Spring 2021

Investigating the Evolutionary Dynamics of Traits in Metazoa

Jennifer L. Spillane
University of New Hampshire, Durham

INVESTIGATING THE EVOLUTIONARY DYNAMICS OF TRAITS IN METAZOA

BY

Jennifer L. Spillane

Bachelor of Arts, Trinity Christian College, 2012

Master of Science, Western Washington University, 2016

DISSERTATION

Submitted to the University of New Hampshire

in Partial Fulfillment of

the Requirements for the Degree of

Doctor of Philosophy

in

Molecular and Evolutionary Systems Biology

May 2021

This dissertation was examined and approved in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Molecular and Evolutionary Systems Biology by:

Dissertation Director, David C. Plachetzki
Associate Professor of Molecular, Cellular, and Biomedical Sciences

Matthew D. MacManes
Associate Professor of Molecular, Cellular, and Biomedical Sciences

M. Sabrina Pankey
Research Scientist

W. Kelley Thomas
Professor of Molecular, Cellular, and Biomedical Sciences
Director, Hubbard Center for Genome Studies

Joseph F. Ryan
Associate Professor of Biology
The Whitney Laboratory for Marine Bioscience, University of Florida

On April 13, 2021

Approval signatures are on file with the University of New Hampshire Graduate School.

DEDICATION

To all the people who taught me how to be a student and a scientist though their classes, books, articles, conversations, and example. Thank you.

ACKNOWLEDGEMENTS

First, I thank my advisors, Dave Plachetzki and Matt MacManes. You took me in when I was on the verge of being a lab-less graduate student. You have always thought of me as a whole person and not just as a scientist or a student. You have guided me through this process with support and compassion in addition to direction and feedback, and I will always be grateful. This dissertation would also not have been possible without the comments, assistance, and thoughtful conversations provided by my other committee members, Sabrina Pankey, Joseph Ryan, and Kelley Thomas. I would also like to thank Toni Westbrook, for teaching me more about bioinformatics and coding than anyone else, and for being such a great help and friend through these years.

I also thank all the funding sources that made this research possible. From the University of New Hampshire, I was supported by the Summer Teaching Assistant Fellowship, Dissertation Year Fellowship, Marine Biology Graduate Program Grant, and Travel Grants. I also benefitted from the Charlotte Magnum Student Support Program, the Society for Systematic Biologists Graduate Student Research Award, the National Science Foundation Collaborative Research: RUI: Biogenesis and evolution of hagfish slime and slime glands (961152880), and Collaborative Research: Dimensions: Evolutionary ecology of sponges and their microbiome drives sponge diversity on coral reefs (1638296).

All of the graduate students who have been on this journey with me have been instrumental to my success. I want to mention specifically Kris Wojtusik, Andrew Lang, Meg Ange-Stark, Curtis Provencher, Sydney Birch, Hannah Pare, Chris Gonzales, Dani Blumstein, and all the regular attendees at the writing group. Your friendship, solidarity, advice, and expertise, both in graduate school and beyond it, have made this experience a joy in the best of times and bearable in the worst of them. Thank you so much. Special thanks goes to the undergraduate students I have worked with on this project, Nhen Hunter and Troy LaPolice. Watching you grow as scientists and getting to participate in that process has been a privilege. I hope my mentorship has been valuable to you during these years, and I'm so proud of you both.

I thank my family: you have supported every degree I've gotten, every new place I've moved, and every new adventure I've gone on. Thank you for giving me the confidence to pursue this dream in the first place, and the perseverance to see it through. And finally, I thank my husband Tyler. You carried me through this dissertation. Thank you for keeping me alive with your delicious food, for calming me down when everything felt like a disaster, for celebrating every little success with me, and for being the absolute best partner I could have. I never, never could have done this without you. Thank you.

TABLE OF CONTENTS

CHAPTER 3

LIST OF TABLES

LIST OF FIGURES

Chapter 1:

Chapter 2:

Chapter 3:

ABSTRACT

Over the last 800 million years, animals have evolved an incredible array of diverse forms, life histories, ecologies, and traits. In the age of genome-scale resources for many animal taxa, researchers have a unique opportunity to investigate animal diversity and evolution through comparative genomic methods. These methods allow for studies not only of current diversity and evolutionary relationships, but also of ancient evolutionary dynamics and genomic repertoire. In order to study the evolution of diverse animal traits in a rigorous way however, researchers must not neglect the fundamental components of a robust comparative genomics study: well-supported phylogenies, high-quality genomic resources, and ways of applying comparative genomic methods to a phylogenetic tree.

Here, I present three studies of animal trait evolution that address each of the three components above. First, I have leveraged current bioinformatic technologies to identify biases in phylogenomic studies stemming from transcriptome assembly errors, and determined the best practices for processing transcriptomic data for these studies (Chapter 1). I found that high-quality transcriptome assemblies yield richer datasets that are less prone to bias and ambiguity when used to create phylogenetic trees. Second, I have sequenced and assembled a new genomic dataset from a unique marine organism which occupies a crucial position for Cnidarian phylogeny (Chapter 2). This new genomic resource is an important contribution to studies of the evolution of novel

cell types and mitochondrial structure. Third, I have investigated the patterns of gene gain and loss that characterize the evolution of one of the earliest-branching metazoan lineages in a well-supported phylogenomic context (Chapter 3). I established that animals in the phylum Porifera have lost traits associated with most other animal lineages, resulting in a derived form in extant sponges. The findings I lay out in this dissertation add to the growing body of knowledge concerning the evolution of non-bilaterian and early-branching metazoan lineages while also providing the scientific community with best practices for the accurate study of diverse traits in Metazoa.

INTRODUCTION

The animals in our world today possess a staggering array of diverse forms and traits. This diversity can manifest in many levels and systems, from specialized protein types like the globins in vertebrates (1), to intricate organ structures as in the compound eyes of insects. Non-bilaterian animals branch from some of the deepest nodes in the Metazoa phylogeny (2). This means that the study of these organisms can provide new information on the origins of traits such as nervous systems and immunity (3,4), but also reveal complexities that are unique to non-bilaterian animals (5–7), giving us a more complete picture of the diversity present within Metazoa.

In order to study the evolution of diverse animal traits in a robust way, we must 1) have *well-supported phylogenies* without which we have no framework on which to place evolutionary changes. We must 2) have *high-quality genomic resources* from taxa spanning the entire diversity of animal life, including those which have historically been overlooked or inaccessible. And we must 3) have ways of applying phylogenetic comparative methods to phylogenies, particularly at important transitions in animal evolution.

The goal of this dissertation is to add to the growing knowledge about the evolution of animal traits by addressing the three needs outlined above. First, I have leveraged current bioinformatic technologies to identify biases in phylogenomic studies stemming from transcriptome assembly errors, and determined the best practices for

processing transcriptomic data for these studies (Chapter 1). Second, I have sequenced and assembled a new genomic dataset from a unique marine organism with implications for Cnidarian phylogeny, as well as the evolution of novel cell types and mitochondrial structure (Chapter 2). Third, I have investigated the patterns of gene gain and loss that characterize the evolution of one of the earliest-branching metazoan lineages (Chapter 3). I have also ensured that my analyses are as reproducible as possible by making all datasets, workflows, and custom scripts for each of my dissertation chapters publicly available.

Chapter 1 – Signal, bias, and the role of transcriptome assembly quality in phylogenomic inference.

Phylogenomics is the necessary first step to studying the evolution of traits in a lineage of organisms. Without a well-supported hypothesis about how animals are related to one another, it is impossible to put traits into an evolutionary context. Transcriptomes have become ubiquitous in current phylogenomic studies (8–12). They provide a means through which researchers can generate a large number of genetic markers without the expense of whole genome sequencing. However, transcriptome assembly is still a complex process, and there are multiple steps at which researchers could introduce bias into their results (13). While many researchers have addressed potential pitfalls in different aspects of phylogenomic data matric construction and

analysis (14–23), few have considered possible biases introduced at the earlier and more fundamental stage of primary transcriptome assembly.

In Chapter 1 I examine the effects of transcriptome assembly quality on the number and identity of orthogroups obtained as well as differences in the quality of the partition alignments compared to those from higher-quality transcriptomes. I used a well-characterized quantitative metric (Transrate score (24)) to evaluate transcriptome assemblies and to construct two separate phylogenomic datasets: one of high quality and one of intentionally low quality. I then performed identical phylogenomic analyses on each dataset and assessed their relative phylogenetic performance. I find that assembly quality, when all other factors are controlled, can have a dramatic impact on phylogenomic analyses in three ways. First, the richness and size of the dataset can differ profoundly when assembly errors are prevalent in the data. Second, alignments created from low-quality assemblies are more prone to ambiguity and compositional bias than their high-quality counterparts. And third, the partitions derived from high-quality assemblies have greater phylogenetic signal to resolve true evolutionary relationships than partitions derived from low-quality assemblies. This work will lead to fewer inaccurate inferences about organisms' evolutionary relationships, and allow the scientific community to ensure that it is using the best information possible to support hypotheses about animal evolution.

Chapter 2 – The first genome assembly of a cerianthid, *Pachycerianthus borealis*

A broad and complete taxonomic sampling is essential to studies of complex trait evolution. The genomics revolution has allowed the sequencing of more and more organisms, however species have not been sequenced evenly across taxonomic groups. Marine invertebrates in particular are underrepresented in genome-scale resources, with some whole phyla lacking genomic representation. Even in clades with more numerous genetic resources overall, there remain unique groups of organisms that are excluded from studies of complex animal traits because of their lack of these resources. If the underrepresented organisms possess unique traits, cell types, or behaviors, overlooking these animals will present a limited view of complex trait evolution in Metazoa.

One such clade is the Ceriantharia, in phylum Cnidaria. These organisms, the tube-dwelling anemones, form their own subclass within the Cnidarian class Anthozoa. While Cnidaria as a whole is represented by a growing number of genomic datasets (6,25), a whole-genome sequence from any member of the cerianthids is lacking, and studies of these organisms and of Cnidaria are hindered by this exception. In this chapter I present the first genome sequence for a member of Ceriantharia, *Pachycerianthus borealis*, which will aid in the study of specialized cell type and gene family evolution, Anthozoa phylogenetics, and mitochondrial genome structure evolution. I used both long- and short-read sequencing technologies to assemble and polish a 492 Mb genome. It has a scaffold N50 of 396 kb and 18.4% of these scaffolds are larger than 100 kb. I also annotated the genome assembly and found 37,856 predicted proteins. The genome of *Pachycerianthus borealis* has contiguity and

completeness comparable to other anthozoan genomes and will be an asset to further studies of complex trait evolution.

Chapter 3 – Evolutionary dynamics of gene family gain and loss near the root of the Metazoa tree

Finally, we must apply comparative methods to the study of gene family evolution in a phylogenetic context. In studies of animal evolution, researchers often focus on the evolution of novelty and gene gain. As we continue to sequence more genomes to fill in the taxonomic gaps in our comparative genomic studies, we are finding more instances in which the loss of genes or gene families may be an important evolutionary force (26). While in some cases gene loss can be neutral or nearly neutral to an organism (27) and result from a relaxation of selective pressure on that gene, in others it can be directly or indirectly adaptive (28,29) by changing a trait to a more favorable variation or by freeing up limited physiological resources for another purpose.

Scientists have long placed sponges (phylum Porifera) as the first branch of the Metazoa phylogenetic tree because of their apparently simple body plan and lack of traits common to many other metazoan clades (30). More recently, phylogenomic studies have called into question this placement of Porifera and suggest instead that Ctenophora constitutes the first branch of the Metazoa tree (2,31). The growing evidence in support of this hypothesis has caused the scientific community to reconsider when early animals may have evolved certain traits and what the genic

repertoire of the animal ancestor may have been. If the first poriferan was relatively simple, modern sponges may have retained that simplicity through evolutionary time, however if the ancestral poriferan had complex traits that were more similar to other animal lineages, then extant sponges may represent a loss of some of those traits.

In Chapter 3, I used a dataset of 114 species from across Metazoa and Holozoa to construct a well-supported phylogeny and identify gene families. I then used a Dollo parsimony approach to detect gains and losses of these gene families at the ancestral Porifera node and other deep nodes of the Metazoa tree. I found that sponges have lost gene families associated with tissue-grade multicellularity, developmental-morphogenic processes, and nervous systems, and have gained gene families that may help facilitate interactions with diverse microbial communities. I also found that the ancestral Metazoa node gains a substantial number of gene families relating to multicellular processes, the branch directly after (Porifera+ParaHoxozoa) gains a greater number, many of which are implicated in the development of sensory mechanisms and nervous systems. While the branching order of Porifera and Ctenophora has little effect on the gains and losses at these two branches, constraining the tree to reflect a Porifera-first hypothesis eliminates the Porifera+ParaHoxozoa node and concentrates its associated gains onto the ancestral Metazoa node instead. These analyses show that modern sponges represent a degeneration of ancestral complexity regardless of phylogeny, but that the topology affects hypotheses about the complex evolutionary history of gene family evolution in animals.

References

1. Gotting M, Nikinmaa M. More than hemoglobin – the unexpected diversity of globins in vertebrate red blood cells. Physiol Rep. 2015;3(2):1–8.

2. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, et al. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature. 2008;452(7188):745–9.

3. Moroz LL, Kocot KM, Citarella MR, Dosung S, Norekian TP, Povolotskaya IS, et al. The ctenophore genome and the evolutionary origins of neural systems. Nature [Internet]. 2014;510(7503):109–14. Available from: http://dx.doi.org/10.1038/nature13400

4. Muller WEG, Muller, Isabel M. Origin of the Metazoan immune system: identification of the molecules and their functions in sponges. Integr Comp Biol. 2003;43:281–92.

5. Babonis LS, Martindale MQ. Old Cell, new trick? Cnidocytes as a model for the evolution of novelty. Integr Comp Biol. 2014;54(4):714–22.

6. Gold DA, Katsuki T, Li Y, Yan X, Regulski M, Ibberson D, et al. The genome of the jellyfish Aurelia and the evolution of animal complexity. Nat Ecol Evol [Internet]. 2019;3:96–104. Available from: http://dx.doi.org/10.1038/s41559-018-0719-8

7. Dunn CW, Leys SP, Haddock SHD. The hidden biology of sponges and ctenophores. Trends Ecol Evol [Internet]. 2015;30(5):282–91. Available from: http://dx.doi.org/10.1016/j.tree.2015.03.003

8. Chen X, Zhao X, Liu X, Warren A, Zhao F, Miao M. Phylogenomics of non-model ciliates based on transcriptomic analyses. Protein Cell [Internet]. 2015;6(5):373–85. Available from: http://dx.doi.org/10.1007/s13238-015-0147-3

9. Reich A, Dunn C, Akasaka K, Wessel G. Phylogenomic Analyses of Echinodermata Support the Sister Groups of Asterozoa and Echinozoa. PLoS One. 2015;1–11.

10. Kutty SN, Wong WH, Meusemann K, Meier R, Cranston PS. A phylogenomic analysis of Culicomorpha (Diptera) resolves the relationships among the eight constituent families. Syst Entomol. 2018;(March):1–14.

11. Washburn JD, Schnable JC, Conant GC, Brutnell TP, Shao Y, Zhang Y, et al. Genome-Guided Phylo-Transcriptomic Methods and the Nuclear Phylogentic Tree of the Paniceae Grasses. Sci Rep [Internet]. 2017;7(1):1–12. Available from: http://dx.doi.org/10.1038/s41598-017-13236-z

12. Yang Y, Moore MJ, Brockington SF, Soltis DE, Wong GK, Carpenter EJ, et al. Dissecting Molecular Evolution in the Highly Diverse Plant Clade Caryophyllales Using Transcriptome Sequencing. Mol Biol Evol. 2015;32(8):2001–14.

13. Vijay N, Poelstra JW, Kunstner A, Wolf JBW. Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico

assessment of RNA-seq experiments. Mol Ecol. 2013;22:620–34.

14. Whelan N V., Kocot KM, Moroz LL, Halanych KM. Error, signal, and the placement of Ctenophora sister to all other animals. Proc Natl Acad Sci [Internet]. 2015;112(18):5773–8. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1503453112

15. Blanquart S, Lartillot N. A site- and time-heterogeneous model of amino acid replacement. Mol Biol Evol. 2008;25(5):842–58.

16. Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol Biol. 2014;14(82):1–14.

17. Philippe H, Delsuc F, Brinkmann H, Lartillot N. Phylogenomics. Annu Rev Ecol Evol Syst. 2005;36:541–62.

18. Feuda R, Dohrmann M, Pett W, Philippe H, Rota-Stabelli O, Lartillot N, et al. Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals. Curr Biol. 2017;27(24):3864-3870.e4.

19. Wang HC, Minh BQ, Susko E, Roger AJ. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. Syst Biol. 2018;67(2):216–35.

20. Liu L, Yu L, Edwards S V. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol Biol. 2010;10(302):25–7.

21. Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics [Internet]. 2018;19(153):15–30. Available from: http://dx.doi.org/10.1186/s12859-018-2129-y

22. Simion P, Phillippe H, Baurain D, Jager M, Richter DJ, Di Franco A, et al. A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. Curr Biol. 2017;27:1–10.

23. Borowiec ML, Lee EK, Chiu JC, Plachetzki DC. Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. BMC Genomics [Internet]. 2015;16(2015):987. Available from: http://dx.doi.org/10.1186/s12864-015-2146-4

24. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. TransRate: reference free quality assessment of de-novo transcriptome assemblies. Genome Res. 2016;26.

25. Jiang JB, Quattrini AM, Francis WR, Ryan JF, Rodríguez E, McFadden CS. A hybrid de novo assembly of the sea pansy (Renilla muelleri) genome. Gigascience. 2019;8(4):1–7.

26. Miller DJ, Ball EE, Technau U. Cnidarians and ancestral genetic complexity in the animal kingdom. Trends Genet. 2005;21(10):533–6.

27. Drouin G, Godin J, Pagé B. The Genetics of Vitamin C Loss in Vertebrates. 2011;371–8.

28. Hoballah ME, Gu T, Stuurman J, Broger L, Barone M, Mandel T, et al. Single Gene – Mediated Shift in Pollinator Attraction in Petunia. 2007;19(March):779–90.

