90
Msp130rel6	70	19	18	51	90
Msp130rel7	43	16	15	31	90

The *msp130* gene tree (Figure 4.1) includes representatives from each extant echinoderm class and is rooted on the cephalochordate *Branchiostoma floridae*. Within the echinoderm ingroup, there are four main nodes with bootstrap values that are ≥ 92%. that support the reconstruction of the evolution of *msp130*. Among these four nodes, one subtends a lineage exclusively of crinoids. This tree also shows a clade of twelve asteroids with a bootstrap support value of 98%. A lineage of echinoids is also observed consisting of *Arbacia*, *Eucidaris* and the *Strongylocentrotus purpuratus* query sequence (SPU_002088) with a bootstrap support value of 97%. The lineage of echinoids is sister to the holothurian *Psolus* with a 92% bootstrap support value. In summary, the radiation of the Msp130 gene in these lineages is result of a speciation event that occurred at the Eleutherozoa-Crinozoa split. Subsequently, as seen within echinoderm classes, we also observe several well-supported gene duplication events (all ≥ 98% bootstrap support

values). A few examples of this phenomenon can be observed in the echinoids *Arbacia*. *Eucidaris*, *Strongylocentrotus*, the asteroids *Pisaster*, *Henricia*, and *Porania*, and the crinoids *Oligometra*, *Ptilometra*, and *Isometra*.

In the additional topology rooted on the bacteria generated for the *msp130* gene (Figure 4.2) several major lineages are observed. One lineage includes Eukaryota (Stramenopiles, Viridiplantae, Metazoa). The split between metazoa and other eukaryotes (green and brown algae) is supported with a bootstrap support value of 76%. Within the metazoan group, these genes form two lineages in Mollusca with strong bootstrap support values (100% and 95%). These *msp130* gene in hemichordates and cephalochordates form a clade with moderate bootstrap support (59%) that is sister to the echinoderm lineages. Within this tree there is strong bootstrap support for the echinoderm lineages of *msp130* (97%). Thus, the radiation of *msp130* are likely the result of ancient speciation events that split the ancestors of higher taxa (Stramenopiles, Viridiplantae, Metazoa) and not multiple independent horizontal transfer events.

Subsequently, within each clade of metazoans (hemichordate, cephalochordates, molluscs, echinoderms), both gene duplication and speciation events are observed. An example of speciation followed by gene duplication occurs within the hemichordates and cephalochordates. A more complex pattern of evolution occurs within the molluscs in which gene duplication precedes speciation, followed by further gene duplication. In echinoderms we observe a speciation event that leads to the formation of single lineage of *msp130* in asteroids that is separate from other echinoderms. In other echinoderm classes, the evolution of *msp130* is more complex and consists of both speciation and duplication events. The relationships within the echinoderm clade remain like that of

Figure 4.1, with only minor fluctuations in bootstrap support values to the relationships. I observe a strongly supported asteroid lineage with 100% bootstrap support value. This topology presents another lineage of nine crinoids with 89% bootstrap support values. A well-supported (bootstrap value of 96%) echinoid lineage is observed consisting of *Arbacia, Eucidaris,* the *Strongylocentrotus purpuratus* query sequence, and two additional echinoid sequences that were previously determined to be *msp130* orthologs (*Heliocardis erythrogamma* and *Heliocardis tuberculata)* were added to the echinoid lineage[87]. This echinoid lineage is sister to a holothurian, *Psolus* with a 94% bootstrap support value.

Figure 4.3 shows the phylogenetic relationships of the *msp130rel1* gene within extant echinoderms. In this topology, the tree is rooted on the cephalochordate *Branchiostoma floridae*. Within the echinoderm clade, the evolution of the *msp130rel1* gene can be best described by four nodes that have strong bootstrap support values. I observe a strongly supported asteroid lineage with 92% bootstrap support value. This topology presents another lineage of nine crinoids with 89% bootstrap support values. A well-supported (bootstrap value of 91%) echinoid lineage that consists of *Arbacia* and *Eucidaris*. This echinoid lineage is sister to a holothurian, *Psolus* with a 91% bootstrap support value. The relationships within the topology presented in this analysis remains consistent with observations made on the *msp130* gene tree. Once again, I observe unambiguously supported gene duplication and speciation events within classes with bootstrap support values no less than 100%. This is observed in four of the five classes. Examples of gene duplication includes the crinoid *Oligometra,* the echinoid *Eucidaris*

and the asteroid *Henricia*. Examples of speciation events can be seen within the ophiuroids between *Ophioderma* and *Astrophyton*.

In Figure 4.4 the relationships of the *msp130rel2* gene within extant echinoderms is presented and is rooted on cephalochordate *Branchiostoma floridae*. The ingroup forms several echinoderm lineages, there are three notable lineages observed that have well supported nodes (≥ 74). Within one lineage, the phylogenetic analysis recovers strong support for a crinoid group (bootstrap support value of 94%), a strongly supported asteroid group (bootstrap value of 97%) and well supported echinoid groups (bootstrap value of 74%). The relationships observed in this topology are like that of *msp130* and *msp130rel1*. Like the previously described gene trees, speciation events among taxa and gene duplications events are also observed within classes.

Figure 4.5 displays the topology for the *msp130rel3* gene within extant echinoderms, rooted on cephalochordate data. Within the ingroup three strongly supported lineages emerge: a lineage in crinoids with a 90% bootstrap support, a lineage in ophiuroids with a 100% bootstrap support value and a lineage in asteroids with a 99% bootstrap support value. The tree presented here contains fewer terminals but still presents gene duplication and speciation events at the class level.

The relationships of extant echinoderms rooted on a cephalochordate for the gene *msp130rel4* is presented in Figure 4.6. Akin to *msp130*, *msp130rel1*, and *msp130 rel4*, there are four strongly supported lineages in this tree. A well-supported lineage of *msp130rel4* in asteroids can be observed with a bootstrap support value of 97%. I also observe a well-supported lineage of *msp130rel4* genes in seven echinoids sister to a holothurian with bootstrap support values of 88% and 87% at each respective node.

Lastly a grouping of *msp130rel4* genes in nine crinoids presents itself as a well-supported lineage with a 91% bootstrap support value. Gene duplication and speciation events are again evident in this topology as with all the previous topologies of the *msp130* gene family.

The relationships of extant echinoderms rooted on a cephalochordate for the gene *msp130rel5* is presented in Figure 4.7. The relationships within the in-group here remain analogous the rest of the *msp130* family. The same well supported four nodes can be used to reconstruct lineages within the tree. Asteroid *msp130rel5* genes form a lineage with 97% bootstrap support value, echinoid *msp130rel5* genes form a lineage with 95% bootstrap support that is sister to a holothurian *msp130rel5* with 77% bootstrap support, *msp130rel5* in crinoids form a nine-terminal lineage with moderate bootstrap support of 65%.

Figure 4.8 shows the topology for the *msp130rel6* gene within extant echinoderm and is rooted on a cephalochordate. This figure contains the most aligned taxa (See Table 4.2) of the *msp130* genes analyzed in this study and presents the same major lineages described in the aforementioned analyses. In summary the four nodes that standout as well supported in this tree include *msp130rel6* genes in asteroid lineage with 75% bootstrap support, *msp130rel6* genes in the echinoid lineage with 90% bootstrap support sister to a holothurian with 67% bootstrap support and a crinoid lineage with 88% bootstrap support.

The evolution of most modern member of the *msp130* gene family within echinoderms is presented in Figure 4.9 [15]. Alignment used to produce this phylogenetic analysis contains the second to least number of terminals with 31 aligned

taxa. Despite the smaller dataset, the same four lineages appear as previously described. The asteroid lineage is well supported with 94% bootstrap support, an echinoid lineage is present with 100% bootstrap support that is sister to a holothurian with 72% bootstrap support and a crinoid lineage of six taxa is supported with a 92% bootstrap value.