29. Jeffrey WR. Regressive evolution in Astyanax cavefish. Annu Rev Genet. 2008;23(1):1–7.

30. Brusca RC, Brusca GJ. Invertebrates. Sinauer Associates, Inc.; 1990. 881–883 p.

31. Ryan JF, Pang K, Schnitzler CE, Nguyen A-D, Moreland RT, Simmons DK, et al. The Genome of the Ctenophore Mnemiopsis leidyi and Its Implications for Cell Type Evolution. Science (80-) [Internet]. 2013;342(6164):1242592–1242592. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1242592

CHAPTER 1

Signal, bias, and the role of transcriptome assembly quality in phylogenomic inference

Abstract

Phylogenomic approaches have great power to reconstruct evolutionary histories, however they rely on multi-step processes in which each stage has the potential to affect the accuracy of the final result. Many studies have empirically tested and established methodology for resolving robust phylogenies, including selecting appropriate evolutionary models, identifying orthologs, or isolating partitions with strong phylogenetic signal. However, few have investigated errors that may be initiated at earlier stages of the analysis. Biases introduced during the generation of the phylogenomic dataset itself could produce downstream effects on analyses of evolutionary history. Transcriptomes are widely used in phylogenomics studies, though there is little understanding of how a poor-quality assembly of these datasets could impact the accuracy of phylogenomic hypotheses. Here we examined how transcriptome assembly quality affects phylogenomic inferences by creating independent datasets from the same input data representing high-quality and low-quality transcriptome assembly outcomes.

By studying the performance of phylogenomic datasets derived from alternative high- and low-quality assembly inputs in a controlled experiment, we show that high-quality transcriptomes produce richer phylogenomic datasets with a greater number of unique partitions than low-quality assemblies. High-quality assemblies also give rise to partitions that have lower alignment ambiguity and less compositional bias. In addition, high-quality partitions hold stronger phylogenetic signal than their low-quality transcriptome assembly counterparts in both concatenation- and coalescent-based analyses.

Our findings demonstrate the importance of transcriptome assembly quality in phylogenomic analyses and suggest that a portion of the uncertainty observed in such studies could be alleviated at the assembly stage.

Introduction

The genomics revolution has resulted in a transformation of the approaches that scientists use to estimate phylogeny by vastly increasing the number of available independent genetic markers (1,2), as well as the number of taxa included in phylogenetic analyses (3). However, for taxa that remain largely unrepresented in publicly available datasets, generating a large number of genetic markers, often accomplished as part of a *de novo* whole genome sequencing project, continues to be a challenge. Transcriptome sequencing is a more accessible method of generating a reduced representation of the nuclear genome that requires fewer sequenced reads and is therefore less expensive than whole genome sequencing (although it is not without its

own challenges, see (4)). In addition, transcriptomes perform comparably to genomes in phylogenomic studies when used with robust methods of ortholog identification (5). For these reasons, data derived from transcriptome assemblies have become widely used in phylogenomic studies and have come to represent a mainstream approach to phylogenetic reconstruction (6–10).

The generation of a phylogenomic data matrix is a complex and critical process, as biases introduced at this point can propagate in downstream analyses in unpredictable ways. Phylogenomic data matrices are composed of multiple (often hundreds of) partitions, alignments of orthologous loci that have been filtered and concatenated together (concatenation-based methods) or analyzed as separate gene trees to inform species trees (coalescent-based methods), resulting in data matrices that are highly dimensional. In addition, phylogenomic datasets are often comprised of an agglomeration of data from multiple research groups that may have leveraged different sequencing and assembly strategies. Therefore it is not surprising that there are still many questions concerning the best practices related to the generation and application of these massive new datasets to phylogenomics (11–13). Many researchers have addressed questions related to the most appropriate modeling schemes for different partitions of the data matrix (14–19). Some have considered the impact of incomplete lineage sorting in phylogenomic reconstruction and have leveraged this property of recently diverged lineages to inform species trees (20,21). Others have sought to examine differential phylogenetic signal among partitions in order to maximize phylogenomic performance (22,23). Increasingly, researchers have added the additional

step of recoding the amino acid data matrix in an attempt to account for saturation and compositional heterogeneity (16,22–24, although see 25). While each of these issues is critical to consider in phylogenomic studies, collectively they deal with aspects of the analyses that occur after transcriptome datasets have been assembled. In most cases, biases introduced during the generation of the primary transcriptome assemblies are not explicitly addressed and may persist in influencing downstream inferences.

Whole transcriptome sequencing is itself a relatively new technology, having gained widespread popularity only in the past decade (28). Therefore, RNA-seq data are commonly treated inconsistently among different phylogenomic studies. While many genomics studies have investigated methodological impacts of read trimming (29,30), error correction (31–33), different approaches to transcriptome assembly (34), and quality assessment (35–37), researchers using transcriptome assemblies for phylogenomic applications have been slow to adopt many of these recommendations (but see 38–41). Phylogenomics studies commonly provide few details regarding the nature and quality of the transcriptome assemblies used as input in phylogenomic workflows.

To date there has been no empirical study of how transcriptome assembly quality may affect downstream phylogenomic analyses, although many impacts are possible. Poor-quality assemblies may alter the accuracy of ortholog prediction, alignment quality, and phylogenetic signal. We predicted that in phylogenomic analyses, poor-quality assemblies would result in differences in the number and identity of orthogroups obtained as well as differences in the quality of the partition alignments compared to

those from higher-quality transcriptomes. Here we examine the effects of transcriptome assembly quality on these metrics. Our research strategy is to eliminate as many variables that arise from phylogenomic workflows as possible so that we can attribute discrepancies in phylogenomic results to the differences in transcriptome assembly quality. We use a well-characterized quantitative metric (*TransRate* score, see Methods; (37)) to evaluate transcriptome assemblies and to systematically construct two separate phylogenomic datasets: one of high quality and one of intentionally low quality. We then perform identical phylogenetic analyses on each dataset, allowing the identification of discrepancies between them and the assessment of their relative phylogenomic performance. We find that high-quality transcriptomes produce larger phylogenomic datasets with partitions that have less alignment ambiguity, weaker compositional bias, and are more concordant with the constraint tree, in both concatenation- and coalescent-based analyses, than datasets derived from low-quality transcriptome assemblies. Our results indicate that a portion of the uncertainty in phylogenomic studies likely stems from issues related to the initial assemblies used in preparing phylogenomic data matrices.

Methods

Read selection and assembly

To understand the effects of transcriptome assembly quality on phylogenomic inference, we created two datasets, one of high and one of low quality, from publicly available transcriptomic reads (see Additional File 1 for more information on data availability). All

read data are available on the European Nucleotide Archive (Table 1). We focused on craniates because there are few remaining disputes on the craniate phylogeny (43) and these well-established phylogenetic relationships serve as a comparison to the topologies found using our high- and low-quality transcriptome assemblies. Our research strategy was to assemble high- and low-quality transcriptomes from the same set of reads. We obtained Illumina-generated paired-end liver transcriptomic reads for 37 vertebrate species spanning the majority of the diversity contained within the clade as well as one craniate outgroup. We assembled each read set using the Oyster River Protocol (ORP) version 2.2.3 (34) on a Linux computer with 24 CPUs and 128GB of RAM. In brief, this protocol begins by adapter- and quality-trimming reads using *Trimmomatic* version 0.38 (54) as per recommendations in MacManes (29), after which it corrects read errors using *Rcorrector* version 1.0.8 (32) following recommendations from MacManes and Eisen (31). The ORP then assembles trimmed and corrected reads using three different assemblers: *Trinity* version 2.8.5 (55) with a kmer length of 25, *Trans-ABySS* version 2.0.1 (56) with a kmer length of 32, and *rnaSPAdes* version 3.14 (57) using kmer lengths of 55 and 75. The protocol continues by merging the resultant four assemblies and clustering them into isoform groups. The ORP then scores all transcripts using *TransRate* version 1.0.3 (37) which maps the read sets onto the assembly and, based on the mapping, detects assembly errors such as fragmentation, chimerism, and local misassembly. *TransRate* then uses this error information to assign quality scores to each transcript before integrating these individual scores into a score for the assembly as a whole. The ORP selects the member of each

isoform group with the highest *TransRate* score and places it into a new file. Finally, the protocol uses *cd-hit-est* version 4.8.1 (58) and a 98% sequence identity threshold to reduce transcript redundancy. The assemblies produced by the ORP are therefore populated by the highest quality, non-redundant sequences produced by any of the five possible assembly strategies (34). A graphical summary of this protocol and our phylogenomic pipeline can be found in Figure 1.

Quality analysis and high- and low-quality dataset construction

We evaluated each of the five assemblies generated from the ORP (from *Trinity*, *TransABySS*, *rnaSPAdes* at two kmer lengths, and the final ORP assembly) for each species in two main ways. We used *BUSCO* version 3.0.1 (59), which uses benchmarking universal single copy orthologs to measure the genic completeness of an assembly. In addition, because we were primarily interested in assessing the structural differences in the transcriptome assemblies arising from errors during the assembly process, we generated *TransRate* scores for each assembly. Of the five assemblies for each species, we chose the assembly with the highest overall *TransRate* score to be part of the high-quality dataset, and the one with the lowest overall score to be part of the low-quality dataset. We selected assemblies for each dataset regardless of which assembler produced them, resulting in datasets that contain transcriptomes from multiple different programs. This was done in part to simulate transcriptomic datasets in other studies that may be constructed from preexisting transcriptome assemblies, rather than those that have reassembled each dataset using the same program and to provide

appropriate contrast between the high- and low-quality datasets. We performed all subsequent steps on both datasets in parallel.

Orthogroup inference, statistics, and data partition creation

We used *TransDecoder* version 5.5.0 (60) to translate all transcript sequences to amino acid sequences. The transcriptome assembly process assigns each new transcript a unique name so that it can be differentiated within the assembly. This means that the high- and low-quality assemblies do not share identical transcripts or names common to both assemblies, making the direct comparison of sequences impossible. To circumvent this issue, we added the *Mus musculus* reference transcriptome (release 96) (61) to both datasets just before the *TransDecoder* step so that a *Mus* sequence would be present in many orthogroups and partitions downstream. This created a common naming system by which we could compare the content of orthogroups and partitions derived from assemblies of high and low quality later in the analysis.

For each dataset (containing either the high-quality or low-quality transcriptome assemblies for the 38 craniate species plus the *Mus* reference transcriptome) we performed a separate *OrthoFinder* version 2.3.3 analysis (48,49). We then used linear regressions in *R* version 3.5.2 (62) to evaluate the relationship between the total number of orthogroups found for each taxon and three other measures: the total number of transcripts in each assembly, the overall *TransRate* score, and the *BUSCO* complete score. We also plotted the distributions of these three measures for each

dataset and performed Wilcoxon rank sum tests in *R* to determine if they were statistically different.

We filtered the resulting orthogroups so that we retained only those that had each taxon represented by at least one sequence. From these, we obtained one-to-one orthologs using *PhyloTreePruner* (63). We realigned these sequences using *MAFFT* version 7.305b using the "auto" setting (64), and filtered the alignments for poorly aligned or divergent regions using *Gblocks* version 0.91b (65,66) with options "-b2=0.65 -b3=10 -b4=5 -b5=a" in the script "gblocks_wrapper.pl" (67). Finally, we concatenated all sequences into a NEXUS file for each dataset. We measured the lengths of the alignments both before and after *Gblocks* and compared the content of both groups of partitions by using the *Mus* sequence headers as common identifiers that were present in both datasets and determined the numbers of unique and shared partitions. We then used *IQ-TREE* version 1.6.12 under the LG model (42) to find individual gene trees for each partition in each dataset.

GO analysis and alignment metrics

To investigate the differences in content and qualities of the partitions between the two datasets, we separated the partitions into groups containing only those that were unique to each dataset, and only those that were shared between the two datasets. We used *InterProScan* version 5.31-70.0 (68) to annotate the partitions unique to each dataset and then performed a gene ontology (GO) analysis with *topGO* version 2.32.0 (69) in *R* version 3.5.2 (62) to check for any functional enrichment or depletion bias in the

partitions of either dataset. For each partition common to both datasets, we extracted various alignment metrics from the log and information files generated while making partition trees in *IQ-TREE*. These included percent constant sites, percent parsimony-informative sites, number of sequences that failed the chi^2 composition test (which we normalized by alignment length), and the number of sequences that contained more than 50% gaps or ambiguity. To test for significant differences, we performed Wilcoxon rank sum tests in *R* version 3.5.2 (62) between the two datasets for each of these measures.

Constraint tree and comparisons of partition trees

The phylogenetic relationships among the 38 craniate species for which we obtained liver RNA-seq data are well-supported by previous work (43). Therefore, we used a tree that reflects the most well-supported hypothesized relationships for comparison against the partition trees. Using *Mesquite* version 3.6 (70), we constructed a constraint tree that reflects the widely accepted topology for craniates. We used the high-quality dataset NEXUS alignment file along with this topology to estimate the constraint tree topology with branch lengths in *IQ-TREE* using the LG model (42). We calculated RF distances (45) from the partition trees in each dataset to the constraint tree using *phangorn* version 2.5.5 (71) in *R* version 3.5.2 (62). This metric measures the differences in topology (RF distance) from the partition trees to the constraint tree, with smaller numbers indicating less conflict between the two trees. We also calculated ICA values between the individual partition trees and the constraint tree using *RAxML*

version 8.2.11 (72). The ICA refers to the degree of certainty for each internal node of the tree compared to the constraint tree when all other conflicting bipartitions are taken into account for that dataset. Numbers close to 1 show a lack of conflict between the partition tree and the constraint tree (46). We tested for significant differences between the two dataset distributions using a Wilcoxon rank sum test in *R* version 3.5.2 (62) for both RF distances and ICA values. Finally, we created species trees using the 332 gene trees that were common to both the high-quality and low-quality datasets with a coalescent method implemented in *ASTRAL* version 5.7.4 (20,47). We calculated the normalized quartet score for each tree, which represents the percentage of quartet trees in the input trees that are satisfied by the species tree and ranges from 0-1, with higher numbers indicating less discordance.

Results

Datasets chosen based on TransRate scores have different numbers of transcripts, but show little variation in BUSCO score

Our study design controls for several factors that could preclude direct comparison between empirical outcomes in phylogenomic analyses. We focus on the craniate phylogeny because there is little debate about the major relationships within the group and because RNA-seq read data are available from the same tissue type (liver) for a wide range of taxa. The read sets used in this study ranged in size from 13.7 million read-pairs (*Calidris pugnax*) to 46.4 million read-pairs (*Ambystoma mexicanum*). We prepared one high-quality dataset and one low-quality dataset from the same read sets

using the Oyster River Protocol (ORP) (34), an assembly pipeline that creates five different transcriptome assemblies for each raw RNA-seq dataset, calculates quality scores for each one, and produces a merged transcriptome assembly consisting of the highest quality unique transcripts (Figure 1). We leverage the ORP here to intentionally create low-quality transcriptome assemblies that represent real-world empirical outcomes, in addition to high-quality transcriptome assemblies, for each taxon. Reads assembled into significantly fewer transcripts in the high-quality dataset compared to the low-quality dataset ($P < 0.001$, Figure 2A), with an average of 178,473 and 321,306 transcripts per assembly respectively. The *BUSCO* scores and numbers of orthogroups recovered from orthology analysis of each assembly were both higher on average in the high-quality dataset (Table 1). We compared the number of transcripts in each assembly with the number of orthogroups found for that assembly and identified a significant relationship between these measures in both datasets (linear regression: high-quality dataset, $P = 0.001$; low-quality dataset, $P = 0.002$; Figure 2B). The high-quality dataset based on overall *TransRate* assembly scores had a median *TransRate* score of 0.47236 (ranging from 0.23542 to 0.68372), while the low-quality dataset's median *TransRate* score was 0.15943 (ranging from 0.09216 to 0.25281), and overall *TransRate* scores of the two datasets were significantly different from one another ($P < 0.001$; Figure 2C). We did not find a significant relationship between the overall *TransRate* scores of assemblies and the number of orthogroups obtained for each assembly (linear regression: high-quality dataset, $P = 0.43$; low-quality dataset, $P = 0.51$; Figure 2D). The number of orthogroups for each dataset was higher in the high-

quality dataset, but still largely comparable to the low-quality dataset with the exception of two low-quality read datasets, *Takifugu rubripes* and *Callorhinchus milii*. Each of these datasets recovered much lower numbers of orthogroups than other taxa in the low-quality dataset. In addition to *TransRate* evaluations, the *BUSCO* scores for the low-quality *T. rubripes* and *C. milii* assemblies were also dramatically lower than all other *BUSCO* scores in both datasets (2.7% and 7.2% respectively, compared to the next lowest score: 42.9% for *Notechis scutatus*). However, the overall *BUSCO* scores for the high- and low-quality datasets were not significantly different (Wilcoxon rank sum: $P = 0.24$, Figure 2E). We observed a significant relationship between *BUSCO* score and number of orthogroups recovered in both datasets (linear regression: high-quality dataset, $P = 0.001$; low-quality dataset, $P = 0.001$; Figure 2F).

High-quality assemblies result in a larger number of partitions after processing

Next, we isolated one-to-one orthologs that were present in 100% of taxa. After aligning and filtering these orthologs into partitions we observed that one major impact of assembly quality on phylogenomic data matrix construction is the scale of the resulting data. We obtained 2,016 data partitions from the high-quality dataset, whereas we recovered only 408 data partitions from the low-quality dataset. 332 data partitions in both the high- and low-quality datasets included an identical reference sequence from the *Mus musculus* reference transcriptome, demonstrating that a majority of the data partitions recovered from the low-quality dataset are also represented in the high-quality dataset (Figure 3A). The high-quality dataset however, included many more unique

sequence partitions (1684 unique partitions compared to 76, Figure 3A). The distributions of alignment lengths between datasets differed significantly before alignment filtering (Wilcoxon rank sum, $P = 0.02$; Figure 3B) with alignments in the high-quality dataset being longer on average, but not after alignment filtering (Wilcoxon rank sum, $P = 0.79$; Figure 3C).