In all nine of the topologies presented in these results, recurring themes emerge within the echinoderm lineage. Either four and three main lineages can be used to explain the major relationships within each tree. These lineages are represented by four classes: asteroids, echinoids, holothurians, and crinoids.

Figure 4.1: The evolution of the *msp130* gene within extant echinoderms, rooted on the cephalochordate *Branchiostoma floridae*. The red circled nodes are described within the text and describe a series of speciation and duplication events, these events are detailed in Figure 1.2 for reference. Final ML Optimization Likelihood: -17860.838230.

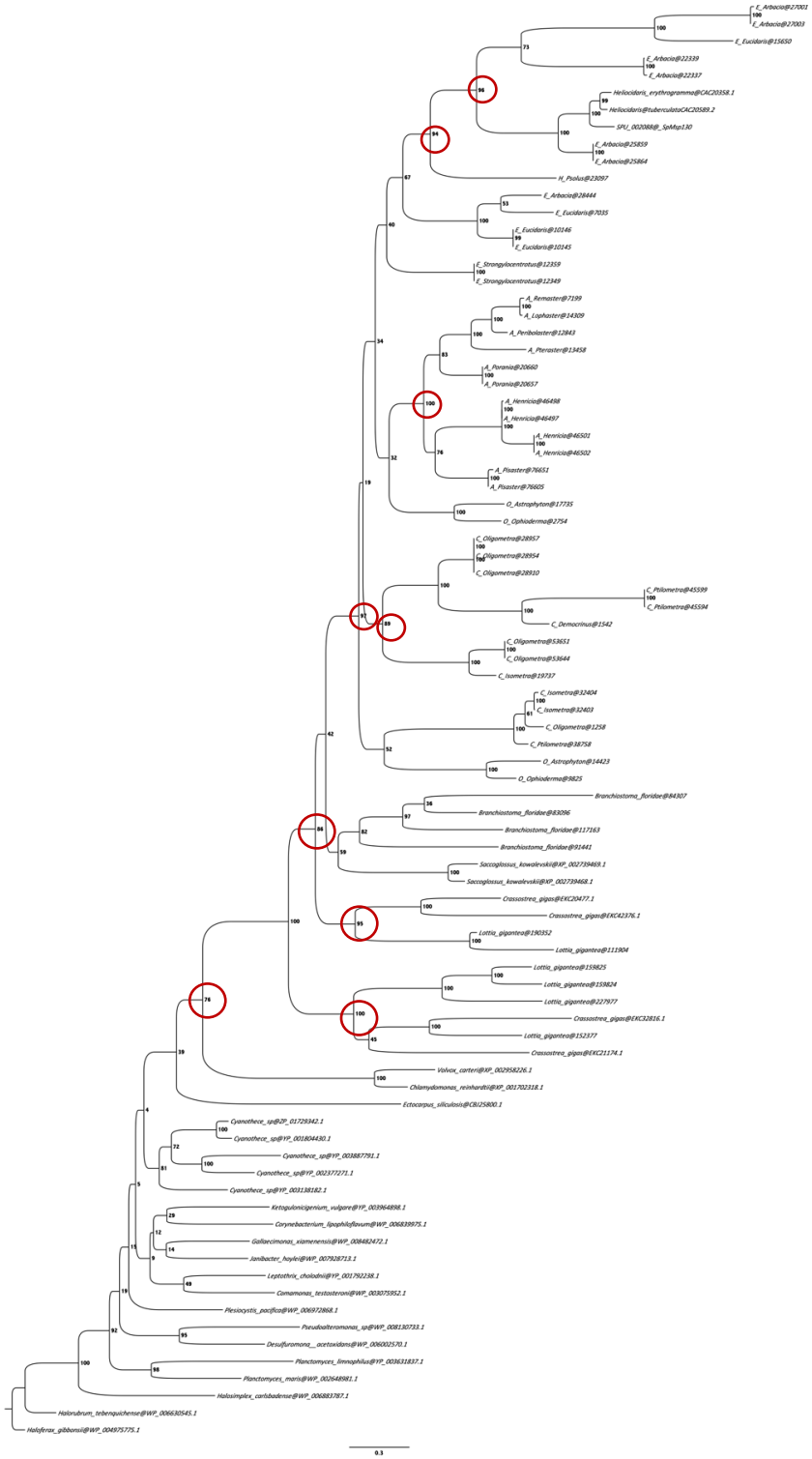

Figure 4.2: The *msp130* gene maximum likelihood tree of extant echinoderm species and their paralogs rooted on bacteria. The red circled nodes are described within the text and describe a series of speciation and duplication events, these events are detailed in Figure 1.2 for reference. Final ML Optimization Likelihood: -40806.433845.

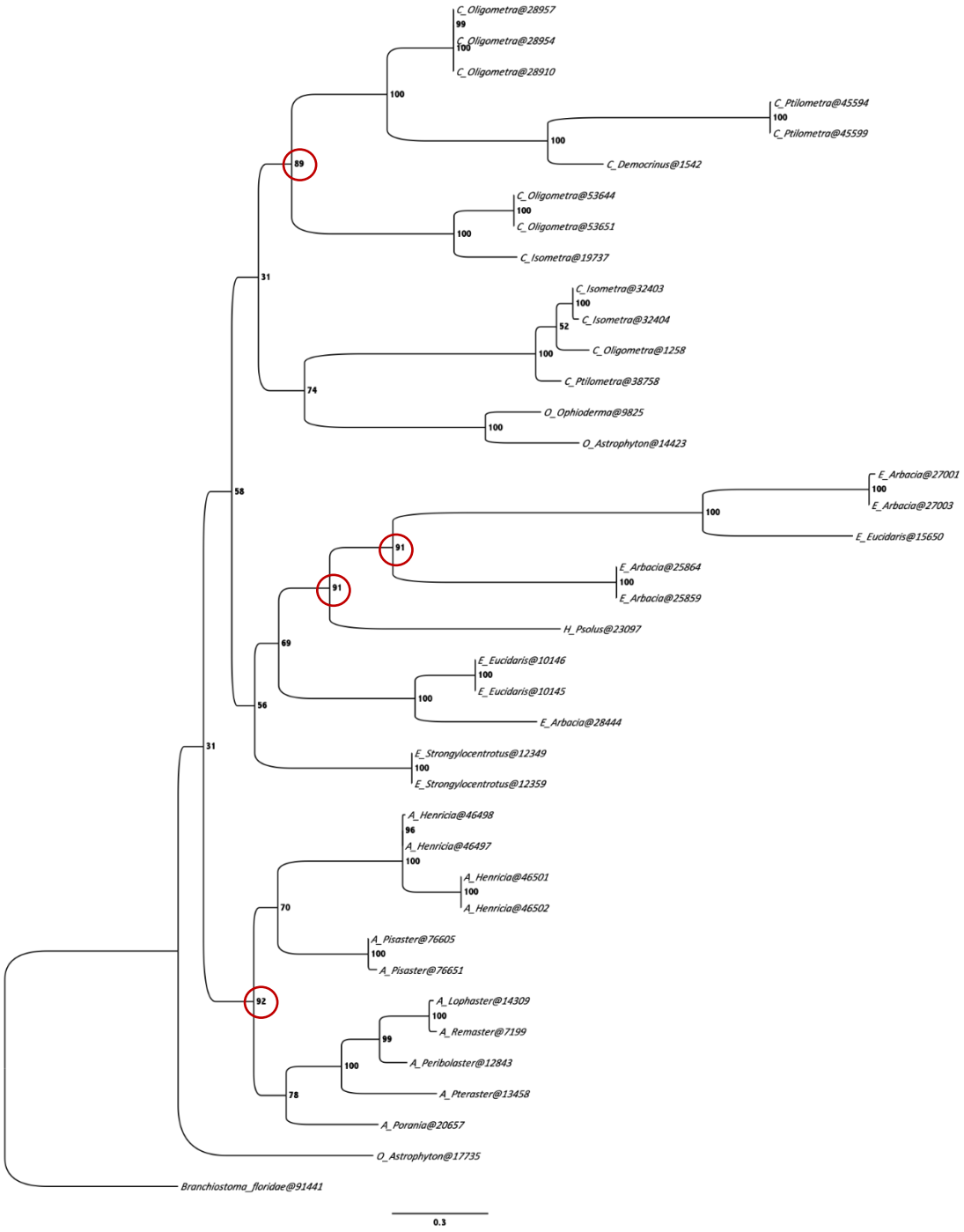

Figure 4.3: *msp130rel1* maximum likelihood tree of extant echinoderms rooted on the cephalochordate *Branchiostoma floridae*. The red circled nodes are described within the text and describe a series of speciation and duplication events, these events are detailed in Figure 1.2 for reference. Final ML Optimization Likelihood: -15807.759502.