High-quality alignments possess reduced compositional bias and alignment ambiguity

In order to draw direct comparisons between the partitions derived from the high- and low-quality datasets, we examined the alignment statistics of the 332 partitions that were shared between them. The percentage of constant sites in each alignment was not significantly different between the high- and low-quality datasets (Wilcoxon rank sum, $P = .37$, Figure 4A). Similarly, the percentage of parsimony-informative sites in the alignments did not differ significantly between the two datasets (Wilcoxon rank sum, $P = .89$, Figure 4B). However, the number of sequences that failed the composition chi^2 test (42) and the number of sequences with over 50% alignment ambiguity were significantly different between the two datasets (composition – Wilcoxon rank sum, $P = .006$, Figure 4C; ambiguity – Wilcoxon rank sum, $P < .001$, Figure 4D), and both of these metrics were higher in the low-quality dataset.

No bias in gene content in partitions from both high- and low-quality datasets

Phylogenetic information content of a given phylogenomic data matrix could be impacted if the partitions themselves are drawn from a biased set of loci. In order to

understand the genetic composition of phylogenomic datasets derived from high- and low-quality assemblies, we conducted gene ontology (GO) analysis of the recovered partitions. We did not observe enrichment for functional category in either the high- or low-quality datasets.

Partitions from high-quality assemblies recapitulate the constraint tree to a larger extent than those from low-quality assemblies in both concatenation- and coalescent-based analyses

Finally, we sought to understand the impact of assembly quality on phylogenetic signal. We first compared the two datasets to a constraint tree representing the current view of craniate relationships (43,44) by using Robinson-Foulds (RF) distances and internode certainty all (ICA) values in concatenation analyses. RF distances reflect topological differences between partition subtrees and the constraint tree (45), whereas ICA values indicate the proportion of data partitions for the high-quality and low-quality datasets that support each node in our constraint tree (46). We found that the high-quality dataset had significantly lower RF values overall than the low-quality dataset (Wilcoxon rank sum, $P < .001$; Figure 5), indicating a shorter distance to the constrained craniate tree for the partitions in the high-quality dataset. The partitions derived from the high-quality dataset possessed characteristically higher ICA values than those from the low-quality dataset, although the distributions of scores were not significantly different (Wilcoxon rank sum, $P = .47$; Figure 6) likely due to low statistical power. We also investigated the relative performance of the two datasets in coalescent-based analyses

using *ASTRAL* (20,47). Similarly, we found that the high-quality dataset produced gene trees with less discordance to the estimated species tree than their low-quality counterparts, with a normalized quartet score of 0.75 for the high-quality partitions compared to 0.73 for the low-quality partitions. Both datasets resolved the same topology in *ASTRAL* analyses (Figure 7).

In summary, we find that datasets derived from high-quality transcriptome assemblies yield larger phylogenomic matrices than those from low-quality transcriptome assemblies. In addition to being more numerous, the data partitions in the high-quality dataset are also less compositionally biased, have less alignment ambiguity, and are less discordant with the constraint tree.

Discussion

Given the ubiquity of transcriptome usage phylogenomics, we sought to understand how sub-optimal data handing practices during the assembly process may affect downstream phylogenomic analyses. We observed a general trend in our analyses where more accurate transcriptome assemblies resulted in phylogenomic datasets with a greater number of unique data partitions, longer alignments, fewer ambiguous regions, less compositional bias, greater consistency with the known phylogeny in concatenation-based analyses, and higher normalized quartet scores in coalescent-based analyses. We did not uncover any functional biases in the GO terms associated with either dataset.

High-quality assemblies result in a larger number of partitions after phylogenomic processing

The most dramatic difference between the high- and low-quality phylogenomic data matrices is the number of orthogroups that contained all species. After estimating one-to-one orthologs, aligning the orthologs, and filtering the alignments, this difference led to ~five times the number of data partitions in the high-quality dataset compared with the low-quality dataset. Transcriptomic assembly errors that are expected to pervade low-quality assemblies include the generation of chimeric transcripts, the generation of incomplete transcripts, or the failure to generate transcripts due to missing data (34,37). Our results from analyses of the low-quality assemblies indicate that incompletely assembled transcripts may be at least partially responsible for the differences in partition number because the partition alignments before filtering are significantly longer in the high-quality dataset, indicating fewer incompletely assembled transcripts in the latter. While *OrthoFinder* (48,49) may be somewhat robust to these issues, when more complete sequence information is provided in high-quality transcripts, *OrthoFinder* analyses identify significantly greater numbers of orthogroups that contain a high proportion of species and therefore greater numbers of orthologs. Missing transcripts could also impact the accuracy of downstream analyses and the establishment of one-to-one orthologs because, depending on what data are missing, orthologs and paralogs could become conflated between taxa. Our results are consistent with this expectation because among partitions that are shared between high- and low-quality datasets,

those from the high-quality dataset show more accurate phylogenetic signal, as measured by constraint tree analyses in concatenation analyses and in coalescent approaches (see below).

We identified two transcriptome assemblies within the low-quality dataset, *Takifugu rubripes* and *Callorhinchus milii*, which have dramatically lower *BUSCO* scores and number of orthogroups recovered than other taxa within the same dataset. We included these two taxa in the analysis despite their extreme *BUSCO* scores for a number of reasons. First, these taxa occupy important phylogenomic positions within the craniate tree and publicly available craniate liver transcriptome datasets are somewhat limited. Second, while the *TransRate* scores for these two taxa are below average for the low-quality dataset (Figure 2C, D), they are well within the distribution of low-quality assembly *TransRate* scores, indicating that these two taxa yield assemblies that are contiguous and correctly assembled to a comparable extent to the other assemblies included in that dataset. While it is standard practice to deposit raw reads into public databases, the read-sets for these two species appeared to have been trimmed prior to public data deposition (50), making them shorter than the other read-sets. We identified average read length as the probable reason for the lack of genic completeness as measured by *BUSCO* for these two taxa. Due to this shorter read length, these two organisms performed especially poorly in *rnaSPAdes* with a kmer length of 75 (only reads of length k+1 are used in assembly), which was subsequently the assembly used in the low-quality dataset for both of these organisms. Importantly, these two species' corresponding assemblies in the high-quality dataset were not

outliers (Figure 2C, D), indicating that a robust assembly strategy can compensate for sub-optimal sequence reads. Therefore, by including these two taxa, we were able to represent a situation commonly encountered in phylogenomic studies that utilize publicly available data – the inclusion of reads of poor quality or that have been previously processed.

The drastic difference in number of partitions in the low-quality dataset compared to the high-quality dataset is due in part to these two taxa having smaller and less complete assemblies than all others. However, when we relax the strict filtering to include orthogroups with up to two missing taxa (thereby giving the low-quality dataset the opportunity to exclude *T. rubripes* and *C. milii*) we find that the high-quality dataset still has over 1600 more partitions than the low-quality dataset, and therefore the inclusion of these taxa is not the only driving force behind the difference in partitions between the datasets. While there are fewer partitions in the low-quality dataset, it is still a sufficient number (408) for most downstream phylogenomic applications. Therefore, we conclude that while the situation encountered with the *T. rubripes* and *C. milii* RNA-seq data has an effect on some aspects of our phylogenomic analysis, their effects are only manifested in analyses of the low-quality assemblies and extend beyond data drop out.

Low-quality assemblies produce alignments with more compositional bias and alignment ambiguity than high-quality assemblies

In the process of making gene trees for each of the data partitions, *IQ-TREE* calculates a number of metrics about the partition alignments and the sequences within them (42). One such test is for compositional homogeneity, which measures the character composition of amino acids in each sequence against the character composition in the whole alignment. Here, we chose to assess changes in compositional heterogeneity using the simple chi^2 test implemented in *IQ-TREE* (42,51). Heterogeneity or bias in amino acid composition can mislead phylogenetic inferences: distantly-related organisms that have high compositional bias may erroneously group together (52). The number of sequences failing the composition test – that is, the number of sequences with higher compositional heterogeneity than expected by chance – was higher in the partitions from the low-quality dataset. Because these partitions have direct counterparts in the high-quality dataset, this difference in compositional heterogeneity is directly attributable to a difference in assembly quality. Similarly, the partitions from the low-quality dataset also contained more sequences with over 50% gaps or ambiguity in the alignment. While global alignments often contain gaps because of insertions or deletions in the sequences, comparison of the two datasets implies that the greater number of gaps in the low-quality dataset also results from incorrect transcriptome assemblies rather than natural variation.

The low-quality dataset contained some partitions that the high-quality dataset did not have. These partitions could be unique transcripts only assembled in the low-quality dataset, or they could be the result of differential pruning of paralogous sequences between the two datasets, resulting in a different *Mus* identifying sequence

in two partitions that represent the same gene family. They might also be erroneous or duplicate partitions that were misidentified during the *OrthoFinder* procedure as separate gene families due to poor assembly quality. In principle, differential data assembly quality could inject bias into the resulting orthogroups if some loci, perhaps short or highly expressed genes, were preferentially assembled among the different datasets, however our GO analyses showed no enrichment or depletion of GO terms in these partitions.

Partitions derived from high-quality assemblies perform better in both concatenation- and coalescent-based phylogenomic analyses

In this study, we used quantitative analyses to assess phylogenomic performance of the high- and low-quality transcriptome assemblies. We showed that the individual partitions included in the high-quality dataset were closer to the constraint tree by calculating RF distances. The high-quality dataset had significantly smaller RF distances to the constraint tree in concatenation-based analyses (Wilcoxon rank sum, $P < .001$) and less discordance in coalescence-based analyses as indicated by normalized quartet score (Figure 7). While the ICA values of the high-quality dataset were not significantly higher than those in the low-quality dataset, the trend shows that ICA values are generally higher among partitions from the high-quality dataset with a greater proportion of partitions falling above 0.6. This indicates that the gene trees estimated from the high-quality dataset partitions are more consistent with the constraint tree of craniates and

show greater phylogenetic signal (53) than the low-quality dataset in concatenated analyses (Figure 6B).

Limitations in data availability and statistical power do not affect our conclusions

Our research strategy was to eliminate as many variables as possible so that we could isolate the effects of assembly quality on phylogenomic performance. These variables include the type of tissue that RNA-seq datasets are derived from and the topology itself. We treat the craniate phylogeny, for which few arguments remain regarding the relationships of the taxa included (43,44), as a "known" parameter to constrain our analyses. In this way we were able assess how close a given analysis accords with that constraint in light of other perturbations like assembly quality. However, it is notable that phylogenomic trees based on the 332 data partitions that are common to both the high-quality and low-quality datasets, using either concatenation- or coalescent-based methods, fail to resolve the craniate phylogeny accurately (Figure 7; Supplementary Figure 1). While this result has no bearing on any of the conclusions presented here, it is likely due to two factors. First, the magnitude of both datasets, 332 partitions, is far fewer than that included in recent well-resolved phylogenomic studies of craniates (43). Here, our utilization of only 332 partitions derives from the necessity that they be shared between the high- and low-quality assemblies, and therefore directly comparable. Second, our taxon sampling is low compared to recent phylogenomic studies of craniates. This is due to the requirement of our study design that RNA-seq reads be derived from a homologous tissue (e.g. liver) across taxa, offering a different type of

direct comparison. While we were able to represent most of the major lineages of craniates with RNA-seq data derived from liver tissue, it was not possible to provide greater taxon sampling given current publicly available data while also preserving taxonomic evenness in sampling across various vertebrate clades.

We also point out that some of the quantitative measures reported here (e.g. ICA) show clear trends that favor the high-quality dataset over the low-quality dataset but are not significantly different. This may be due to intrinsic differences in statistical power that make it unlikely that a significant difference would be identified between datasets for those measures that have fewer data points (RF distances yield one data point per gene tree (332) while ICA scores provide one data point per node (36)). However, we do not observe a single instance of the low-quality dataset being quantitatively or qualitatively better than the high-quality dataset in terms of phylogenetic signal for any of our measures.

Conclusions

Phylogenomic approaches leverage great power to resolve phylogenetic relationships, but they also include many analytical pitfalls associated with ortholog identification, alignment filtering, and model selection. While these pitfalls have been well-characterized, we chose to focus on transcriptome assembly quality – a more fundamental and largely overlooked aspect of phylogenomic analyses. We addressed this problem empirically using a study design that controls for variables including taxon selection, data type, data provenance, and phylogenetic uncertainty. We show that

assembly quality, when all other factors are controlled, can have a dramatic impact on phylogenomic analyses in three ways. First, the richness and size of the dataset can differ profoundly when assembly errors are prevalent in the data. Second, alignments created from low-quality assemblies are more prone to ambiguity and compositional bias than their high-quality counterparts. And third, the partitions derived from high-quality assemblies have greater phylogenetic signal to resolve true evolutionary relationships than partitions derived from low-quality assemblies. We conclude that additional analytical interventions aimed at improving assembly quality, such as the Oyster River Protocol (34), are likely worth the additional effort.

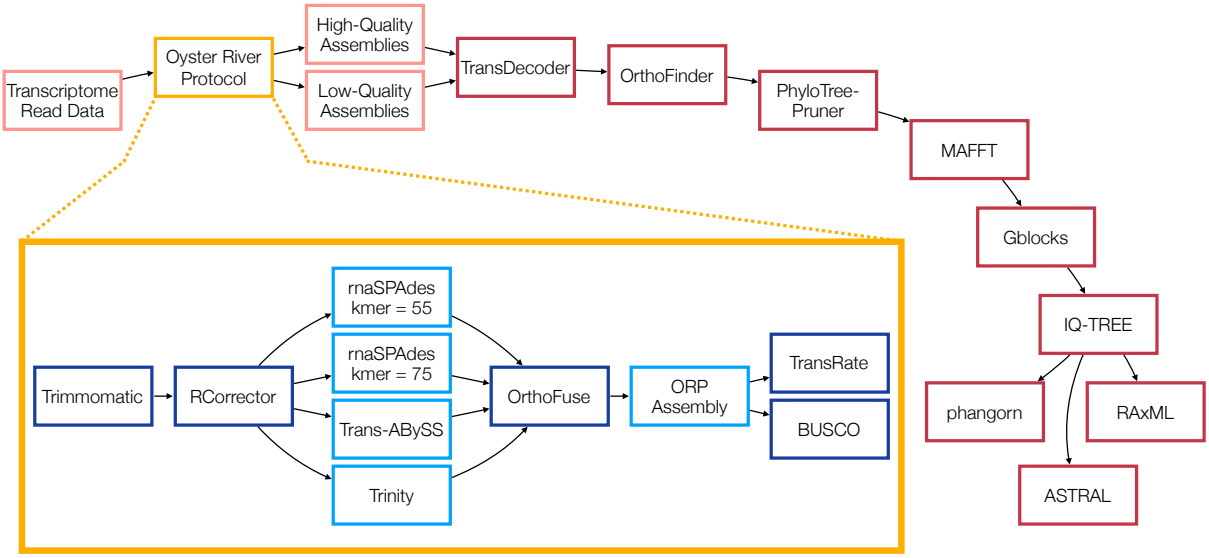

Figure 1: The phylogenomic pipeline used in this analysis from publicly available transcriptomic datasets to partition tree statistics. In the top flowchart red borders indicate bioinformatic tools used while pink ones depict datasets. The Oyster River Protocol is highlighted in yellow, and in the inset: darker blue borders represent steps of the protocol while the resulting transcriptome assemblies are outlined in lighter blue.

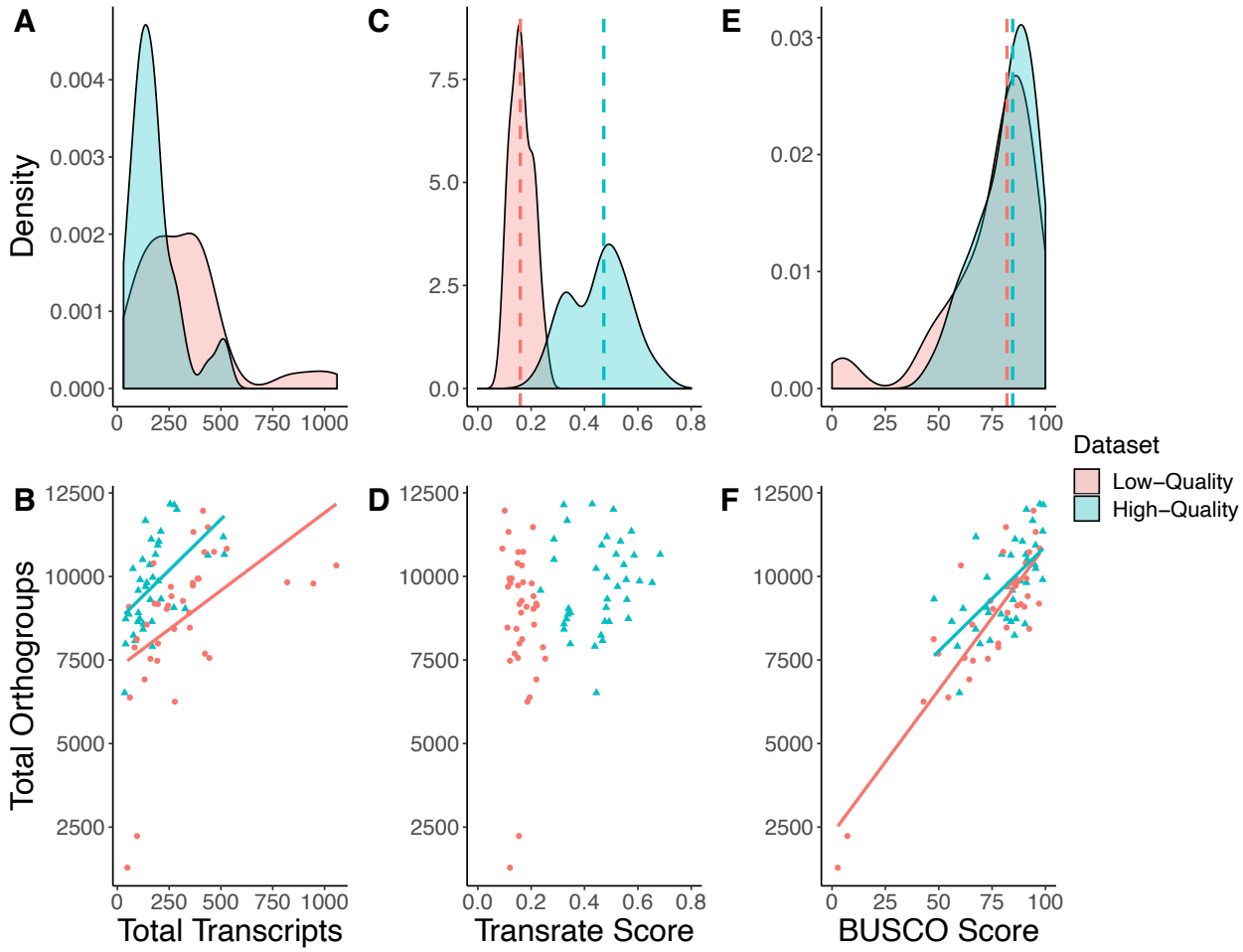

Figure 2: Summary statistics for the high- and low-quality datasets produced. We selected high- and low- quality datasets based on *TransRate* score. This resulted in transcriptome assemblies with both high and low completeness, according to complete *BUSCO* score, in each dataset. Larger assembles in the low-quality dataset did not lead to higher *BUSCO* or *TransRate* scores. Dotted lines in density plots represent medians for each dataset. **A:** Density plot of the total number of transcripts (in thousands) in each transcriptome. **B:** Relationship between the total number of transcripts (in thousands) and the total number of orthogroups. **C:** Density plot of overall *TransRate* scores for each assembly. **D:** Relationship between the overall *TransRate* score and the total number or orthogroups. **E:** Density plot of complete *BUSCO* score for each transcriptome assembly. **F:** Relationship between *BUSCO* score and total number of orthogroups.