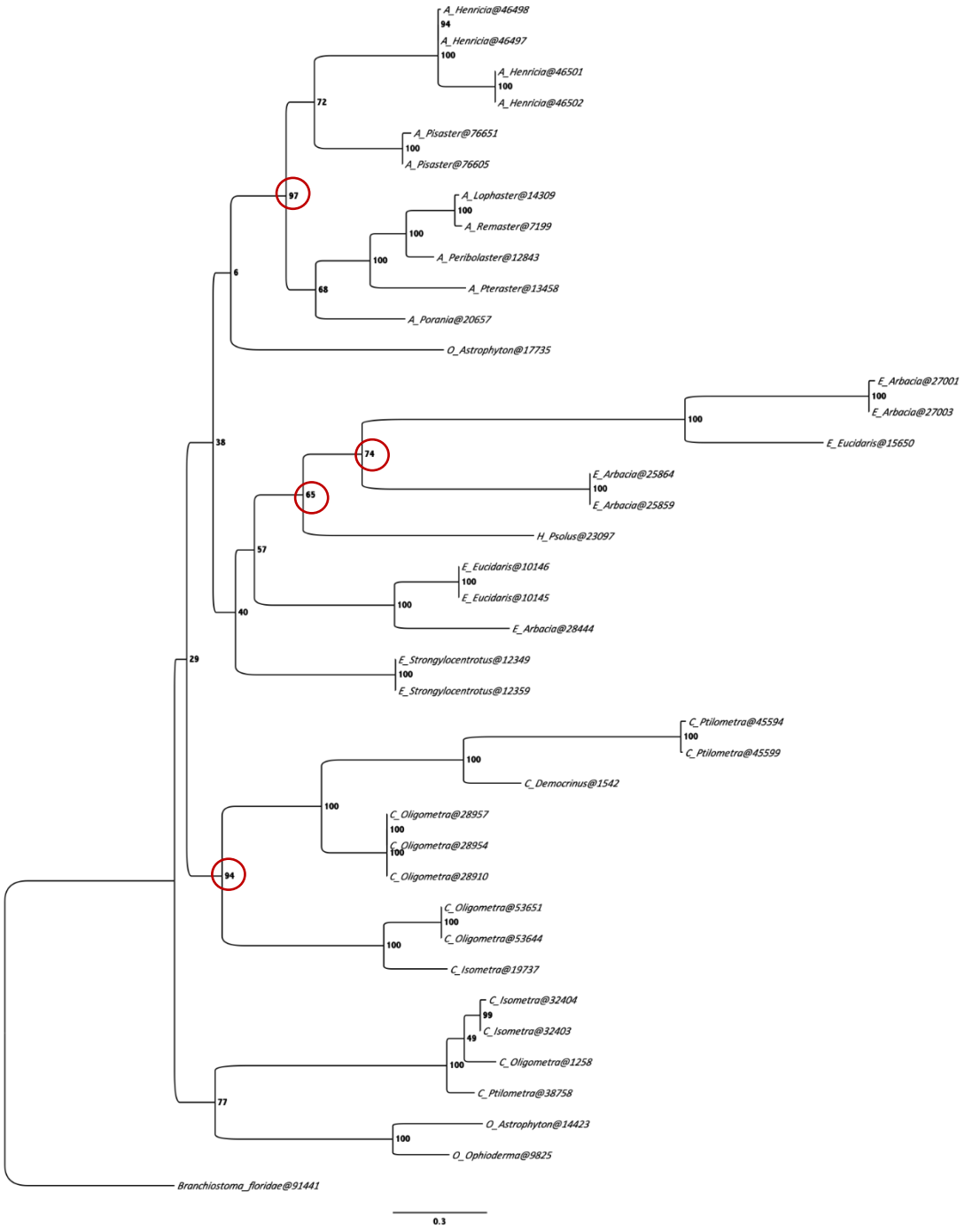

Figure 4.4: *msp130rel2* maximum likelihood tree of extant echinoderms rooted on the cephalochordate *Branchiostoma floridae*. The red circled nodes are described within the text and describe a series of speciation and duplication events, these events are detailed in Figure 1.2 for reference. Final ML Optimization Likelihood: -15986.297024.

Figure 4.5: *msp130rel3* maximum likelihood tree of extant echinoderms rooted on the cephalochordate *Branchiostoma floridae*. The red circled nodes are described within the text and describe a series of speciation and duplication events, these events are detailed in Figure 1.2 for reference. Final ML Optimization Likelihood: -13675.964464.

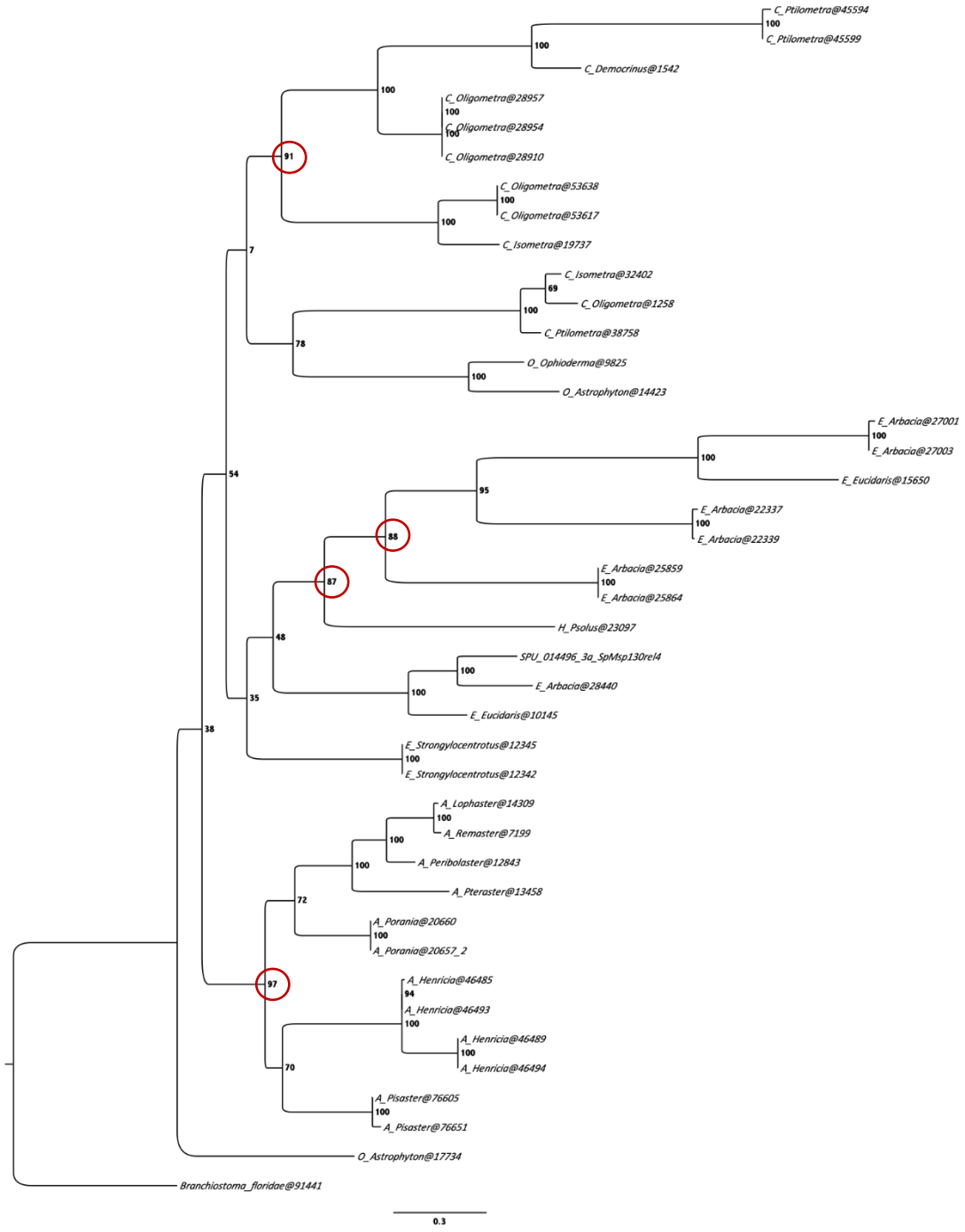

Figure 4.6: *msp130rel4* maximum likelihood tree of extant echinoderms rooted on the cephalochordate *Branchiostoma floridae*. The red circled nodes are described within the text and describe a series of speciation and duplication events, these events are detailed in Figure 1.2 for reference. Final ML Optimization Likelihood: -18477.452657.

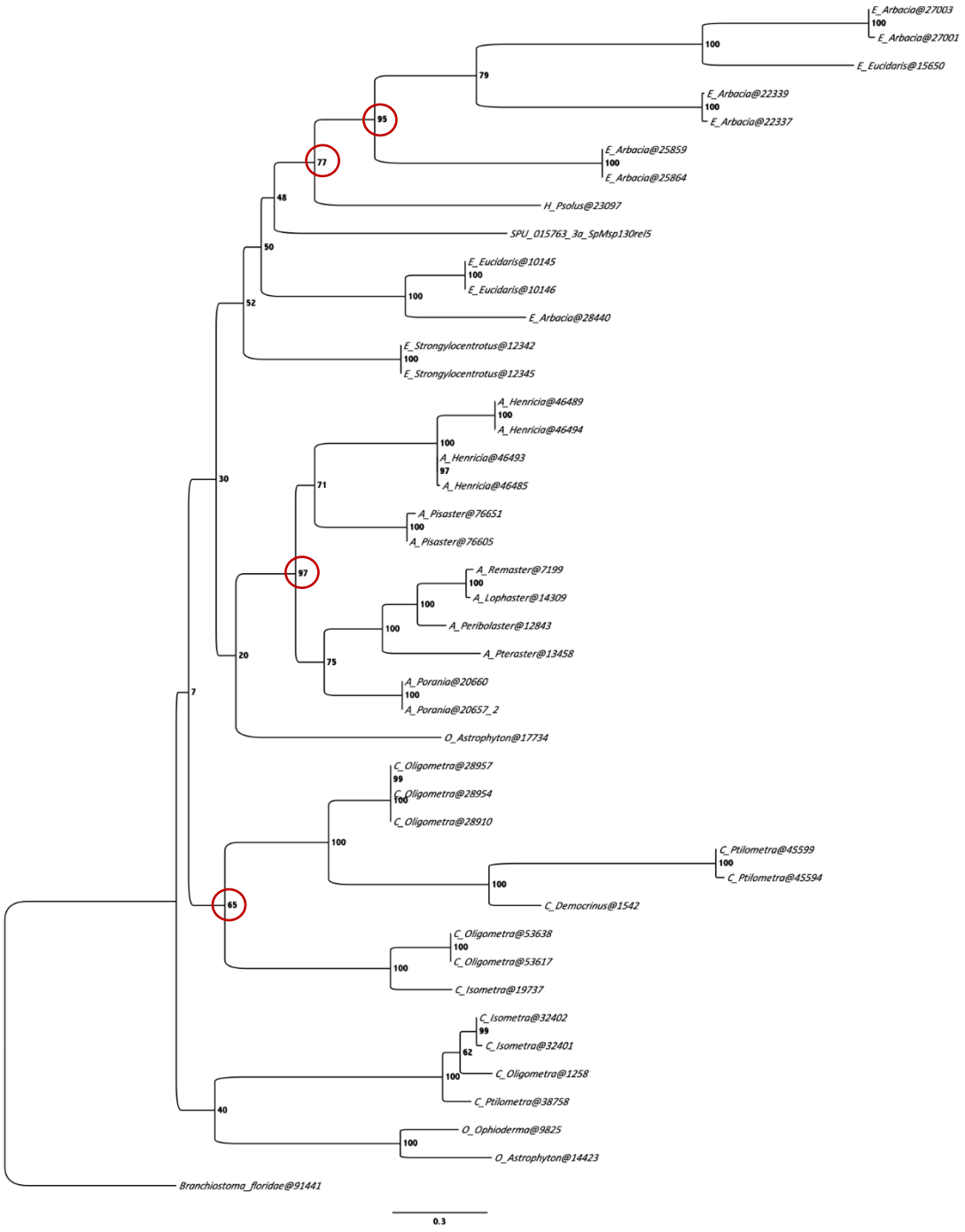

Figure 4.7: *msp130rel5* maximum likelihood tree of extant echinoderms rooted on the cephalochordate *Branchiostoma floridae*. The red circled nodes are described within the text and describe a series of speciation and duplication events, these events are detailed in Figure 1.2 for reference. Final ML Optimization Likelihood: -18161.884908.

Figure 4.8: *msp130rel6* maximum likelihood tree of extant echinoderms rooted on the cephalochordate *Branchiostoma floridae*. The red circled nodes are described within the text and describe a series of speciation and duplication events, these events are detailed in Figure 1.2 for reference. Final ML Optimization Likelihood: -35180.540418.