Table 1: Read set information and transcriptome assembly metrics. For each species, we assembled the transcriptomic reads using the Oyster River Protocol. Of the five resulting transcriptome assemblies, we chose the one with the highest overall *TransRate* score and the one with the lowest overall *TransRate* score to use in the high- and low-quality datasets, respectively. We also quantified the number of transcripts in each assembly, calculated the complete *BUSCO* score, and inferred orthogroups using *OrthoFinder*.

				High-quality Dataset					Low-quality Dataset				
Species	Accession	Read Length	Number of Reads	Number of Transcripts	*BUSCO* complete	*TransRate* score	Orthogroups	Species-Specific Orthogroups	Number of Transcripts	*BUSCO* complete	*TransRate* score	Orthogroups	Species-Specific Orthogroups
Alligator mississippiensis	SRR629636	100	36,130,137	287695	91.1	0.50848	12004	32	466618	80.2	0.16986	10737	20
Ambystoma mexicanum	SRR5341572	101	46,417,978	209702	98.7	0.57581	11350	59	528158	97.4	0.09216	10832	61
Anas platyrhynchos	SRR7127376	101	20,486,658	142201	91.1	0.65376	9813	8	244848	86.8	0.2212	9129	11
Anolis carolinensis	SRR391653	101	17,152,427	40327	86.2	0.33263	8729	9	56207	90.1	0.18273	9093	9
Astyanax mexicanus	SRR2045431	100	32,893,691	110132	98.7	0.55641	9902	44	180139	97	0.21902	9187	36
Balaenoptera acutotostrata	SRR919296	100	23,923,194	200511	89.2	0.53496	11048	10	364918	86.1	0.15086	9729	14
Bufo bufo	ERR1331718	126	37,410,097	135770	94	0.33512	11671	57	413473	94.4	0.10086	11968	34
Caecilia tentaculata	SRR5591453	101	28,784,422	107413	81.8	0.56427	8737	35	196546	77.9	0.15651	7993	29
Caiman crocodilus	ERR2198478	variable	31,864,053	163595	85.8	0.28529	11113	3	436573	81.5	0.20671	11475	6
Calidris pugnax	ERR1018151	150	13,725,659	78074	85.5	0.46221	8239	10	83535	77.9	0.24439	7880	2
Callorhinchus milii	SRR513760	76	35,000,000	124415	67.3	0.32314	8418	17	95463	7.2	0.15425	2232	13
Canis lupus familiaris	ERR1331673	100	36,371,999	437158	83.8	0.58601	10633	3	819785	86.5	0.1697	9826	10
Dasypus novemcinctus	SRR494766	101	31,705,473	55634	79.2	0.33783	8868	7	192657	66	0.12049	7478	6
Felis catus	ERR1331679	100	40,228,662	516209	79.5	0.51854	10659	8	945952	85.2	0.20215	9790	4
Gadhus morhua	SRR2045420	100	18,943,673	85927	74	0.46787	8082	29	131171	64.3	0.21936	6919	15
Gallus gallus	ERR1298598	100	14,955,711	272485	72.3	0.48137	9069	8	444042	62.1	0.15068	7562	9
Haplochromis burtoni	SRR387451	101	16,142,312	40240	69.3	0.34653	7981	14	60824	54.5	0.19438	6379	48
Homo sapiens	SRR5576267	101	20,633,201	171048	72.6	0.48352	9971	5	317048	74.2	0.16465	9271	8

Ictalurus punctatus	SRR917955	100	28,319,586	99232	83.8	0.49223	8645	32	159608	73	0.25281	7538	34
Latimeria menadoensis	SRR576100	109	39,788,120	101337	73.3	0.34696	8913	69	258443	82.2	0.11311	9692	13
Lepidophyma flavimaculatum	DRR034613	variable	20,350,517	121895	91.4	0.28563	10505	59	174935	90.7	0.14923	10395	37
Lepisosteus oculatus	SRR1287992	101	22,992,842	75239	95.4	0.44361	10235	55	195782	88.5	0.15598	9172	42
Lethenteron camtschaticum	SRR3223459	125	29,559,367	125856	90.4	0.32322	8577	292	274262	92.4	0.14484	8431	93
Lissotriton montandoni	SRR3299753	100	32,548,205	195142	95.4	0.46462	10934	68	387445	91.1	0.12708	9939	56
Notamacropus eugenii	DRR013408, DRR013409, DRR013410	100	24,378,361	198447	88.5	0.60651	9859	24	347172	82.2	0.16235	8917	28
Notechis scutatus	SRR519122	90	25,626,764	168738	58.7	0.43875	7908	31	277137	42.9	0.18596	6254	32
Oophaga sylvatica	SRR9120851	100	22,858,029	166747	56.1	0.47685	8650	18	423029	49.8	0.13789	7690	24
Oryctolagus cuniculus	ERR1331669	100	22,037,691	158880	84.8	0.5591	9304	4	349879	81.8	0.11102	8469	5
Parus major	SRR1847228	101	35,000,000	155826	95	0.54739	10349	16	261539	91.7	0.20877	9408	18
Pelodiscus sinensis	SRR6157006	150	24,740,727	274343	99	0.32231	12143	40	367085	95.4	0.11519	11332	23
Pelusios castaneus	SRR629649	100	45,163,324	254815	97.4	0.42891	12168	31	419831	92.4	0.15182	10728	19
Protopterus sp.	ERR2202465	150	18,298,224	327343	61.4	0.34033	9036	127	141824	65.7	0.21121	8558	89
Rana pipiens	SRR1185245	101	35,791,829	136439	82.2	0.52391	9695	36	238110	75.4	0.20868	9029	16
Rhinella marina	SRR6311453	100	27,446,915	511551	67.4	0.48377	11184	48	1056698	60.4	0.16511	10330	32
Rhinolophus sinicus	SRR2273875	101	30,559,494	184384	90.8	0.68372	10658	14	392613	86.2	0.1185	9933	9
Squalus acanthias	ERR1525379	variable	35,000,000	101153	84.5	0.23542	9582	25	363863	83.2	0.12189	9803	79
Takifugu rubripes	SRR1005688	76	35,796,911	35375	59.7	0.44456	6518	13	48271	2.7	0.1206	1287	22
Trachemys scripta	ERR2198830	150	22,741,770	210713	47.8	0.48531	9322	6	94129	47.6	0.16631	8123	3

Table 1: Read set information and transcriptome assembly metrics. For each species, we assembled the transcriptomic reads using the Oyster River Protocol. Of the five resulting transcriptome assemblies, we chose the one with the highest overall *TransRate* score and the one with the lowest overall *TransRate* score to use in the high- and low-quality datasets, respectively. We also quantified the number of transcripts in each assembly, calculated the complete *BUSCO* score, and inferred orthogroups using *OrthoFinder*.

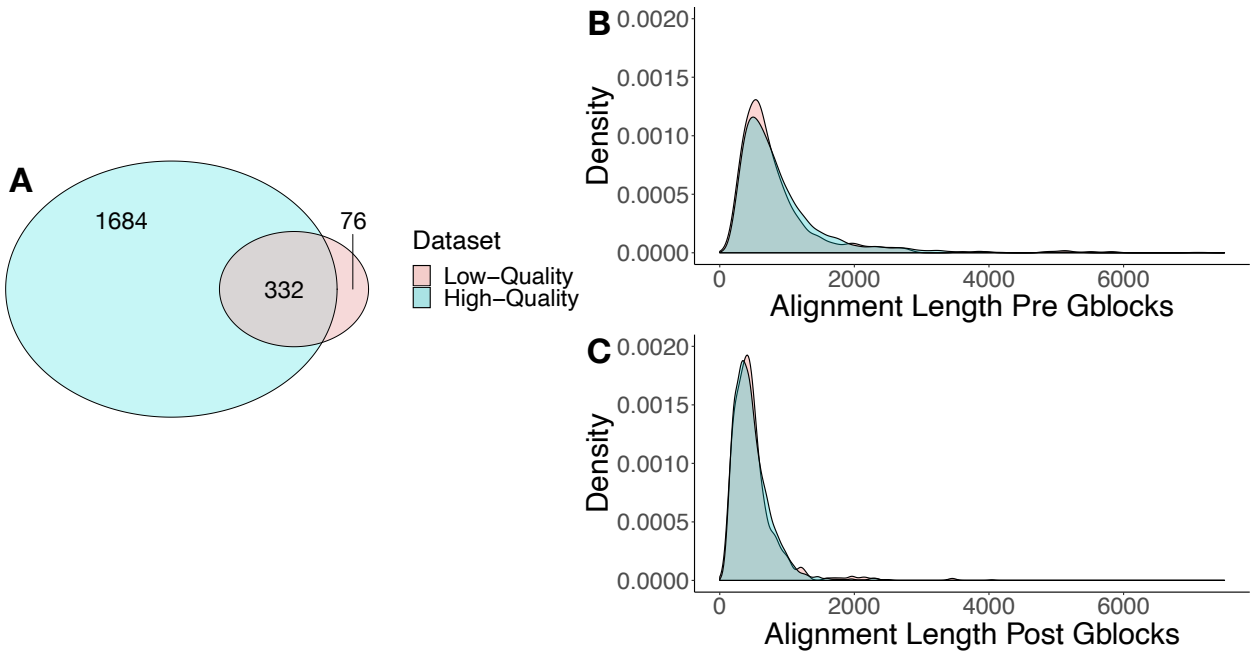

Figure 3: Length of alignments and number of partitions for each dataset. **A:** Venn diagram showing number of partitions unique to each dataset, and common between them. The number of partitions recovered through the phylogenomic analysis pipeline is fivefold higher when the dataset is made up of high-quality transcripts compared to lower-quality ones. **B:** Density plot of alignment lengths of each partition before filtering with *Gblocks*. **C:** Density plot of alignment lengths of each partition after filtering with *Gblocks*. While the lengths of the individual alignments are significantly different before *Gblocks* filtering, they are similar afterwards.

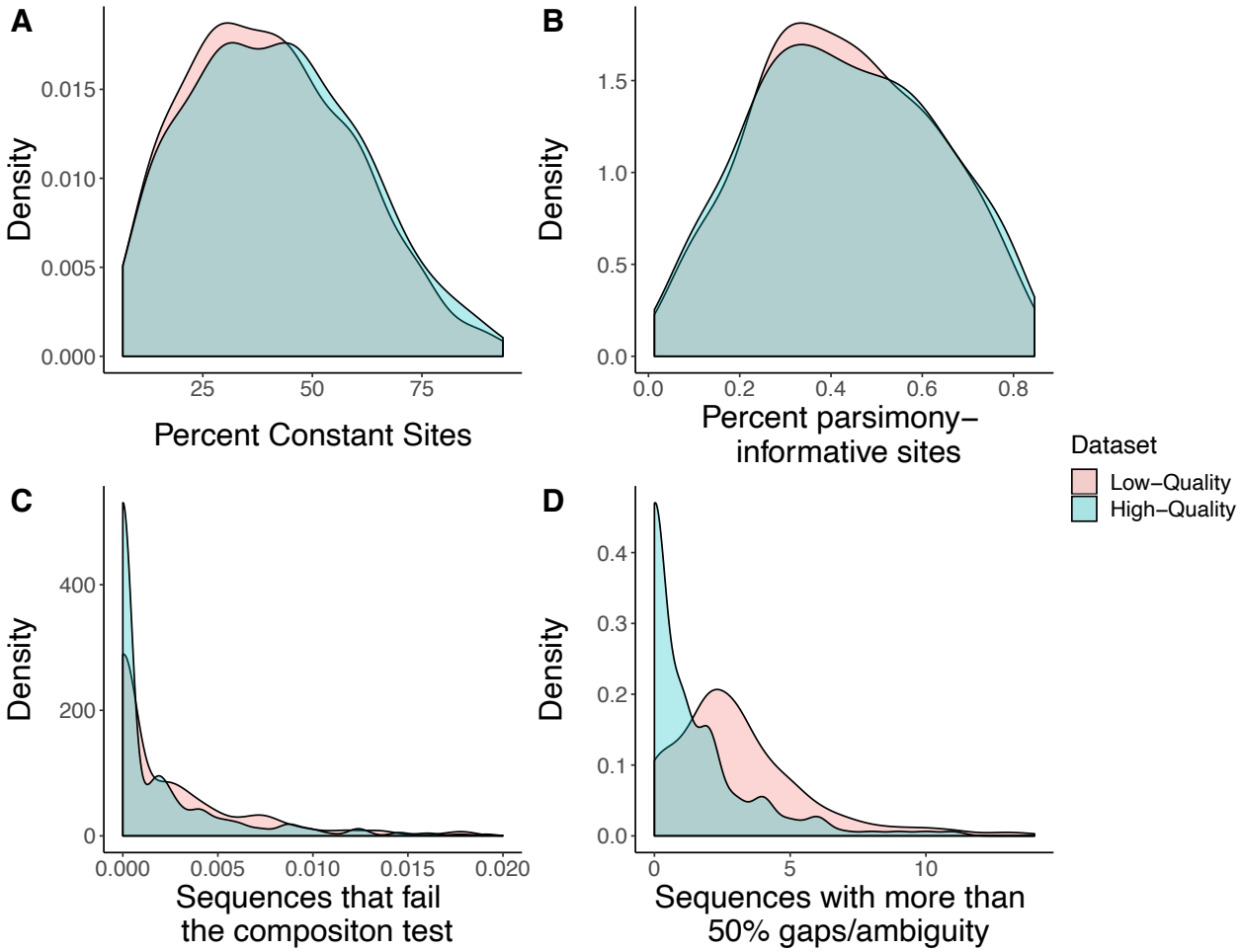

Figure 4: Density plots of four alignment metrics for both datasets. Alignments created from low-quality transcriptome assemblies have similar percentages of constant and parsimony-informative sites, but higher compositional bias and ambiguity when compared to alignments from high-quality assemblies. **A:** Percentage of constant sites in each partition alignment. **B:** Percentage of parsimony-informative sites in each partition alignment. **C:** Number of sequences that fail the composition test, normalized by partition alignment length. **D:** Number of sequences that contain more than 50% gaps/ambiguity in each partition alignment.