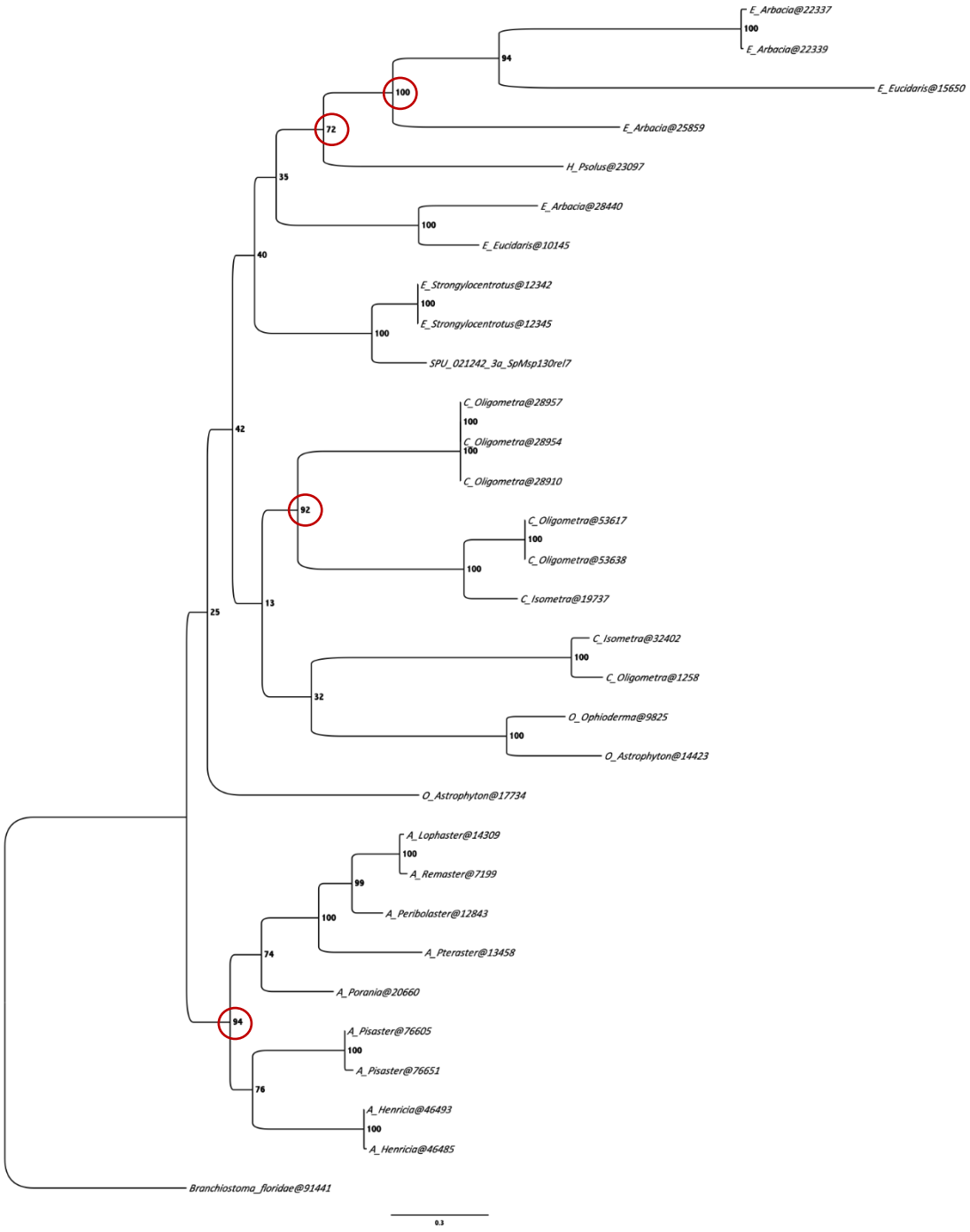

Figure 4.9: *msp130rel7* maximum likelihood tree of extant echinoderms rooted on the cephalochordate *Branchiostoma floridae*. The red circled nodes are described within the text and describe a series of speciation and duplication events, these events are detailed in Figure 1.2 for reference. Final ML Optimization Likelihood: -15619.011501.

4.4 Discussion

The phylogenetic relationships of the *msp130* gene family within echinoderms was previously limited to three species; *Heliocardis erythrogamma, Heliocardis tuberculata* and *Strongylocentrotus purpuratus* [15]. The results of this work have expanded the taxonomic coverage of *msp130* gene family by 17 new echinoderms species. This list includes eight asteroids (*Pisaster, Pteraster, Remaster, Peribolaster, Lophaster, Henricia, Porania, Labidiaster*), four crinoids (*Isometra, Ptilometra, Oligometra, Democrinus*), two echinoids (*Arbacia, Eucidaris*), two ophiuroids (*Ophioderma, Astrophyton*) and one holothuroid (*Psolus*). All previously mentioned taxa were present in the analysis of MSP130rel6. MSP130rel1, MSP130rel2, MSP130rel4, and MSP130rel5 contained one less asteroid, *Labidiaster*. MSP130rel3 and MSP130rel7 contained less taxa overall, generally, less asteroid representatives. These results reflect that the *msp130* gene family is more prevalent in echinoderms than previously thought.

I expected to recover BLASTP hits predominantly consisting of Msp130, Msp130rel1-3 proteins as they are more likely to be expressed in the adult echinoderms from which this sequence data was produced [15]. However, after the removal of duplicate sequences, the proteins Msp130rel4-6 had the most hits (See Table 4.1). This result suggests that within this dataset for 40 echinoderm transcriptomes, Msp130rel4-6 are more likely to be expressed in adult tissues than Msp130, and Msp130rel1-3.

The multiple phylogenetic analyses displayed a consistent pattern: within the echinoderm ingroup, there are four main lineages that explain the evolution of *msp130* gene family with high bootstrap support. Among these four lineages, one subtends a

lineage exclusively of crinoids, another shows a clade of asteroids, a third lineage of echinoids is also observed which is sister to the holothuroid *Psolus*. The diversification of the *msp130* gene in these lineages are a likely result of a speciation event that occurred at the Eleutherozoa-Crinozoa split rather than the repeated horizontal transfer of *msp130* genes. In all nine topologies, both gene duplication and speciation events were present within echinoderms. Within the *msp130* gene family, two main models of evolution can be observed. One model is speciation followed by gene duplication and the alternative model is a duplication event that precedes speciation, followed by further duplication. In echinoderms we observe a speciation event that leads to the formation of single cluster of Msp130 in asteroids. In other echinoderm classes, the relationships are more complex and consisting of both speciation and duplication events.

A previous worker who studies the evolution of the *msp130* gene family proposed that the existence of the Msp130 in eukaryotes and prokaryotes are an effect of multiple cases of horizontal gene transfer [15]. With the additional echinoderm representatives from the current dataset, I recovered a pattern of evolutionary events in the echinoderm *msp130* gene family that would be inconsistent with horizontal gene transfer. The hypothesis of interdomain horizontal gene transfer between bacteria and animals is likely a reflect a random occurrence happening at a very low frequency in the population [88]. Most examples of HGT often occur within bacteria and very few examples with little evidence have supported this phenomenon occurring between the vastly different lineages of bacteria and animals. Furthermore, if HGT was rampant in *msp130,* increased mixing of taxa (e.g. algae/animal sister taxa) in gene trees would be observed. The topologies produced by the phylogenetic analyses in this study provide evidence for idea that the

radiation of *msp130* in prokaryotes and eukaryotes are better described by by descent through series of speciation and gene duplication events that occur naturally. The absence of *msp130* in other major metazoan lineages can be alternatively described by gene loss rather than multiple instances of horizontal gene transfer [89,90]. However, we believe that this notion is testable by increasing genomic and transcriptomic sampling of animal phyla. An example of this is the Msp130 protein sequence of the brachiopod *Lingula anatina*, which was discovered after this analysis (NCBI protein sequence accession XP_013418960) [91]. As more *msp130* representatives around found in different lineages, HGT becomes more improbable.

CHAPTER 5: CONCLUSIONS

In my analyses I used RNA-Seq generated transcriptomes to develop an improved understanding of three major aspects of echinoderm biology: echinoderm phylogeny, echinoderm transcriptome gene content, and the evolutionary relationships of echinoderms within the *msp130* family. The analyses that I performed addressed the following questions related to these three main topics: What phylogeny of extant echinoderms is supported by a large transcriptome dataset? Will the resulting phylogeny support the Cryptosyringid or Asterozoa-Echinozoa hypothesis? Can this derived phylogeny be used to study the variation of 40 transcriptomes assemblies among echinoderm clades? Will certain functions be enriched in certain taxa? I also asked a question involving the biomineralization related gene family of *msp130* and whether it was introduced into echinoderms via multiple instances of horizontal gene transfer or simple descent from a common ancestor with speciation and gene duplication events among animal lineages. The advent of next-generation sequencing data has made these studies tractable providing results that are novel and challenge previous studies.

In the study of echinoderm phylogenetics 40 novel transcriptomes were used to test the Asterozoa-Echinozoa versus Cryptosyringid hypotheses. Using novel methods that created 19 distinct data subsets based on alignment occupancy (BOXER) the results of the largest datasets of show support for the Asterozoa-Echinozoa hypothesis. The topologies of the five most inclusive datasets recovered Echinozoan and Asterozoan groups with high bootstrap support. This is consistent with previous works but at a greater scale of taxon sampling and diversity [33,43–45]. The topologies that were

created from the most gap permissive datasets containing the highest amounts of loci (i.e. >500 loci and 45-95% allowable gaps) placed *Xyloplax* is sister to asteroids, a relationship that has been supported by previous work (Baker et al., 1986; Gale, 1987; Mah, 2006) but refuted by more tightly scoped analyses of asteroids and ophiuroids in which *Xyloplax* is a velatid asteroid (Linchangco et al., 2017; Janies et al., 2011).

In the mapping of echinoderm of gene function via annotation, clades resulting from the preferred phylogeny in chapter 2 were used to search for class-specific genes within echinoderms. This analysis lead to the detection of GO enriched terms in crinoids, and ophiuroids relating to biological processes. This revealed that members of crinoids could serve as models in nervous system regeneration studies and that ophiuroids may be valuable to basic research in defining the mechanisms that govern the oncogenesis of colorectal cancer in humans.

In the study of the evolution of the *msp130* gene family within echinoderms, the multiple phylogenetic analyses of the RNA-Seq data displayed a recurring pattern of four major lineages representing asteroids, crinoids, echinoids, and holothuroids. These lineages are supported with high bootstrap values and can be explained by gene duplication and speciation events. Across the echinoderm classes, a complex series of speciation and duplication events are observed leading to these four lineages. This study also considered non-echinoderm taxa in the evolution of *msp130* via an additional analysis rooted on the bacteria. This work resulted in a lineage that includes Eukaryota (Stramenopiles, Viridiplantae, Metazoa). Within this metazoan group comprised of Mollusca, cephalochordates, hemichordates and echinoderms the absence of *msp130* in other metazoan lineages can be described by gene loss or lack of adequate genomic

sampling thus far in contrast to the rare possibility of multiple instances of horizontal gene transfer [15].

5.1 Significance

The studies performed in this work contributes a deeper understanding of the evolutionary biology of echinoderms in three different ways.

The phylogenetic reconstruction of echinoderms addressed the two debates of Asterozoa-Echinozoa versus Cryptosyringid and the placement of *Xyloplax* within the phylum. Using a novel sensitivity analysis, the most inclusive datasets of 1256 loci supported the Asterozoa Echinozoa hypothesis and the placement of *Xyloplax* as sister to all asteroids. This demonstrates that pipeline used to generate these findings can be effectively used on other datasets to resolve debated relationships at the class level.

The annotation by similarity of 40 transcriptomes uncovered potential alternative model organisms for regenerative nervous system studies and colorectal cancer research.

The phylogenetic reconstruction of the eight members of the *msp130* gene family provides a newly expanded taxonomic and transcriptomic view of gene family evolution. The presence of *msp130* gene family members in 17 new echinoderms uncovers that a biomineralization toolkit was present in metazoans prior to echinoderms and radiated with speciation and gene duplication events rather than via horizontal gene transfer.

5.2: Future work

Further study in transcriptomics can benefit from additional sampling of more echinoderms and other metazoan phyla. The addition of additional echinoderm transcriptomes from classes with limited sampling may affect the evolutionary relationships within Eleutherozoa and our view of the radiation of the *msp130* gene family. Additional samples would also provide better support for function of gene protein products of under-represented classes.

The phylogenetic methods in this work used ancestry to extract patterns of relationships between taxa as well as infer the function of protein products, providing new insights on echinoderm evolution that remain important in developmental studies. I created a pipeline that uses phylogenetic methods and a novel sensitivity analysis, taking in an input of raw RNA-Seq data with an output of a maximum likelihood phylogeny. The next steps involve the development of a publicly accessible, web-facing application and backend server that automates this process for any users with who would like to infer patterns of relationships between taxa or gene content of their sequence data.

I also performed sequence similarly based annotation of echinoderm transcriptomes revealing enriched gene families in crinoids and ophiuroids. These annotations information can be made public and added to EchinoDB to provide enhanced annotations for the existing echinoderm database [10].

References

[1] Carnveli C, Daniella M. Regeneration in Echinoderms: repair, regrowth, cloning. Invertebr Surviv J 2006. doi:10.1016/j.foodqual.2015.02.012.

[2] McClay DR. Evolutionary crossroads in developmental biology: sea urchins. Development 2011;138:2639–48. doi:10.1242/dev.048967.

[3] Mashanov VS, Zueva OR, García-Arrarás JE. Transcriptomic changes during regeneration of the central nervous system in an echinoderm. BMC Genomics 2014. doi:10.1186/1471-2164-15-357.

[4] Clouse RM, Linchangco Jr G V, Kerr AM, Reid RW, Janies DA, Linchangco G V., et al. Phylotranscriptomic analysis uncovers a wealth of tissue inhibitor of metalloproteinases variants in echinoderms. R Soc Open Sci 2015;2:150377. doi:10.1098/rsos.150377.

[5] Uthicke S, Schaffelke B, Byrne M. A boom-bust phylum? Ecological and evolutionary consequences of density variations in echinoderms. Ecol Monogr 2009. doi:10.1890/07-2136.1.

[6] Sodergren E, Weinstock GM, Davidson EH, Cameron RA, Gibbs RA, Angerer RC, et al. The genome of the sea urchin Strongylocentrotus purpuratus. Science (80-) 2006;314:941–52. doi:10.1126/science.1133609.

[7] Dunn CW, Luo X, Wu Z. Phylogenetic analysis of gene expression. Integr Comp Biol 2013;53:847–56. doi:10.1093/icb/ict068.

[8] Janies DA, Voight JR, Daly M. Echinoderm Phylogeny Including Xyloplax, a Progenetic Asteroid. Syst Biol 2011;60:420–38. doi:10.1093/sysbio/syr044.

[9] Hibino T, Loza-Coll M, Messier C, Majeske AJ, Cohen AH, Terwilliger DP, et al. The immune gene repertoire encoded in the purple sea urchin genome. Dev Biol 2006. doi:10.1016/j.ydbio.2006.08.065.

[10] Janies DA, Witter Z, Linchangco GV, Foltz DW, Miller AK, Kerr AM, et al. EchinoDB, an application for comparative transcriptomics of deeply-sampled clades of echinoderms. BMC Bioinformatics 2016;17. doi:10.1186/s12859-016-0883-2.

[11] Livingston BT, Killian CE, Wilt F, Cameron A, Landrum MJ, Ermolaeva O, et al. A genome-wide analysis of biomineralization-related proteins in the sea urchin Strongylocentrotus purpuratus. Dev Biol 2006;300:335–48. doi:10.1016/j.ydbio.2006.07.047.

[12] Reinardy HC, Emerson CE, Manley JM, Bodnar AG. Tissue regeneration and biomineralization in sea urchins: Role of Notch signaling and presence of stem cell markers. PLoS One 2015;10:e0133860. doi:10.1371/journal.pone.0133860.

[13] Leaf DS, Anstrom JA, Chin JE, Harkey MA, Showman RM, Raff RA. Antibodies to a fusion protein identify a cDNA clone encoding msp130, a primary mesenchyme-specific cell surface protein of the sea urchin embryo. Dev Biol 1987;121:29–40. doi:10.1016/0012-1606(87)90135-7.

[14] Illies MR, Peeler MT, Dechtiaruk AM, Ettensohn CA. Identification and developmental expression of new biomineralization proteins in the sea urchin Strongylocentrotus purpuratus. Dev Genes Evol 2002;212:419–31. doi:10.1007/s00427-002-0261-0.

[15] Ettensohn CA. Horizontal transfer of the msp130 gene supported the evolution of

metazoan biomineralization. Evol Dev 2014;16:139–48. doi:10.1111/ede.12074.

[16] Smith AB. Classification of the Echinodermata. Palaeontology 1984;27:431–59.

[17] Sumrall CD, Sprinkle J. Phylogenetic analysis of living Echinodermata based on primitive fossil taxa. Echinoderms San Fr., Rotterdam: A.A. Balkema; 1998, p. 81–7.

[18] Littlewood DTJ, Smith AB, Clough KA, Emson RH. The interrelationships of the echinoderm classes: morphological and molecular evidence. Biol J Linn Soc 1997;61:409–38. doi:10.1111/j.1095-8312.1997.tb01799.x.