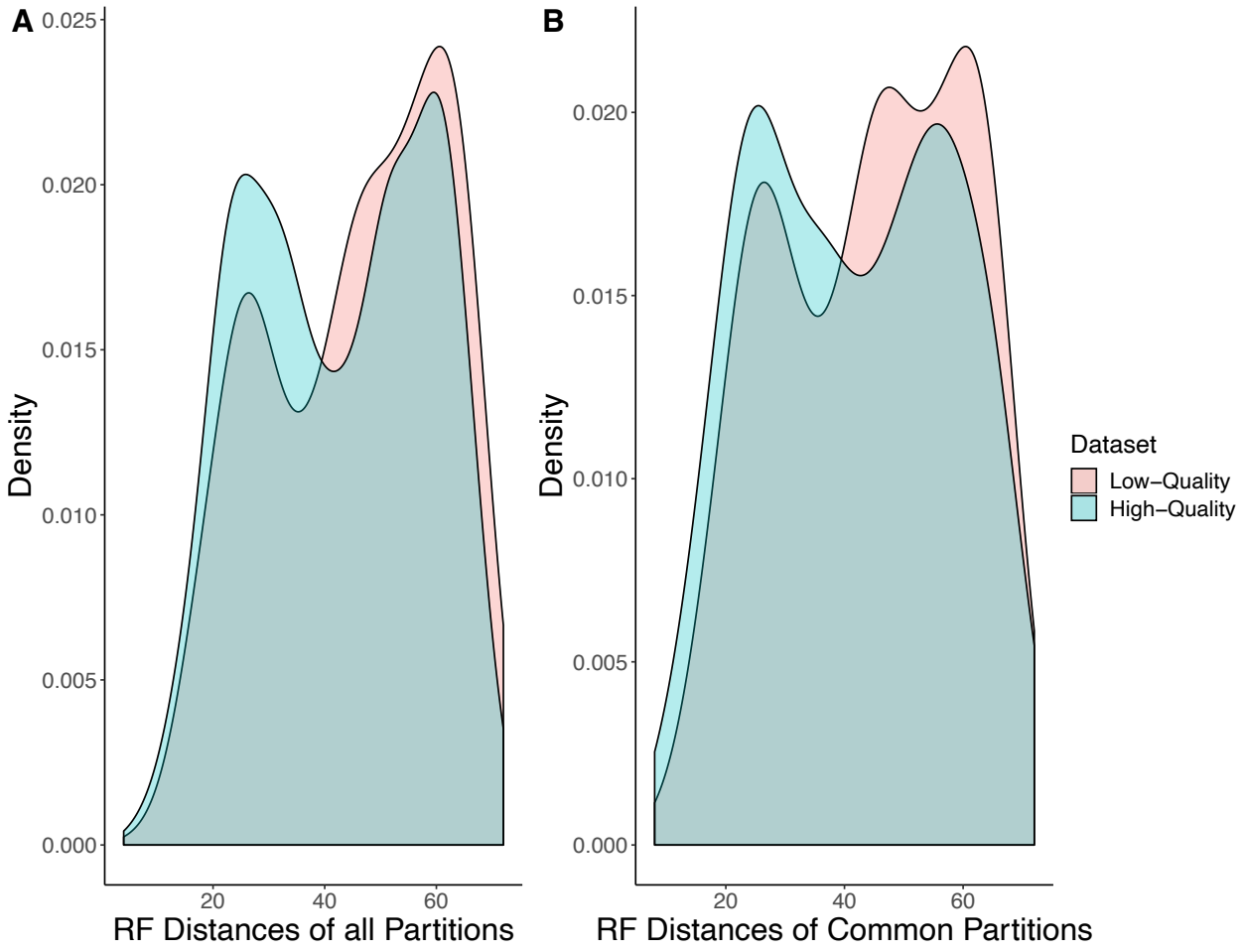

Figure 5: Per partition Robinson-Foulds (RF) distances to the constraint tree are significantly shorter in the high-quality dataset compared with the low-quality dataset. **A:** Density plot for all partitions from both datasets. **B:** Density plot for only those 332 partitions that are shared between the two datasets

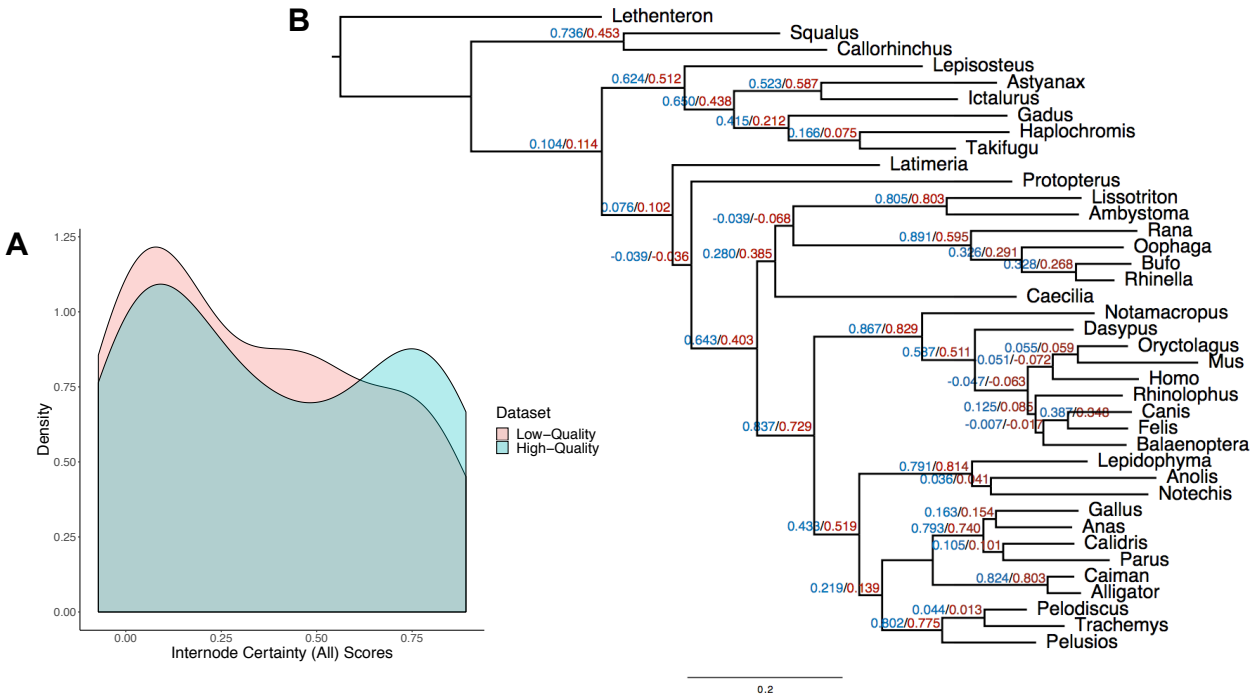

Figure 6: Partitions derived from the high-quality dataset have higher internode certainty all (ICA) values than those derived from the low-quality dataset when compared to the constraint tree. **A:** Density plot of ICA values **B:** Average ICA values for each node. Blue represents the high-quality dataset, red represents the low-quality dataset. Negative ICA values suggest that the node conflicts with at least one other node that has a higher support.

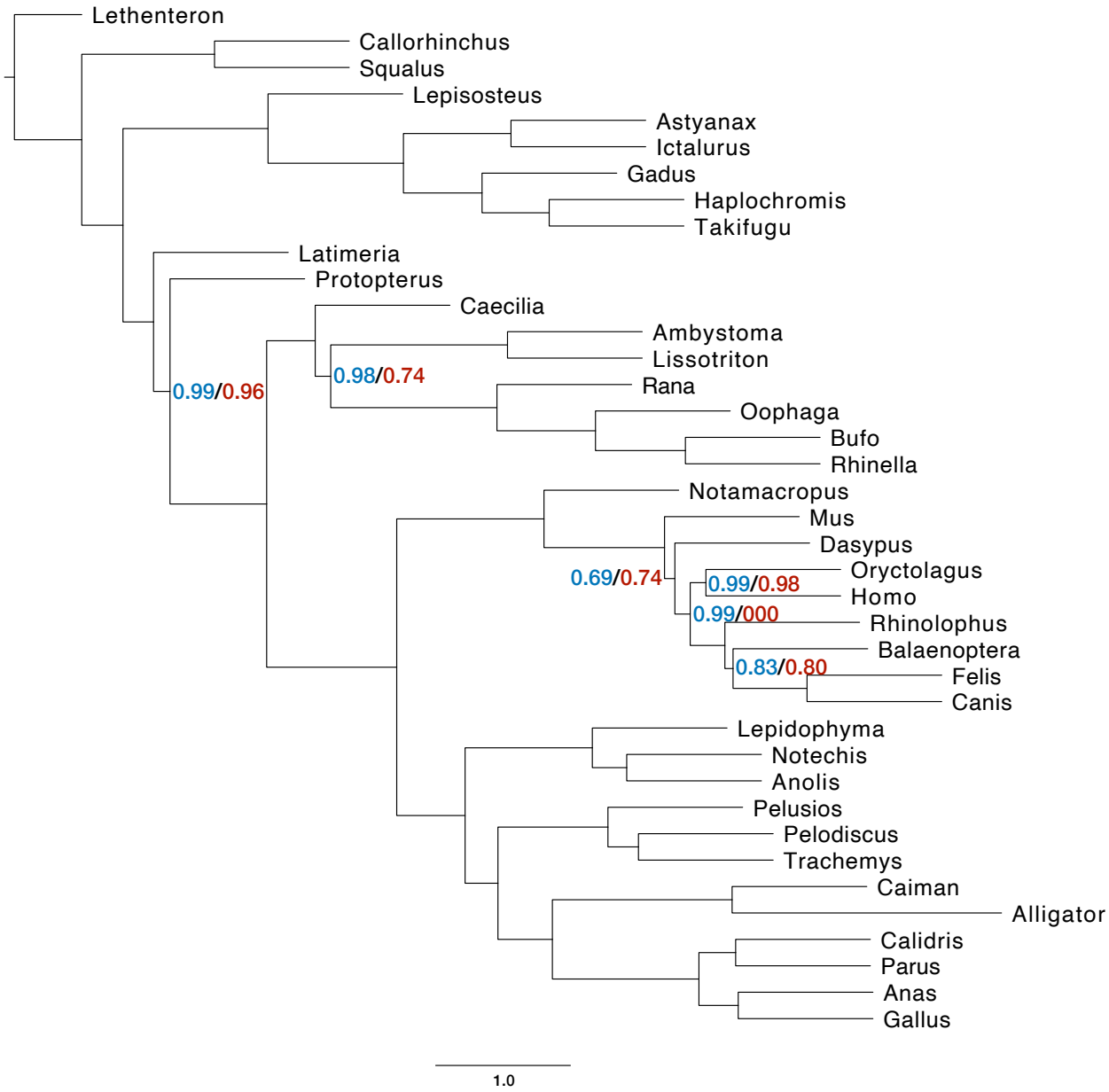

Figure 7: Species tree analysis in *ASTRAL* reveals a similar pattern to concatenation analyses. *ASTRAL* analyses of gene trees from 332 shared partitions from the high- and low-quality datasets result in identical topologies. In addition to normalized quartet scores being higher for gene trees derived from the high-quality dataset, node support values for the high-quality dataset are marginally stronger than those from the low-quality dataset. Support values represent support for quadripartitions of the tree, and only those that were less than 1 are represented.

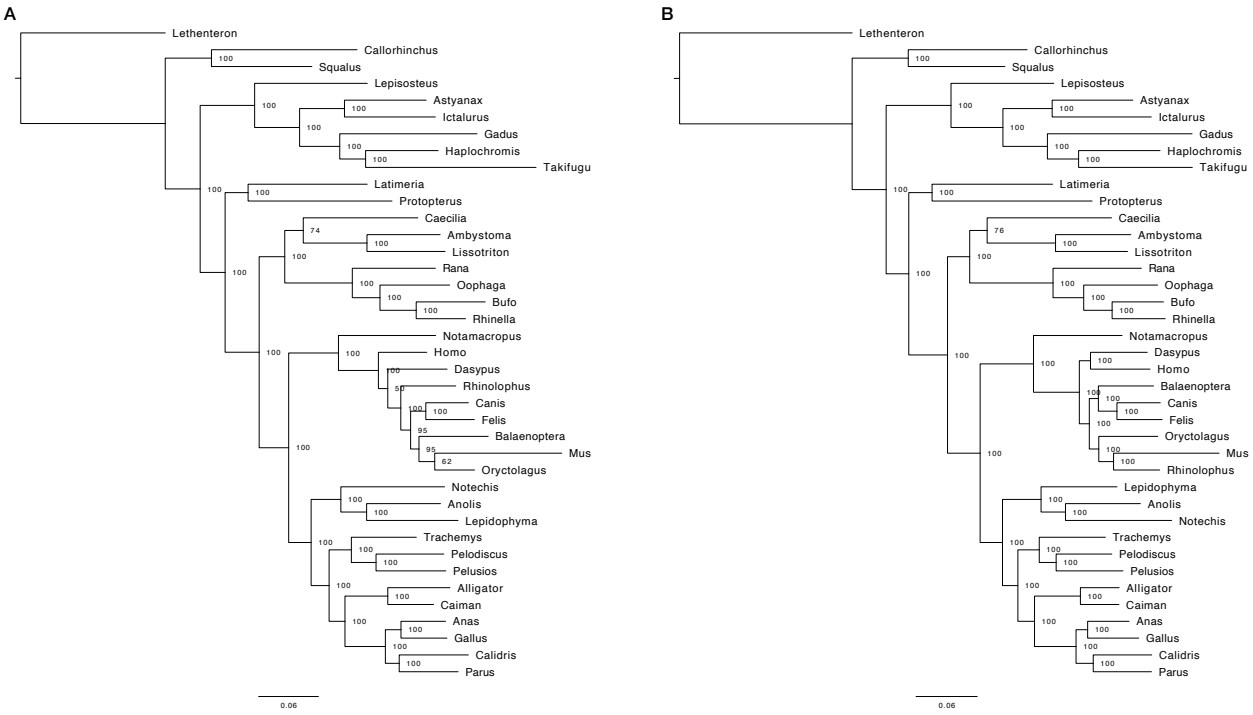

Figure S1: Phylogenetic trees created using the 332 data partitions shared between the two datasets and concatenation methods do not resolve the accepted craniate phylogeny but produce differing topologies. The trees were built in *IQ-TREE* using an LG model and nodes are labeled with ultrafast bootstrap approximated branch supports using the "-bnni" (a hill-climbing nearest neighbor interchange search) to reduce the impact of severe model violations. **A:** Phylogenetic tree for the low-quality dataset. **B:** Phylogenetic tree for the high-quality dataset.

Table S1: Accession numbers and associated studies of RNA-seq read sets used in these analyses

Species	Accession	Reference
Alligator mississippiensis	SRR629636	McGaugh SE, Bronikowski AM, Kuo CH, Reding DM, Addis EA, Flagel LE, Janzen FJ, Schwartz TS. Rapid molecular evolution across amniotes of the IIS/TOR network. Proceedings of the National Academy of Sciences. 2015 Jun 2;112(22):7055-60. http://dx.doi.org/10.1073/pnas.1419659112
Ambystoma mexicanum	SRR5341572	Nowoshilow S, Schloissnig S, Fei JF, Dahl A, Pang AW, Pippel M, Winkler S, Hastie AR, Young G, Roscito JG, Falcon F. The axolotl genome and the evolution of key tissue formation regulators. Nature. 2018 Feb;554(7690):50-5. http://dx.doi.org/10.1038/nature25458
Anas platyrhynchos	SRR7127376	Hérault F, Houée-Bigot M, Baéza E, Bouchez O, Esquerré D, Klopp C, Diot C. RNA-seq analysis of hepatic gene expression of common Pekin, Muscovy, mule and hinny ducks fed ad libitum or overfed. BMC genomics. 2019 Dec;20(1):1-4. http://dx.doi.org/10.1186/s12864-018-5415-1
Anolis carolinensis	SRR391653	Eckalbar WL, Hutchins ED, Markov GJ, Allen AN, Corneveaux JJ, Lindblad-Toh K, Di Palma F, Alföldi J, Huentelman MJ, Kusumi K. Genome reannotation of the lizard Anolis carolinensis based on 14 adult and embryonic deep transcriptomes. BMC genomics. 2013 Dec 1;14(1):49. http://dx.doi.org/10.1186/1471-2164-14-49
Astyanax mexicanus	SRR2045431	Pasquier J, Cabau C, Nguyen T, Jouanno E, Severac D, Braasch I, Journot L, Pontarotti P, Klopp C, Postlethwait JH, Guiguen Y. Gene evolution and gene expression after whole genome duplication in fish: the PhyloFish database. BMC genomics. 2016 Dec;17(1):1-0. http://dx.doi.org/10.1186/s12864-016-2709-z
Balaenoptera acutotostrata	SRR919296	Yim HS, Cho YS, Guang X, Kang SG, Jeong JY, Cha SS, Oh HM, Lee JH, Yang EC, Kwon KK, Kim YJ. Minke whale genome and aquatic adaptation in cetaceans. Nature genetics. 2014 Jan;46(1):88-92. http://dx.doi.org/10.1038/ng.2835
Bufo bufo	ERR1331718	Jin L, Yu JP, Yang ZJ, Merilä J, Liao WB. Modulation of gene expression in liver of hibernating Asiatic Toads (Bufo gargarizans). International journal of molecular sciences. 2018 Aug;19(8):2363. http://dx.doi.org/10.3390/ijms19082363
Caecilia tentaculata	SRR5591453	Torres-Sánchez M, Creevey CJ, Kornobis E, Gower DJ, Wilkinson M, San Mauro D. Multi-tissue transcriptomes of caecilian amphibians highlight incomplete knowledge of vertebrate gene families. DNA Research. 2019 Feb 1;26(1):13-20. http://dx.doi.org/10.1093/dnares/dsy034

Caiman crocodilus	ERR2198478	No associated article. Study accession: PRJEB21261
Calidris pugnax	ERR1018151	Küpper C, Stocks M, Risse JE, Dos Remedios N, Farrell LL, McRae SB, Morgan TC, Karlionova N, Pinchuk P, Verkuil YI, Kitaysky AS. A supergene determines highly divergent male reproductive morphs in the ruff. Nature genetics. 2016 Jan;48(1):79-83. http://dx.doi.org/10.1038/ng.3443
Callorhinchus milii	SRR513760	Venkatesh B, Lee AP, Ravi V, Maurya AK, Lian MM, Swann JB, Ohta Y, Flajnik MF, Sutoh Y, Kasahara M, Hoon S. Elephant shark genome provides unique insights into gnathostome evolution. Nature. 2014 Jan;505(7482):174-9. http://dx.doi.org/10.1038/nature12826
Canis lupus familiaris	ERR1331673	Berthelot C, Villar D, Horvath JE, Odom DT, Flicek P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. Nature ecology & evolution. 2018 Jan;2(1):152-63. http://dx.doi.org/10.1038/s41559-017-0377-2
Dasypus novemcinctus	SRR494766	No associated article. Study accession: PRJNA163137
Felis catus	ERR1331679	Berthelot C, Villar D, Horvath JE, Odom DT, Flicek P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. Nature ecology & evolution. 2018 Jan;2(1):152-63. http://dx.doi.org/10.1038/s41559-017-0377-2
Gadhus morhua	SRR2045420	Pasquier J, Cabau C, Nguyen T, Jouanno E, Severac D, Braasch I, Journot L, Pontarotti P, Klopp C, Postlethwait JH, Guiguen Y. Gene evolution and gene expression after whole genome duplication in fish: the PhyloFish database. BMC genomics. 2016 Dec;17(1):1-0. http://dx.doi.org/10.1186/s12864-016-2709-z
Gallus gallus	ERR1298598	Kuo RI, Tseng E, Eory L, Paton IR, Archibald AL, Burt DW. Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. BMC genomics. 2017 Dec 1;18(1):323. http://dx.doi.org/10.1186/s12864-017-3691-9
Haplochromis burtoni	SRR387451	Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S, Simakov O, Ng AY, Lim ZW, Bezault E, Turner-Maier J. The genomic substrate for adaptive radiation in African cichlid fish. Nature. 2014 Sep;513(7518):375-81. http://dx.doi.org/10.1038/nature13726
Homo sapiens	SRR5576267	Kim DS, Ryu JW, Son MY, Oh JH, Chung KS, Lee S, Lee JJ, Ahn JH, Min JS, Ahn J, Kang HM. A liver-specific gene expression panel predicts the differentiation status of in vitro hepatocyte models. Hepatology. 2017 Nov;66(5):1662-74. http://dx.doi.org/10.1002/hep.29324