[19] Pisani D, Feuda R, Peterson KJ, Smith AB. Resolving phylogenetic signal from noise when divergence is rapid: A new look at the old problem of echinoderm class relationships. Mol Phylogenet Evol 2012;62:27–34. doi:10.1016/j.ympev.2011.08.028.

[20] Bather FA, Goodrich ES (Edwin S, Lankester Sir, ER (Edwin R. A treatise on zoology / edited by E. Ray Lankester. . vol. 2. London : A. & C. Black,; 1900.

[21] Mooi R, David B. What a New Model of Skeletal Homologies Tells Us About Asteroid Evolution. Am Zool 2000;40:326–39. doi:10.1668/0003-1569(2000)040[0326:WANMOS]2.0.CO;2.

[22] Janies D, Mooi R. Xyloplax is an asteroid. Echinoderm Res 1998:311–6.

[23] Janies D. Phylogenetic relationships of extant echinoderm classes. Can J Zool 2001;79:1232–50. doi:10.1139/z00-215.

[24] Candia-Carnevali MD, Thorndyke MC, Matranga V. Regenerating Echinoderms: A Promise to Understand Stem Cells Potential. Stem Cells Mar. Org., Dordrecht: Springer Netherlands; 2009, p. 165–86. doi:10.1007/978-90-481-2767-2_7.

[25] Culver SJ, Pojeta J, Repetski JE. First record of Early Cambrian shelly microfossils from west Africa. Geology 1988;16:596–9. doi:10.1130/0091-7613(1988)016<0596:FROECS>2.3.CO;2.

[26] Cohen PA, Schopf JW, Butterfield NJ, Kudryavtsev AB, Macdonald FA. Phosphate biomineralization in mid-neoproterozoic protists. Geology 2011;39:539–42. doi:10.1130/G31833.1.

[27] Yue JX, Holland ND, Holland LZ, Deheyn DD. The evolution of genes encoding for green fluorescent proteins: Insights from cephalochordates (amphioxus). Sci Rep 2016. doi:10.1038/srep28350.

[28] Fitch WM. Distinguishing Homologous from Analogous Proteins. Syst Zool 1970. doi:10.2307/2412448.

[29] Koonin E V. Orthologs, Paralogs, and Evolutionary Genomics. Annu Rev Genet 2005. doi:10.1146/annurev.genet.39.073003.114725.

[30] Smith A. To group or not to group: The taxonomic position of Xyloplax. 6th Int Echinoderm Conf 1988:17–23.

[31] Satoh N, Rokhsar D, Nishikawa T. Chordate evolution and the three-phylum system. Proceedings Biol Sci 2014;281:20141729. doi:10.1098/rspb.2014.1729.

[32] Rouse GW, Jermiin LS, Wilson NG, Eeckhaut I, Lanterbecq D, Oji T, et al. Fixed, free, and fixed: The fickle phylogeny of extant Crinoidea (Echinodermata) and their Permian-Triassic origin. Mol Phylogenet Evol 2013. doi:10.1016/j.ympev.2012.09.018.

[33] Telford MJ, Lowe CJ, Cameron CB, Ortega-Martinez O, Aronowicz J, Oliveri P, et al. Phylogenomic analysis of echinoderm class relationships supports Asterozoa.

Proc R Soc B Biol Sci 2014;281:20140479–20140479. doi:10.1098/rspb.2014.0479.

[34] Mariko Kondo, Koji Akasaka. Current Status of Echinoderm Genome Analysis - What do we Know? Curr Genomics 2012. doi:10.2174/138920212799860643.

[35] Matsubara M, Komatsu M, Wada H. Close Relationship between Asterina and Solasteridae (Asteroidea) Supported by Both Nuclear and Mitochondrial Gene Molecular Phylogenies. Zoolog Sci 2004;21:785–93. doi:10.2108/zsj.21.785.

[36] Perseke M, Fritzsch G, Ramsch K, Bernt M, Merkle D, Middendorf M, et al. Evolution of mitochondrial gene orders in echinoderms. Mol Phylogenet Evol 2008;47:855–64. doi:10.1016/j.ympev.2007.11.034.

[37] Perseke M, Bernhard D, Fritzsch G, Br??mmer F, Stadler PF, Schlegel M. Mitochondrial genome evolution in Ophiuroidea, Echinoidea, and Holothuroidea: Insights in phylogenetic relationships of Echinodermata. Mol Phylogenet Evol 2010;56:201–11. doi:10.1016/j.ympev.2010.01.035.

[38] Scouras A, Smith MJ. A novel mitochondrial gene order in the crinoid echinoderm Florometra serratissima. Mol Biol Evol 2001;18:61–73. doi:10.1093/oxfordjournals.molbev.a003720.

[39] Scouras A, Beckenbach K, Arndt A, Smith MJ. Complete mitochondrial genome DNA sequence for two ophiuroids and a holothuroid: The utility of protein gene sequence and gene maps in the analyses of deep deuterostome phylogeny. Mol Phylogenet Evol 2004;31:50–65. doi:10.1016/j.ympev.2003.07.005.

[40] Scouras A, Smith MJ. The complete mitochondrial genomes of the sea lily Gymnocrinus richeri and the feather star Phanogenia gracilis: Signature nucleotide

bias and unique nad4L gene rearrangement within crinoids. Mol Phylogenet Evol 2006;39:323–34. doi:10.1016/j.ympev.2005.11.004.

[41] Shen X, Tian M, Liu Z, Cheng H, Tan J, Meng X, et al. Complete mitochondrial genome of the sea cucumber Apostichopus japonicus (Echinodermata: Holothuroidea): The first representative from the subclass Aspidochirotacea with the echinoderm ground pattern. Gene 2009;439:79–86. doi:10.1016/j.gene.2009.03.008.

[42] Linchangco GV, Foltz DW, Reid R, Williams J, Nodzak C, Kerr AM, et al. The phylogeny of extant starfish (Asteroidea: Echinodermata) including Xyloplax, based on comparative transcriptomics. Mol Phylogenet Evol 2017;115. doi:10.1016/j.ympev.2017.07.022.

[43] Reich A, Dunn C, Akasaka K, Wessel G. Phylogenomic Analyses of Echinodermata Support the Sister Groups of Asterozoa and Echinozoa. PLoS One 2015;10:e0119627. doi:10.1371/journal.pone.0119627.

[44] O'Hara TDD, Hugall AFF, Thuy B, Moussalli A, O'Hara TD, Hugall AFF, et al. Phylogenomic Resolution of the Class Ophiuroidea Unlocks a Global Microfossil Record. Curr Biol 2014;24:1874–9. doi:10.1016/j.cub.2014.06.060.

[45] Cannon JTT, Kocot KMM, Waits DSS, Weese DAA, Swalla BJJ, Santos SRR, et al. Phylogenomic Resolution of the Hemichordate and Echinoderm Clade. Curr Biol 2015;24:2827–32. doi:10.1016/j.cub.2014.10.016.

[46] Baker AN, Rowe FWE, Clark HES. A new class of Echinodermata from New Zealand. Nature 1986;321:862–4. doi:10.1038/321862a0.

[47] Janies DA, McEdward LR. A Hypothesis for the Evolution of the

Concentricycloid Water-Vascular System. Reprod. Dev. Mar. Invertebr. Pap. from a Symp. held Friday Harb. Lab. Univ. Washington, June 9 - 11, 1994, p. 246–57.

[48] Mooi R, Rowe FWE, David B. Application of a theory of axial and extraxial skeletal homologies to concentricycloid morphology. In: Mooi R, Telford M, editors. Echinoderms, San Fr., Rotterdam: Balkema; 1998, p. 61–2.

[49] Pearse VB, Pearse JS. Echinoderm phylogeny and the place of concentricycloids. Echinoderms through Time 1994:121–6.

[50] Rowe FWE, Baker AN, Clark HES. The Morphology, Development and Taxonomic Status of Xyloplax Baker, Rowe and Clark (1986) (Echinodermata: Concentricycloidea), with the Description of a New Species. Proc R Soc London B Biol Sci 1988;233:431–59. doi:10.1098/rspb.1988.0032.