Ictalurus punctatus	SRR917955	Liu S, Wang X, Sun F, Zhang J, Feng J, Liu H, Rajendran KV, Sun L, Zhang Y, Jiang Y, Peatman E. RNA-Seq reveals expression signatures of genes involved in oxygen transport, protein synthesis, folding, and degradation in response to heat stress in catfish. Physiological genomics. 2013 Jun 15;45(12):462-76. http://dx.doi.org/10.1152/physiolgenomics.00026.2013
Latimeria menadoensis	SRR576100	Pallavicini A, Canapa A, Barucca M, Alföldi J, Biscotti MA, Buonocore F, De Moro G, Di Palma F, Fausto AM, Forconi M, Gerdol M. Analysis of the transcriptome of the Indonesian coelacanth Latimeria menadoensis. BMC genomics. 2013 Dec 1;14(1):538. http://dx.doi.org/10.1186/1471-2164-14-538
Lepidophyma flavimaculatum	DRR034613	No associated article. Study accession: PRJDB3883
Lepisosteus oculatus	SRR1287992	No associated article. Study accession: PRJNA247500
Lethenteron camtschaticum	SRR3223459	Du K, Zhong Z, Fang C, Dai W, Shen Y, Gan X, He S. Ancient duplications and functional divergence in the interferon regulatory factors of vertebrates provide insights into the evolution of vertebrate immune systems. Developmental & Comparative Immunology. 2018 Apr 1;81:324-33. http://dx.doi.org/10.1016/j.dci.2017.12.016
Lissotriton montandoni	SRR3299753	Stuglik MT, Babik W. Genomic heterogeneity of historical gene flow between two species of newts inferred from transcriptome data. Ecology and evolution. 2016 Jul;6(13):4513-25. http://dx.doi.org/10.1002/ece3.2152
Notamacropus eugenii	DRR013408, DRR013409, DRR013410	Deakin JE. Genome Sequence of an Australian Kangaroo, Macropus eugenii. eLS. 2013. http://dx.doi.org/10.1186/gb-2011-12-8-r81
Notechis scutatus	SRR519122	No associated article. Study accession: PRJNA170152
Oophaga sylvatica	SRR9120851	Caty SN, Alvarez-Buylla A, Byrd GD, Vidoudez C, Roland AB, Tapia EE, Budnik B, Trauger SA, Coloma LA, O'Connell LA. Molecular physiology of chemical defenses in a poison frog. Journal of Experimental Biology. 2019 Jun 15;222(12):jeb204149. http://dx.doi.org/10.1242/jeb.204149
Oryctolagus cuniculus	ERR1331669	Berthelot C, Villar D, Horvath JE, Odom DT, Flicek P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. Nature ecology & evolution. 2018 Jan;2(1):152-63. http://dx.doi.org/10.1038/s41559-017-0377-2
Parus major	SRR1847228	Charmantier A, Gienapp P. Climate change and timing of avian breeding and migration: evolutionary versus plastic changes. Evolutionary Applications. 2014 Jan;7(1):15-28. http://dx.doi.org/10.1111/eva.12126
Pelodiscus sinensis	SRR6157006	Zeng D, Li X, Wang XQ, Xiong G. Development of SNP markers associated with growth-related genes of

		Pelodiscus sinensis. Conservation Genetics Resources. 2020 Mar;12(1):87-92. http://dx.doi.org/10.1007/s12686-018-1065-5
Pelusios castaneus	SRR629649	McGaugh SE, Bronikowski AM, Kuo CH, Reding DM, Addis EA, Flagel LE, Janzen FJ, Schwartz TS. Rapid molecular evolution across amniotes of the IIS/TOR network. Proceedings of the National Academy of Sciences. 2015 Jun 2;112(22):7055-60. http://dx.doi.org/10.1073/pnas.1419659112
Protopterus sp.	ERR2202465	Chana-Muñoz A, Jendroszek A, Sønnichsen M, Wang T, Ploug M, Jensen JK, Andreasen PA, Bendixen C, Panitz F. Origin and diversification of the plasminogen activation system among chordates. BMC evolutionary biology. 2019 Dec 1;19(1):27. http://dx.doi.org/10.1186/s12862-019-1353-z
Rana pipiens	SRR1185245	Christenson MK, Trease AJ, Potluri LP, Jezewski AJ, Davis VM, Knight LA, Kolok AS, Davis PH. De novo assembly and analysis of the northern leopard frog Rana pipiens transcriptome. Journal of genomics. 2014;2:141. http://dx.doi.org/10.7150/jgen.9760
Rhinella marina	SRR6311453	Russo AG, Eden JS, Tuipulotu DE, Shi M, Selechnik D, Shine R, Rollins LA, Holmes EC, White PA. Viral discovery in the invasive Australian cane toad (Rhinella marina) using metatranscriptomic and genomic approaches. Journal of virology. 2018 Sep 1;92(17). http://dx.doi.org/10.1128/JVI.00768-18
Rhinolophus sinicus	SRR2273875	Dong D, Lei M, Hua P, Pan YH, Mu S, Zheng G, Pang E, Lin K, Zhang S. The genomes of two bat species with long constant frequency echolocation calls. Molecular Biology and Evolution. 2016 Oct 26:msw231. http://dx.doi.org/10.1093/molbev/msw231
Squalus acanthias	ERR1525379	Chana-Munoz A, Jendroszek A, Sønnichsen M, Kristiansen R, Jensen JK, Andreasen PA, Bendixen C, Panitz F. Multi-tissue RNA-seq and transcriptome characterisation of the spiny dogfish shark (Squalus acanthias) provides a molecular tool for biological research and reveals new genes involved in osmoregulation. PloS one. 2017 Aug 23;12(8):e0182756. http://dx.doi.org/10.1371/journal.pone.0182756
Takifugu rubripes	SRR1005688	No associated article. Study accession: PRJNA222262
Trachemys scripta	ERR2198830	Chana-Muñoz A, Jendroszek A, Sønnichsen M, Wang T, Ploug M, Jensen JK, Andreasen PA, Bendixen C, Panitz F. Origin and diversification of the plasminogen activation system among chordates. BMC evolutionary biology. 2019 Dec 1;19(1):27. http://dx.doi.org/10.1186/s12862-019-1353-z

References

1. Dopazo H, Santoyo J, Dopazo J. Phylogenomics and the number of characters required for obtaining an accurate phylogeny of eukaryote model species. Bioinformatics. 2004;20:116–21.

2. Blair JE, Ikeo K, Gojobori T, Hedges SB. The evolutionary position of nematodes. BMC Evol Biol. 2002;2(7):1–7.

3. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, et al. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature. 2008;452(7188):745–9.

4. Vijay N, Poelstra JW, Kunstner A, Wolf JBW. Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. Mol Ecol. 2013;22:620–34.

5. Cheon S, Zhang J, Park C. Is phylotranscriptomics as reliable as phylogenomics? Mol Biol Evol. 2020;

6. Chen X, Zhao X, Liu X, Warren A, Zhao F, Miao M. Phylogenomics of non-model ciliates based on transcriptomic analyses. Protein Cell [Internet]. 2015;6(5):373–85. Available from: http://dx.doi.org/10.1007/s13238-015-0147-3

7. Reich A, Dunn C, Akasaka K, Wessel G. Phylogenomic Analyses of Echinodermata Support the Sister Groups of Asterozoa and Echinozoa. PLoS One. 2015;1–11.

8. Kutty SN, Wong WH, Meusemann K, Meier R, Cranston PS. A phylogenomic analysis of Culicomorpha (Diptera) resolves the relationships among the eight constituent families. Syst Entomol. 2018;(March):1–14.

9. Washburn JD, Schnable JC, Conant GC, Brutnell TP, Shao Y, Zhang Y, et al. Genome-Guided Phylo-Transcriptomic Methods and the Nuclear Phylogentic Tree of the Paniceae Grasses. Sci Rep [Internet]. 2017;7(1):1–12. Available from: http://dx.doi.org/10.1038/s41598-017-13236-z

10. Yang Y, Smith SA. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: Improving accuracy and matrix occupancy for phylogenomics. Mol Biol Evol. 2014;31(11):3081–92.

11. Mckain MR, Johnson MG, Urive-Convers S, Eaton D, Yang Y. Practical considerations for plant phylogenomics. Appl Plant Sci. 2018;6(3):1–15.

12. Yu X, Yang D, Guo C, Gao L. Plant phylogenomics based on genome-partitioning strategies: Progress and prospects. Plant Divers [Internet]. 2018;40(4):158–64. Available from: https://doi.org/10.1016/j.pld.2018.06.005

13. Wen J, Egan AN, Dikow RB, Zimmer EA. Utility of transcriptome sequencing for phylogenetic inference and character evolution. In: Next-Generation Sequencing in Plant

Systematics. 2015. p. 1–42.

14. Whelan N V., Kocot KM, Moroz LL, Halanych KM. Error, signal, and the placement of Ctenophora sister to all other animals. Proc Natl Acad Sci [Internet]. 2015;112(18):5773–8. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1503453112

15. Blanquart S, Lartillot N. A site- and time-heterogeneous model of amino acid replacement. Mol Biol Evol. 2008;25(5):842–58.

16. Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol Biol. 2014;14(82):1–14.

17. Philippe H, Delsuc F, Brinkmann H, Lartillot N. Phylogenomics. Annu Rev Ecol Evol Syst. 2005;36:541–62.

18. Feuda R, Dohrmann M, Pett W, Philippe H, Rota-Stabelli O, Lartillot N, et al. Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals. Curr Biol. 2017;27(24):3864-3870.e4.

19. Wang HC, Minh BQ, Susko E, Roger AJ. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. Syst Biol. 2018;67(2):216–35.

20. Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics [Internet]. 2018;19(153):15–30. Available from: http://dx.doi.org/10.1186/s12859-018-2129-y

21. Liu L, Yu L, Edwards S V. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol Biol. 2010;10(302):25–7.

22. Borowiec ML, Lee EK, Chiu JC, Plachetzki DC. Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. BMC Genomics [Internet]. 2015;16(2015):987. Available from: http://dx.doi.org/10.1186/s12864-015-2146-4

23. Simion P, Phillippe H, Baurain D, Jager M, Richter DJ, Di Franco A, et al. A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. Curr Biol. 2017;27:1–10.

24. Masta SE, Longhorn SJ, Boore JL. Arachnid relationships based on mitochondrial genomes: Asymmetric nucleotide and amino acid bias affects phylogenetic analyses. Mol Phylogenet Evol [Internet]. 2008;50(1):117–28. Available from: http://dx.doi.org/10.1016/j.ympev.2008.10.010

25. Lasek-Nesselquist E. A Mitogenomic Re-Evaluation of the Bdelloid Phylogeny and Relationships among the Syndermata. PLoS One. 2012;7(8):1–11.

26. Marletaz F, Peijnenburg KT, Goto T, Satoh N, Rokhsar DS. A New Spiralian Phylogeny Places the Enigmatic Arrow Worms among Gnathiferans. Curr Biol. 2019;29:312–8.

27. Hernandez AM, Ryan JF. Six-state amino acid recoding is not an effective strategy to offset the effects of compositional heterogeneity and saturation in phylogenetic analyses. bioRxiv. 2019;(Table 1):1–16.

28. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63.

29. MacManes MD. On the optimal trimming of high-throughput mRNA sequence data. Front Genet. 2014;1–7.

30. Mbandi SK, Hesse U, Rees DJG, Christoffels A. A glance at quality score: implication for de novo transcriptome reconstruction of Illumina reads. Front Genet. 2014;5:1–5.

31. MacManes MD, Eisen MB. Improving transcriptome assembly through error correction of high-throughput sequence reads. PeerJ. 2013;1(e113):1–15.

32. Song L, Florea L. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. Giga Sci. 2015;4(48):1–8.

33. Le H, Schulz MH, Mccauley BM, Hinman VF, Bar-Joseph Z. Probabilistic error correction for RNA sequencing. Nucleic Acids Res. 2013;41(10):1–11.

34. MacManes MD. The Oyster River Protocol : a multi-assembler and kmer approach for de novo transcriptome assembly. PeerJ. 2018;6(e5428):1–18.

35. Li B, Dewey CN. RSEM : accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12(323):1–16.

36. Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, et al. Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biol. 2014;15(553):1–21.

37. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. TransRate: reference free quality assessment of de-novo transcriptome assemblies. Genome Res. 2016;26.

38. Parks MB, Wickett NJ, Alverson AJ. Signal, Uncertainty, and Conflict in Phylogenomic Data for a Diverse Lineage of Microbial Eukaryotes (Diatoms, Bacillariophyta). Mol Biol Evol. 2017;35(1):80–93.

39. Karmeinski D, Meusemann K, Goodheart JA, Schroedi M, Martynov A, Korshunova T, et al. Transcriptomics provides a robust framework for the relationships of the mojor clades of cladobranch sea slugs (Mollusca, Gastropoda, Heterobranchia), but fails to resolve the position of the enigmatic genus Embletonia. bioRxiv. 2020;

40. Yang Y, Smith SA. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. BMC Genomics. 2013;14(328):1–11.

41. Dunn CW, Howison M, Zapata F. Agalma: An automated phylogenomics workflow. BMC Bioinformatics. 2013;14(1).

42. Nguyen L, Schmidt HA, Haeseler A Von, Minh BQ. IQ-TREE : A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol. 2014;32(1):268–74.

43. Irisarri I, Baurain D, Brinkmann H, Delsuc F, Sire J, Kupfer A, et al. Phylotranscriptomic consolidation of the jawed vertebrate timetree. Nat Ecol Evol. 2017;1(9):1370–8.

44. Chen M-Y, Liang D, Zhang P. Phylogenomic resolution of the phylogeny of laurasiatherian mammals: Exploring phylogenetic signals within coding and noncoding sequences. Genome Biol Evol. 2017;9(8):1998–2012.

45. Robinson DF, Foulds LR. Comparison of phylogenetic trees. Math Biosci. 1981;(53):131–41.

46. Salichos L, Stamatakis A, Rokas A. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol Biol Evol. 2014;31(5):1261–71.

47. Mirarab S, Reaz R, Bayzid MS, Zimmermann T, S. Swenson M, Warnow T. ASTRAL: Genome-scale coalescent-based species tree estimation. Bioinformatics. 2014;30(17):i541–8.

48. Emms DM, Kelly S. OrthoFinder : solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol [Internet]. 2015;16(157):1–14. Available from: http://dx.doi.org/10.1186/s13059-015-0721-2

49. Emms DM, Kelly S. OrthoFinder : phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(238):1–14.

50. Venkatesh B, Lee AP, Ravi V, Maurya AK, Lian MM, Swann JB, et al. Elephant shark genome provides unique insights into gnathostome evolution. Nature. 2014;505(7482):174–9.

51. Puig Giribets M, Pilar García Guerreiro M, Santos M, Ayala FJ, Tarrío R, Rodríguez-Trelles F. Chromosomal inversions promote genomic islands of concerted evolution of Hsp70 genes in the Drosophila subobscura species subgroup. Mol Ecol. 2019;28(6):1316–32.

52. Foster PG, Hickey DA. Compositional Bias May Affect Both DNA-Based and Protein-Based Phylogenetic Reconstructions. J Mol Evol. 1999;48:284–90.

53. Revell LJ, Harmon LJ, Collar DC. Phylogenetic signal, evolutionary process, and rate. Syst Biol. 2008;57(4):591–601.

54. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.

55. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript reconstruction from RNA-Seq: reference generation and analysis with Trinity.

Nat Protoc. 2013;8(8):1–43.

56. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, et al. De novo assembly and analysis of RNA-seq data. Nat Methods. 2010;7(11):909–12.

57. Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. Giga Sci. 2019;8:1–13.

58. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658–9.

59. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.

60. Haas BJ, Papanicolaou A. TransDecoder [Internet]. 2018. Available from: https://github.com/TransDecoder/TransDecoder/wiki

61. Howe KL, Contreras-moreira B, De Silva N, Maslen G, Akanni W, Allen J, et al. Ensembl Genomes 2020 — enabling non-vertebrate genomic research. Nucleic Acids Res. 2020;48:689–95.

62. R Core Team. R: a language and environment for statistical computing [Internet]. Vienna, Austria; 2018. Available from: https://www.r-project.org/

63. Kocot KM, Citarella MR, Moroz LL, Halanych KM. PhyloTreePruner: A phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. Evol Bioinforma. 2013;2013(9):429–35.

64. Katoh K, Toh H. Parallelization of the MAFFT multiple sequence alignment program. Bioinformatics. 2010;26(15):1899–900.

65. Castresana J. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Mol Biol Evol. 2000;17(4):540–52.

66. Talavera G, Castresana J. Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. Syst Biol. 2007;56(4):564–77.

67. Dunn C, Smith S, Ryan J. Gblockswrapper [Internet]. Bitbucket; 2009. Available from: https://bitbucket.org/caseywdunn/labcode/src/master/scripts_phylogenomics_21Feb2009/Gblockswrapper

68. Jones P, Binns D, Chang H, Fraser M, Li W, Mcanulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236–40.

69. Alexa A, Rahnenfuhrer J. Gene set enrichment analysis with topGO. Bioconductor Improv. 2009;27.

70. Maddison WP, Maddison DR. Mesquite: a modular system for evolutionary analysis [Internet]. 2018. Available from: http://www.mesquiteproject.org

71. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27(4):592–3.

72. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.