[51] Gordon A, Hannon GJ. Fastx-toolkit. FASTQ/A short-reads pre-processing tools. Unpubl Http//Hannonlab Cshl Edu/Fastx_ Toolkit 2010.

[52] Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 2013;8:1494–512. doi:10.1038/nprot.2013.084.

[53] Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, et al. The Pfam protein families database. Nucleic Acids Res 2010;38:D211–22.

[54] Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res 2003;13:2178–89. doi:10.1101/gr.1224503.

[55] Katoh K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 2002;30:3059–66.

doi:10.1093/nar/gkf436.

[56] Stamatakis A. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 2006;22:2688–90. doi:10.1093/bioinformatics/btl446.

[57] Kocot KM, Moroz LL, Citarella MR, Halanych KM, Moroz LL, Halanych KM. PhyloTreePruner: A Phylogenetic Tree-Based Approach for Selection of Orthologous Sequences for Phylogenomics. Evol Bioinforma 2013;2013:429. doi:10.4137/EBO.S12813.

[58] Struck TH. The Impact of Paralogy on Phylogenomic Studies - A Case Study on Annelid Relationships. PLoS One 2013;8. doi:10.1371/journal.pone.0062892.

[59] Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 2009;25:1972–3. doi:10.1093/bioinformatics/btp348.

[60] Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl Tutorial. Bioinformatics 2009;25:1972–3. doi:10.1093/bioinformatics/btp348.

[61] Kück P, Meusemann K. FASconCAT: Convenient handling of data matrices. Mol Phylogenet Evol 2010;56:1115–8. doi:10.1016/j.ympev.2010.04.024.

[62] Mah CL. A new species of Xyloplax (Echinodermata: Asteroidea: Concentricycloidea) from the northeast Pacific: comparative morphology and a reassessment of phylogeny. Invertebr Biol 2006;125:136–53. doi:10.1111/j.1744-7410.2006.00048.x.

[63] Wang Y, Coleman-Derr D, Chen G, Gu YQ. OrthoVenn: a web server for genome wide comparison and annotation of orthologous clusters across multiple species.

Nucleic Acids Res 2015;43:W78-84. doi:10.1093/nar/gkv487.

[64] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: Tool for the unification of biology. Nat Genet 2000;25:25–9. doi:10.1038/75556.

[65] Gene Ontology Consortium TGO. The Gene Ontology project in 2008. Nucleic Acids Res 2008;36:D440-4. doi:10.1093/nar/gkm883.

[66] Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot. Plant Bioinforma., 2007. doi:10.1007/978-1-59745-535-0_4.

[67] Tepass U, Truong K, Godt D, Ikura M, Peifer M. Cadherins in embryonic and neural morphogenesis. Nat Rev Mol Cell Biol 2000;1:91–100. doi:10.1038/35040042.

[68] Alimperti S, Andreadis ST. CDH2 and CDH11 act as regulators of stem cell fate decisions. Stem Cell Res 2015. doi:10.1016/j.scr.2015.02.002.

[69] Suzuki SC, Takeichi M. Cadherins in neuronal morphogenesis and function. Dev Growth Differ 2008;50:S119–30. doi:10.1111/j.1440-169X.2008.01002.x.

[70] Will CL, Lührmann R. Spliceosome Structure and Function. Cold Spring Harb Monogr Arch 2006. doi:10.1101/cshperspect.a003707.

[71] Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. QuickGO: A web-based tool for Gene Ontology searching. Bioinformatics 2009. doi:10.1093/bioinformatics/btp536.

[72] Vail L. Arm growth and regeneration in Oligometra serripinna (Carpenter) (Echinodermata : Crinoidea) at Lizard Island, Great Barrier Reef. J Exp Mar Bio Ecol 1989. doi:10.1016/0022-0981(89)90203-7.

[73] Nemet J, Jelicic B, Rubelj I, Sopta M. The two faces of Cdk8, a positive/negative regulator of transcription. Biochimie 2014. doi:10.1016/j.biochi.2013.10.004.

[74] Bradham C, Foltz KR, Beane WS, Arnone MI, Rizzo F, Coffman JA, et al. The sea urchin kinome: a first look. Dev Biol 2006. doi:10.1016/j.ydbio.2006.08.074.

[75] Firestein R, Bass AJ, Kim SY, Dunn IF, Silver SJ, Guney I, et al. CDK8 is a colorectal cancer oncogene that regulates β-catenin activity. Nature 2008. doi:10.1038/nature07179.

[76] Fryer CJ, White JB, Jones KA. Mastermind recruits CycC:CDK8 to phosphorylate the Notch ICD and coordinate activation with turnover. Mol Cell 2004. doi:10.1016/j.molcel.2004.10.014.

[77] Murdock DJE, Donoghue PCJ. Evolutionary origins of animal skeletal biomineralization. Cells Tissues Organs, vol. 194, 2011, p. 98–102. doi:10.1159/000324245.

[78] Knoll AH. Biomineralization and Evolutionary History. Rev Mineral Geochemistry 2003. doi:10.2113/0540329.

[79] Bottjer DJ, Davidson EH, Peterson KJ, Cameron RA. Paleogenomics of echinoderms. Science (80-) 2006. doi:10.1126/science.1132310.

[80] Gilbert PUPA, Wilt FH. Molecular aspects of biomineralization of the echinoderm endoskeleton. Prog Mol Subcell Biol 2011;52:199–223. doi:10.1007/978-3-642-21230-7_7.

[81] Jackson DJ, MacIs L, Reitner J, Wörheide G. A horizontal gene transfer supported the evolution of an early metazoan biomineralization strategy. BMC Evol Biol 2011;11. doi:10.1186/1471-2148-11-238.

[82] Kudtarkar P, Cameron RA. Echinobase: an expanding resource for echinoderm genomic information. Database (Oxford) 2017. doi:10.1093/database/bax074.

[83] Cameron RA, Samanta M, Yuan A, He D, Davidson E. SpBase: The sea urchin genome database and web site. Nucleic Acids Res 2009. doi:10.1093/nar/gkn887.

[84] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: Architecture and applications. BMC Bioinformatics 2009. doi:10.1186/1471-2105-10-421.

[85] Jia F, Lo N, Ho SYW. The impact of modelling rate heterogeneity among sites on phylogenetic estimates of intraspecific evolutionary rates and timescales. PLoS One 2014. doi:10.1371/journal.pone.0095722.

[86] Tu Q, Cameron RA, Worley KC, Gibbs RA, Davidson EH. Gene structure in the sea urchin Strongylocentrotus purpuratus based on transcriptome analysis. Genome Res 2012. doi:10.1101/gr.139170.112.

[87] Klueg KM, Harkey MA, Raff RA. Mechanisms of evolutionary changes in timing, spatial expression, and mRNA processing in the msp130 gene in a direct-developing sea urchin, Heliocidaris erythrogramma. Dev Biol 1997. doi:10.1006/dbio.1996.8431.

[88] Dunning Hotopp JC. Horizontal gene transfer between bacteria and animals. Trends Genet 2011. doi:10.1016/j.tig.2011.01.005.

[89] Albalat R, Cañestro C. Evolution by gene loss. Nat Rev Genet 2016. doi:10.1038/nrg.2016.39.

[90] Szabó R, Ferrier DEK. Another biomineralising protostome with an Msp130 gene and conservation of Msp130 gene structure across Bilateria. Evol Dev 2015.

doi:10.1111/ede.12122.

[91] Luo YJ, Takeuchi T, Koyanagi R, Yamada L, Kanda M, Khalturina M, et al. The Lingula genome provides insights into brachiopod evolution and the origin of phosphate biomineralization. Nat Commun 2015. doi:10.1038/ncomms9301.

[92] Gale AS. Phylogeny and classification of the Asteroidea (Echinodermata). Zool J Linn Soc 1987;89:107–32. doi:10.1111/j.1096-3642.1987.tb00652.x.