73. Spillane JL, LaPolice TM, MacManes MD, Plachetzki DC. High- and low-quality assemblies for 38 craniate species [Internet]. 2020 [cited 2020 Jul 14]. Available from: https://doi.org/10.5281/zenodo.3939160

74. Spillane JL. Repository for analysis of high- and low-quality transcriptome assemblies [Internet]. 2019 [cited 2020 Jul 28]. Available from: http://github.com/jls943/quality_review

CHAPTER 2

The first genome assembly of a cerianthid, *Pachycerianthus borealis*

Abstract

While there are many established model organisms within Cnidaria, there are still entire clades of organisms that are not represented in scientific studies due to the difficulty in sampling them or cryptic species and subspecies. Ceriantharia holds a unique position within Cnidaria, as the sister group to the remaining hexacorals according to the most recent phylogenomics analyses. Up to this point however, the data available for cerianthids has been either transcriptomic, or from a small subset of genes. Here we report the draft genome from a cerianthid species, *Pachycerianthus borealis*. We used a combination of long and short-read sequencing technologies to produce a highly contiguous genome assembly that is 492 Mb in length and has a scaffold N50 of 396 kb. The assembly has a high level of completeness as measured by BUSCO score, and its predicted proteins are placed into orthogroups at comparable rates to other cnidarian genomes. This new cerianthid genome will provide a resource to investigate questions about the evolutionary history of unique traits, gene families, and the phylogenomic

distribution and ancestral state of mitochondrial genome structure within cnidarians, among others.

Context

The genomics revolution has drastically expanded the number of genome-scale datasets that are publicly available to researchers. However, this expansion has not been evenly distributed across all taxa. Marine invertebrates, which include the vast majority of animal life on earth, remain underrepresented in these critical genomic resources. Even within phyla that have many sequenced genomes and transcriptomes, such as Cnidaria, there are whole clades that have thus far been overlooked but that merit a closer study.

The class Anthozoa (Phylum Cnidaria) is further divided into three subclasses: Hexacorallia, Octocorallia, and Ceriantharia. Hexacorallia contains many familiar and ecologically crucial species, such as stony corals and sea anemones, and genomic resources for these clades continue to grow rapidly. However the other two subclasses remain under-represented in scientific studies and in publicly available genome-scale resources. Octocorallia is comprised of sea pens, sea fans, and soft corals, and while recent work has sought to add genomes and transcriptomes to the smaller datasets already available (1,2), it remains far less represented than hexacorals. Currently, a whole genome sequence from any member of the anthozoan subclass Ceriantharia is lacking, and studies of these organisms and of Anthozoa are hindered by this exception.

Cerianthids are tube-dwelling anemones that possess a host of unique traits that set them apart from other cnidarians. They possess a unique cnidocyte called a ptychocyte that lacks spines along its tubule, and is folded (instead of coiled) inside its capsule (3). Cerianthids use these distinctive cnidae to help construct the tubes in which they live, though they use differing methods and materials in this construction (3). Minicollagen genes code for the structural casing that encloses the dynamic structure of all cnidocyte cells, as well as the tubules that the cells secrete. The number of distinct minicollagen genes present in a cnidarian is strongly correlated to the diversity of its cnidae (4). Since ptychocytes are characteristic only of cerianthids, they present a unique opportunity to study the expansion of the minicollagen gene family.

The phylogenetic position of Ceriantharia within Anthozoa remains uncertain. Recent studies leveraging evidence from a limited numbers of nuclear or mitochondrial markers have found conflicting results, placing Ceriantharia as the either sister group to Hexacorallia, the sister group to Octocorallia, or the sister to the remaining Anthozoa (5). Still others have concluded that Ceriantharia is not a monophyletic clade, instead having some of its members in the other two subclasses of Anthozoa (6). Resolving the phylogenetic position of Ceriantharia with certainty will require data from a much greater number of genomic loci, and is key to answering questions about the evolution of complex traits within Cnidaria.

In addition, questions related to the mitochondrial genome of cerianthids have captivated biologists. According to one previous study, cerianthids have an unusual mitochondrial chromosome structure (7) unlike that of any other anthozoan. While linear

mitochondrial chromosomes are the rule in medusazoans (8,9), they had never been observed in an anthozoan previous to this study, which found that the cerianthid mitochondrial genome was unusually large (~80,000 bp) and was contained in multiple linear chromosomes. However, the study was unable to determine the number of chromosomes definitively, or whether this structure is typical of all cerianthid mitochondrial genomes. Developing more robust genome-scale resources for this group will help to resolve these questions with more certainty.

Despite their phylogenetically important position and their singular ecology, cerianthids remain understudied. Four transcriptomes for the group have recently been released (10,11), however it persists as one of the only major lineages within Cnidaria without a full genome sequence. Here, we rectify this exception by releasing the first genome sequence for a member of Ceriantharia, *Pachycerianthus borealis*. This genome fills a critical gap in the genomic resources of Cnidaria. It will aid in the study of cnidocyte diversity and gene family evolution, Anthozoa phylogenetics, and mitochondrial genome structure evolution.

Methods

Sample collection, library preparation, and sequencing

We collected a single adult sample of *Pachycerianthus borealis* via SCUBA near Shoals Marine Laboratories, Appledore Island, Maine, USA in 2016. To obtain DNA from this individual we performed four separate DNA extractions using a Qiagen DNeasy Blood and Tissue Kit and followed the standard protocol with the exception that we used

higher centrifuge speeds (12,000 rpm) to ensure the samples flowed through the spin column completely. We then ran the samples through the Blue Pippin High-Pass Filtering with a 0.75% agarose gel cassette to remove DNA fragments less than 6 kb in length. We allowed the samples to remain in the collection well overnight to maximize yield of high molecular weight fragments. We constructed two libraries for the samples using the Oxford Nanopore Technologies (ONT) Genomic DNA by Ligation protocol (GDE_9063_v109_revA_23May2018) and sequenced the libraries on an ONT MinION (one FLO-MIN106 flow cell per library).

Sequence assembly, quality checks, and annotation

We performed a preliminary assembly of the resulting ONT reads in *Flye* version 2.3.5 (12) and found that this assembly was 544Mb long. We then assembled the same ONT reads using *wtdbg2* version 2.5 (13) using the default parameters, and estimating the genome size at 544Mb, based on the preliminary assembly. It was this second assembly that we used for the remainder of our analyses. We ran *QUAST* version 4.6.0 and Assemblathon_stats.pl (14,15) to assess genome size and contiguity. From previous Illumina sequencing (SRA Number) of the same individual (10) we obtained short, high accuracy reads, and these we used to polish the assembly using five iterations of *BWA* version 0.7.17-r1188 (16) and *Pilon* version 1.23 (17). We incorporated transcriptomic reads (SRR11802643) from Klompen et al. (11) for the same species into a sixth iteration of polishing using the same tools and settings. We used *SAMtools* version 1.10 (18) to measure the mapping rate of the Illumina reads to

the genome assembly, and *BUSCO* version 3.0.0 (19) with the metazoan database to gauge its genic completeness.

We assembled the *P. borealis* reads from Klompen et al. (11) into a transcriptome using the *Oyster River Protocol* version 2.2.3 (20). We then used this transcriptome, along with all *P. borealis* transcriptomic reads that we used for polishing the assembly to annotate the genome using *MAKER* version 3.01.02 (21,22). We also included all *P. borealis* transcriptomic reads that we used for polishing the assembly in the EST Evidence section of *MAKER*, as well as many other transcriptome and protein datasets from other members of Cnidaria in the Alt EST Evidence section (Table 1). And finally, we included the output of *RepeatModeler* version open-1.0.8 (23), which we ran on the assembly to identify transposable elements.

To compare the protein predictions of the *P. borealis* genome to other anthozoan genomes, we performed an orthogroup analysis in *OrthoFinder* version 2.3.3 (Emms and Kelly 2019).

Data Validation and Quality Control

We generated 3.5 million reads through ONT (SRR13639782) with an N50 of 7682 bp. The assembled genome (PRJNA699032) has a total length of 492 Mb, and a scaffold N50 of 396 kb. Of its 5833 scaffolds, 18.4% are larger than 100 kb, with 48 above 1 Mb. After six rounds of polishing the assembled genome with Illumina reads from the same species, 99.33% of these reads mapped to the genome, and through *MAKER*, we found 37,856 predicted proteins. Using the Metazoa database, we identified 87.6% complete

BUSCOs in the genome assembly, and 72.1% complete BUSCOs in the predicted proteins.

In orthogroup analysis, we found that *OrthoFinder* sorted 99.7% of *P. borealis* genes into shared orthogroups and species-specific orthogroups in similar proportions to other anthozoan genomes (Figure 1). This indicates that this genome contains recognizable orthogroups and performs at the same level in orthogroup analysis as publicly available genomic resources for Cnidaria.

Re-use Potential

Here we have sequenced the first genome of a cerianthid, *Pachycerianthus borealis*. We show that our hybrid sequencing and assembly strategy is effective for generating genomes of marine invertebrates and other organisms that are currently under-represented in genome-scale datasets. The *P. borealis* genome has contiguity and completeness comparable to other anthozoan genomes, and performs well in preliminary orthogroup analysis. The genome we present will be an asset to studies investigating the phylogenetics of Anthozoa, diverse mitochondrial genome evolution within Cnidaria, and novel gene evolution.

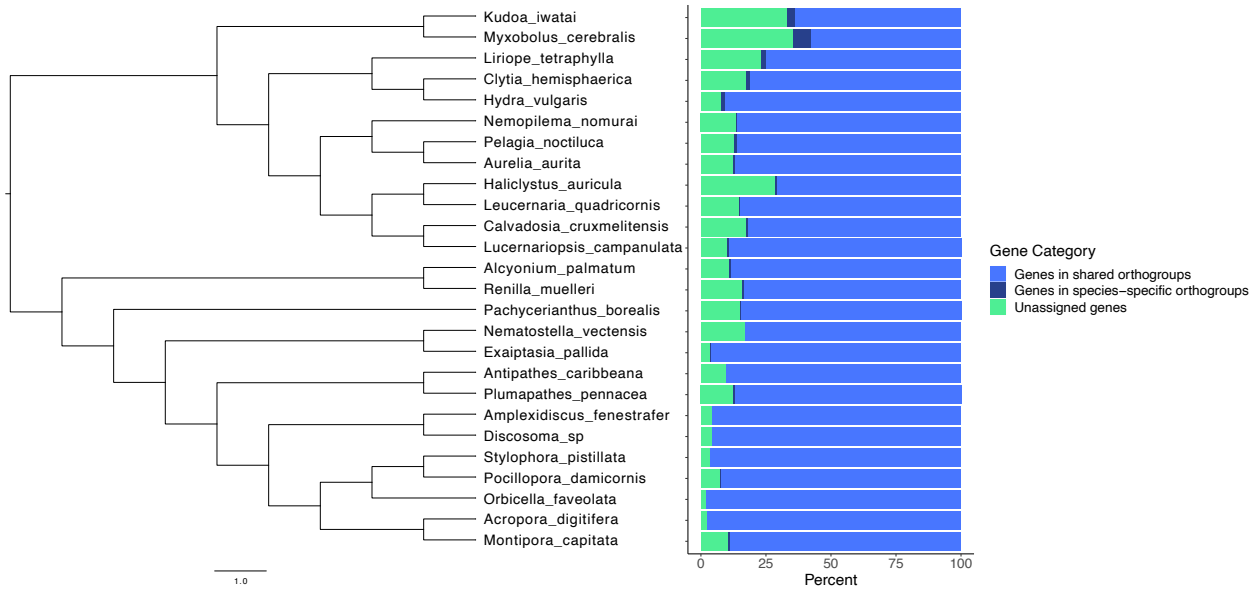

Figure 1: In orthology analysis, genes in the Pachycerianthus borealis genome are placed into orthogroups in similar proportions to other Cnidarian genomes of similar genic completeness. The tree shows Cnidarian genomes with at least 70% complete BUSCOs. The corresponding bars represent the proportion of genes from protein predictions from each genome that are placed into orthogroups with other species, orthogroups with only a single species, and unassigned genes.

Table 1: All datasets used during annotation of the genome, with accession numbers and associated references.

EST Evidence		
Species/Tissue	Accession	Reference
Body	SRR13639783-SRR13639786	Unpublished data
Hypostome	SRR13639783-SRR13639786	Unpublished data
Tentacle	SRR13639783-SRR13639786	Unpublished data
Pachycerianthus borealis	SRR11802643	Klompen, A. M., Macrander, J., Reitzel, A. M., & Stampar, S. N. (2020). Transcriptomic analysis of four cerianthid (Cnidaria, Ceriantharia) venoms. *Marine drugs*, *18*(8), 413.
Alt EST		
Species	Accession	Reference
Clytia hemisphaerica	SRR5814971	Artigas, G. Q., Lapébie, P., Leclère, L., Takeda, N., Deguchi, R., Jékely, G., ... & Houliston, E. (2018). A gonad-expressed opsin mediates light-induced spawning in the jellyfish Clytia. *Elife*, *7*, e29555.
Hydra vulgaris	HAEP_T-CDS_120217	Hemmrich, G., & Bosch, T. C. (2008). Compagen, a comparative genomics platform for early branching metazoan animals, reveals early origins of genes regulating stem-cell differentiation. *Bioessays*, *30*(10), 1010-1018.
Alatina alata	SRR1952741	Zapata, F., Goetz, F. E., Smith, S. A., Howison, M., Siebert, S., Church, S. H., ... & Cartwright, P. (2015). Phylogenomic analyses support traditional relationships within Cnidaria. *PloS one*, *10*(10), e0139068.
Liriope tetraphylla	SRR3407335	Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D. J., Di Franco, A., ... & Manuel, M. (2017). A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Current Biology*, *27*(7), 958-967.
Alcyonium palmatum	SRR3407216	Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D. J., Di Franco, A., ... & Manuel, M. (2017). A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Current Biology*, *27*(7), 958-967.
Lucernariopsis campanulata	SRR3407219	Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D. J., Di Franco, A., ... & Manuel, M. (2017). A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Current Biology*, *27*(7), 958-967.
Antipathes caribbeana	SRR3407160	Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D. J., Di Franco, A., ... & Manuel, M. (2017). A large and consistent phylogenomic dataset supports sponges

		as the sister group to all other animals. *Current Biology*, *27*(7), 958-967.
Myxobolus cerebralis	SRR1557039	Chang, E. S., Neuhof, M., Rubinstein, N. D., Diamant, A., Philippe, H., Huchon, D., & Cartwright, P. (2015). Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. *Proceedings of the National Academy of Sciences*, *112*(48), 14912-14917.
Chironex fleckeri	SRR1819888	Brinkman, D.L., Jia, X., Potriquet, J. *et al.* Transcriptome and venom proteome of the box jellyfish *Chironex fleckeri* . *BMC Genomics* **16,** 407 (2015). https://doi.org/10.1186/s12864-015-1568-3
Pelagia noctiluca	SRR3407257	Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D. J., Di Franco, A., ... & Manuel, M. (2017). A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Current Biology*, *27*(7), 958-967.
Corallium rubrum	SRR1552944	M. Pratlong, A. Haguenauer, O. Chabrol, C. Klopp, P. Pontarotti, et al.. The red coral (Coralliumrubrum) transcriptome: a new resource for population genetics and local adaptation studies. MolecularEcology Resources, Wiley/Blackwell, 2015, 15 (5), pp.1205–1215. 10.1111/1755-0998.12383. hal-01445149
Plumapathes pennacea	SRR3407161	Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D. J., Di Franco, A., ... & Manuel, M. (2017). A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Current Biology*, *27*(7), 958-967.
Hydractinia polyclina	SRR923509	Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D. J., Di Franco, A., ... & Manuel, M. (2017). A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Current Biology*, *27*(7), 958-967.
Stomolophus meleagris	SRR1168418	Li, R., Yu, H., Xue, W., Yue, Y., Liu, S., Xing, R., & Li, P. (2014). Jellyfish venomics and venom gland transcriptomics analysis of Stomolophus meleagris to reveal the toxins associated with sting. *Journal of Proteomics*, *106*, 17-29.
Ceriantheomorphe brasiliensis	SRR11802642	Klompen, A. M., Macrander, J., Reitzel, A. M., & Stampar, S. N. (2020). Transcriptomic analysis of four cerianthid (Cnidaria, Ceriantharia) venoms. *Marine drugs*, *18*(8), 413.
Isarachnanthus nocturnus	SRR11802641	Klompen, A. M., Macrander, J., Reitzel, A. M., & Stampar, S. N. (2020). Transcriptomic analysis of four cerianthid (Cnidaria, Ceriantharia) venoms. *Marine drugs*, *18*(8), 413.
Pachycerianthus maua	SRR11802640	Klompen, A. M., Macrander, J., Reitzel, A. M., & Stampar, S. N. (2020). Transcriptomic analysis of four cerianthid (Cnidaria, Ceriantharia) venoms. *Marine drugs*, *18*(8), 413.

Protein		
Species	Accession	Reference
Acropora digitifera	ADIG_G-PEP_111201	Hemmrich, G., & Bosch, T. C. (2008). Compagen, a comparative genomics platform for early branching metazoan animals, reveals early origins of genes regulating stem-cell differentiation. *Bioessays*, *30*(10), 1010-1018. Shinzato, C., Shoguchi, E., Kawashima, T., Hamada, M., Hisata, K., Tanaka, M., ... & Satoh, N. (2011). Using the Acropora digitifera genome to understand coral responses to environmental change. *Nature*, *476*(7360), 320-323.
Acropora millepora	AMIL_T-PEP_051019	Hemmrich, G., & Bosch, T. C. (2008). Compagen, a comparative genomics platform for early branching metazoan animals, reveals early origins of genes regulating stem-cell differentiation. *Bioessays*, *30*(10), 1010-1018.
Hydra magnipapillata	HMAG_G-PEP_111130	Hemmrich, G., & Bosch, T. C. (2008). Compagen, a comparative genomics platform for early branching metazoan animals, reveals early origins of genes regulating stem-cell differentiation. *Bioessays*, *30*(10), 1010-1018.
Nematostella vectensis	NVEC_G-PEP_111130	Hemmrich, G., & Bosch, T. C. (2008). Compagen, a comparative genomics platform for early branching metazoan animals, reveals early origins of genes regulating stem-cell differentiation. *Bioessays*, *30*(10), 1010-1018.
Thelohanellus kitauei	ASM82789v1	Kevin L Howe, Bruno Contreras-Moreira, Nishadi De Silva, Gareth Maslen, Wasiu Akanni, James Allen, Jorge Alvarez-Jarreta, Matthieu Barba, Dan M Bolser, Lahcen Cambell, Manuel Carbajo, Marc Chakiachvili, Mikkel Christensen, Carla Cummins, Alayne Cuzick, Paul Davis, Silvie Fexova, Astrid Gall, Nancy George, Laurent Gil, Parul Gupta, Kim E Hammond-Kosack, Erin Haskell, Sarah E Hunt, Pankaj Jaiswal, Sophie H Janacek, Paul J Kersey, Nick Langridge, Uma Maheswari, Thomas Maurel, Mark D McDowall, Ben Moore, Matthieu Muffato, Guy Naamati, Sushma Naithani, Andrew Olson, Irene Papatheodorou, Mateus Patricio, Michael Paulini, Helder Pedro, Emily Perry, Justin Preece, Marc Rosello, Matthew Russell, Vasily Sitnik, Daniel M Staines, Joshua Stein, Marcela K Tello-Ruiz, Stephen J Trevanion, Martin Urban, Sharon Wei, Doreen Ware, Gary Williams, Andrew D Yates, Paul Flicek, Ensembl Genomes 2020—enabling non-vertebrate genomic research, *Nucleic Acids Research*, Volume 48, Issue D1, 08 January 2020, Pages D689–D695, https://doi.org/10.1093/nar/gkz890

		Yang, Y., Xiong, J., Zhou, Z., Huo, F., Miao, W., Ran, C., ... & Yao, B. (2014). The genome of the myxosporean Thelohanellus kitauei shows adaptations to nutrient acquisition within its fish host. *Genome biology and evolution*, *6*(12), 3182-3198.

References

1. Jiang JB, Quattrini AM, Francis WR, Ryan JF, Rodríguez E, McFadden CS. A hybrid de novo assembly of the sea pansy (Renilla muelleri) genome. Gigascience. 2019;8(4):1–7.

2. Rivera-García L, Rivera-Vicéns RE, Veglia AJ, Schizas N V. De novo transcriptome assembly of the digitate morphotype of Briareum asbestinum (Octocorallia: Alcyonacea) from the southwest shelf of Puerto Rico. Mar Genomics [Internet]. 2019;47(April):0–1. Available from: https://doi.org/10.1016/j.margen.2019.04.001

3. Stampar SN, Beneti JS, Acuna FH, Morandini AC. Ultrastructure and tube formation in Ceriantharia (Cnidaria, Anthozoa). Zool Anz. 2015;254:67–71.

4. David CN, Suat O, Adamczyk P, Meier S, Pauly B, Chapman J, et al. Evolution of complex structures: minicollagens shape the cnidarian nematocyst. Trends Genet. 2008;24(9):431–8.

5. Stampar SN, Maronna MM, Kitahara M V, Reimer JD, Beneti JS, Morandini AC. Ceriantharia in Current Systematics: Life Cycles, Morphology and Genetics. In: Goffredo S, Dubinsky Z, editors. The Cnidaria, Past, Present and Future. Springer International Publishing Switzerland; 2016. p. 61–72.

6. Mejia ACF, Molodtsova T, Östman C, Bavestrello G, Rouse GW. Molecular phylogeny of Ceriantharia (Cnidaria: Anthozoa) reveals non-monophyly of traditionally accepted families. Zool J Linn Soc. 2019;20:1–20.

7. Stampar SN, Broe MB, Macrander J, Reitzel AM, Brugler MR, Daly M. Linear mitochondrial genome in Anthozoa (Cnidaria): A case study in ceriantharia. Sci Rep. 2019;9(6094):1–12.

8. Kayal E, Bentlage B, Collins AG, Kayal M, Pirro S, Lavrov D V. Evolution of linear mitochondrial genomes in medusozoan cnidarians. Genome Biol Evol. 2011;4(1):1–12.

9. Kayal E, Roure B, Philippe H, Collins AG, Lavrov D V. Cnidarian phylogenetic relationships as revealed by mitogenomics. BMC Evol Biol. 2013;13(5):1–18.

10. Kayal E, Bentlage B, Sabrina Pankey M, Ohdera AH, Medina M, Plachetzki DC, et al. Phylogenomics provides a robust topology of the major cnidarian lineages and insights on the origins of key organismal traits. BMC Evol Biol. 2018;18(1):1–18.

11. Klompen AML, Macrander J, Reitzel AM, Stampar SN. Transcriptomic Analysis of Four Cerianthid (Cnidaria, Ceriantharia) Venoms. Mar Drugs. 2020;18(413):1–24.

12. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol [Internet]. 2019;1–10. Available from: http://dx.doi.org/10.1038/s41587-019-0072-8

13. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 2019;17:155–8.

14. Earl D, Bradnam K, St. John J, Darling A, Lin D, Fass J, et al. Assemblathon 1: A competitive assessment of de novo short read assembly methods. Genome Res. 2011;21(12):2224–41.

15. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: Quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072–5.

16. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.

17. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9(11).

18. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.

19. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.

20. MacManes MD. The Oyster River Protocol : a multi-assembler and kmer approach for de novo transcriptome assembly. PeerJ. 2018;6(e5428):1–18.

21. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008;18:188–96.

22. Holt C, Yandell M. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12(1).

23. Smit AF, Hubley R. RepeatModeler Open-1.0. 2008.

CHAPTER 3

Evolutionary dynamics of gene family gain and loss near the root of the Metazoa tree

Abstract

As knowledge of diverse organisms' genic repertoire grows, the scientific community has had cause to reevaluate the role of gene loss as a major influence in shaping the evolutionary dynamics of animals. Some metazoan lineages in particular, such as Porifera, lack many traits that nearly all other animals possess including a nervous system, gut, or bodily symmetry. Sponges may have always lacked these traits and represent a state of ancestral simplicity, or it is possible that they formerly possessed traits in common with other animals and have since lost them, reflecting a degeneration of complexity. Here, we examine the evolutionary dynamics of gene family gain and loss near the root of the Metazoa tree and show that sponges previously possessed the genic repertoire of other early-branching animal lineages. They lose gene families associated with tissue-grade multicellularity, development and morphology, and nervous systems, while gaining families that could help facilitate interactions with a microbial community. These results are not dependent on the topology of the animal tree, although Ctenophora shows a greater number of gene family losses when the Metazoa

phylogeny is constrained to reflect a hypothesis of Porifera as the sister to the remaining animals. We find that gene family gains typically ascribed to the ancestral metazoan node are divided between that and the node leading to Porifera+ParaHoxozoa, though this pattern shifts in the constrained Porifera-first tree. Our results demonstrate that sponges previously possessed the gene families necessary to have traits similar to other animals, but have since lost them. Though our results with regard to sponges do not change under a Porifera-first hypothesis of animal evolution, these findings will ameliorate concerns on the phylogenetic position of sponges that are based on organismal complexity.

Introduction

Animal traits often arise through genomic novelty (1,2). This novelty results when an animal lineage co-opts genes for new purposes, or neofunctionalizes gene duplicates. Novelty may lead to lineages forming new associations between existing genes, proteins, regulatory networks, or organisms, and these new associations are critical to generating greater animal diversity. While many studies have characterized genetic novelty in various animal clades, it is not the only driver of adaptive shifts in animal evolution.

Gene losses can change the course of evolution in a very different way than does genetic novelty. Rather than provide the raw material for duplications, neofunctionalizations, and co-option, the loss of genes may redirect evolution in new directions by eliminating adaptive possibilities. In some cases, gene loss can be directly

adaptive, as in the Petunia genus, in which *Petunia axillaris* has lost a functional copy of AN2, which codes for a red flower pigment. This results in a white bloom in *P.axillaris* individuals, which in turn makes them more likely to be pollinated by their main pollinator, the nocturnal hawk moth (3). In other cases, however, gene loss could occur because the characteristics a gene provides for are unnecessary to the survival of the organism. This would cause the selective pressure maintaining those genes to relax, and genetic drift could expose them to potential loss-of-function mutations. For example, certain vertebrate lineages have lost the ability to synthesize vitamin C when the lineage also has a diet rich in that vitamin (4). Numerous studies have shown that a significant portion of genes are dispensable (5,6) either though robustness to mutations because of alternative molecular pathways and genetic redundancy (7), or through a lack of relevant environmental pressures needed for expression of that particular gene (8,9). While these gene losses can be nearly neutral, many of them together could open up more energetic or cellular resources, allowing an organism to evolve a more selectively favorable trait. Astyanax cavefish, for example, may have enhanced forebrains and tastebuds through overexpression of shh, which can inhibit the development of eyes, so that the loss of functional eyes could be necessary to acquire these other traits (10,11). Over time, gene losses can compound, leading to an organism whose traits do not reflect its ancestors' level of complexity.

Since the time that sponges have been recognized as animals, scientists have placed the phylum Porifera at the base of the animal tree, as the sister to all other extant metazoans. This is largely because sponges lack many traits that nearly all other

animals have, such as a nervous system, complex body plan, or gut. The apparent simplicity of sponges represented early scientists' ideas about what the first animal must have looked like, over 600 million years ago, before it evolved the organ systems and body structures of more familiar animals.

More recently, phylogenomic studies with more data from underrepresented animal phyla have called into question the placement of Porifera as sister to other animals, suggesting instead that Ctenophora is the first branch of the Metazoa tree (12,13). The growing evidence for this hypothesis and new research on close animal relatives has caused the scientific community to reevaluate when early animals evolved certain traits and to reconsider ideas about the apparent simplicity of the animal ancestor, or Urmetazoan (2,14,15). If the ancestral poriferan was relatively simple, in terms of body plan and tissue complexity, it may be that modern sponges reflect a level of this ancestral simplicity. However if the first poriferans had traits similar to other animal lineages, then extant sponges may represent a degeneration of those traits.

Here, we hypothesize that sponges have lost traits over evolutionary time to become animals without characteristics common to other extant metazoans. We use a dataset composed of 114 species from across Metazoa and Holozoa to construct a well-supported phylogeny using site-heterogenous models and identify gene families present and absent across animal clades. We then use a Dollo parsimony approach to detect gains and losses of these gene families within Porifera and other early branches in the metazoan tree. We find that sponges have lost a substantial amount of gene families, and that the majority of these families are not sponge-specific. The ancestral

Porifera node shows losses in gene families that are associated with tissue-grade multicellularity, such as components of the extracellular matrix and hyaluronic acid binding. Sponges have also lost gene families that are important for developmental-morphogenic processes including the apoptotic process, cell morphogenesis, and the mitotic cell cycle, and those that are related to nervous systems, such as vesicle-mediated transport, receptor clustering, motor activity, and chemotaxis. Gene families gained at the Porifera node include many that help to facilitate interactions with microbes, including caveola assembly, endocytosis involved in viral entry to host, and ectoine transport and binding. Whether Ctenophora or Porifera branches first at the start of Metazoa has little effect on these gains and losses, but does have implications for gene family gains and losses at the ancestral Metazoa node.

Methods

Collection of sequences

In order to sample metazoan diversity, we gathered publicly available genome-scale datasets from metazoan representative organisms. For the genomic datasets, we downloaded protein models directly, and filtered them using *cd-hit* version 4.7 (15) with a threshold of 98% similarity. For the transcriptomic datasets, we downloaded raw Illumina sequence reads and subsampled them down to 35 million read pairs using *seqtk* version 1.2-r94 if there were more reads than that available. We trimmed, error corrected, and assembled the reads using the *Oyster River Protocol* version 2.2.3 (16), and used the final orthomerged assembly in all further analyses. The *Oyster River*

Protocol also runs *TransRate* version 1.0.3 (17) on the finished assemblies, which we used to gauge the assembly quality. We used *TransDecoder* to translate the transcriptome assemblies into predicted proteins and *cd-hit* to filter them, again at a threshold of 98% similarity. We ran *BUSCO* version 3.0.1 (18) with both the eukaryotic database and the metazoan database on all of the protein models from both genomes and transcriptomes, and used only those datasets with at least 80% complete BUSCOs in either of these databases for further analyses. One exception to this was the Hexactinellida; we included three members of this class of sponges despite their lower BUSCO scores because we wanted to make sure that the group was represented in our analysis, and no higher-quality datasets were available.

Phylogenomic analyses and character mapping

For phylogenomic analyses, we first constructed a phylogenomic data matrix including 114 protein sets from transcriptome and whole genome datasets (Table 1) using a best reciprocal *BLAST* approach and the ortholog set determined in Borowiec et al. (19). Here, we searched the *Nematostella* sequence from each of 1080 partitions against each of the current 114 datasets and the top sequence hit for each taxon was then reciprocally searched against the *Nematostella vectensis* genome (19). We retained sequences for which the reciprocal *BLAST* best hit matched the original *Nematostella* sequence query genome locus as orthologs in partition alignments. We did not include sequences for which the reciprocal *BLAST* hit matched a different *Nematostella* genome locus in the data partitions. After eliminating resulting data

partitions that included less than 75% taxon occupancy, filtering individual partition alignments using *Gblocks wrapper* (20), and concatenating partitions into a data matrix, our resulting phylogenomic matrix included 214,569 amino acid positions divided into 704 individual data partitions, each with at least 75% taxon occupancy. Other attempts to produce a *de novo* phylogenomic data matrix using the *OrthoFinder-PhyloTreePruner* (21–23) approach described in Kayal et al. 2018 (24) produced a much smaller dataset (90 data partitions) at 75% taxon occupancy that we did not explore in depth.

We conducted phylogenomic analyses in *IQ-TREE* (25) under the MFP+c60 model, which applies the best fitting model to each partition and approximates a site heterogeneous model by accommodating 60 categories of per-site amino acid equilibrium frequencies (25). Initial analyses under this model produced a topology with maximum support for most nodes, including ctenophores as the sister to the other Metazoa, but failed to recover the monophyly of a few well-accepted, but long-branch clades. Specifically, nematodes, tardigrades, acanthocephalans and platyhelminths fell out into a clade with low support and, as in Borowiec et al. 2015 (19), the position of *Strigamia*, again the sole myriapod in our dataset, favored the Paradoxopoda hypothesis (26) (myriapods sister to chelicerates) rather than the accepted Mandibulata hypothesis (27) (myriapods sister to Pancrustacea). Because these arrangements are likely erroneous and also not pursuant to the present hypotheses, we constrained these taxa using the -g option in *IQ-TREE* to reflect the accepted view that platyhelminths are lophotrochozoans (28) and myriapods are madibulates (29). Additionally, the constrained topology is not significantly less likely than the unconstrained topology. In

either case, both the constrained and unconstrained topologies show maximum support for the ctenophores as sister to the remaining Metazoa in analyses conducted under a site-heterogeneous model. Because the metazoan root is still the subject of controversy, we also analyzed our dataset under the constraint that sponges were the sister to the remaining Metazoa using the -g option in *IQ-TREE*. We conducted likelihood comparisons of topologies in *IQ-TREE* using the -au option to perform an approximately unbiased (AU) test, which tests multiple tree topologies and rejects those that have a p-value less than 0.05 (30).

We found orthogroups in all of the datasets using *OrthoFinder* version 2.3.3 (21,22). In order to see if different clades of organisms were being placed into orthogroups in similar proportions, we created density plots of orthogroup statistics. We created these plots in *ggplot2* version 3.2.1 in *R* version 3.5.2 (31) which include number of orthogroups, percentage of species-specific orthogroups, and percentage of genes in orthogroups. We also tested whether the distributions in these plots were significantly different from one another using Wilcoxon rank sum tests implemented in *R* version 3.5.2 (31). Next, contamination of the genomes and transcriptomes by microbial genetic material could mask gene family losses, or present as gene family gains. We performed alien indexing analysis using *Alien Index* (32) to remove putative contaminate sequences from the orthogroups of interest.

We then used an updated Dollo parsimony procedure (originally described in Plachetzki et al. 2020 (33)) which leverages the raw *OrthoFinder* output and our phylogenetic trees to analyze gain and loss dynamics of gene families for each

phylogeny. Under this procedure orthogroups may evolve once and be lost multiple times, but never re-evolve. Phylogenomic data matrices and all scripts used to create and analyze them are located at https://github.com/jls943/sponge_evol_dynamics.

Analysis of gene family gains and losses

To investigate gene family dynamics, we isolated orthogroups that were gained and lost at the Porifera and Ctenophora ancestral nodes for each topology. We also found the numbers of orthogroups that had been gained and lost at the Metazoa ancestral node and the intermediate node between the first and second branches of Metazoa for both the Ctenophora-first and Porifera-first trees. We compared orthogroups that had been lost at the Porifera node in each topology to one another, and also performed comparisons between the orthogroups gained at the Metazoa node with those lost at Porifera and Ctenophora in each tree.

Many nodes within the Porifera clade also lost orthogroups. To discover if these orthogroups were sponge-specific ones, we found all orthogroups that were gained on each node throughout the Porifera tree. Next, we identified all internal nodes in the Porifera tree that are subtended by a minimum of three tips (Table 2) and identified orthogroups that each of these nodes had lost. We compared these losses to the orthogroups gained at all internal Porifera nodes to determine what proportion of the losses were of sponge-specific orthogroups, and what proportion of lost orthogroups originated at an earlier node.

We also annotated the orthogroups gained and lost at Porifera and Ctenophora in each topology, as well as the orthogroups gained at the Metazoa and sponges and the remaining Metazoa (Porifera+ParaHoxozoa) nodes in the Ctenophora-first tree. We used *usearch* version 9.2.64 (34) to identify centroid sequences in each orthogroup of interest, and *InterProScan* version 5.44-79.0 (35) to annotate the centroid sequences for each orthogroup. From these annotations, we extracted the gene ontology (GO) terms associated with each orthogroup and combined them in groups that correspond to gains and losses at our nodes of interest. We isolated unique GO terms in each of these groups and compared the terms in the gains to the corresponding losses at the same node, eliminating any overlapping GO terms. These unique and non-overlapping GO terms we clustered using *REVIGO* (36) using the "small" setting (allowing 50% similarity between terms) for all GO sets except those for Ctenophora losses and Porifera+ParaHoxozoa gains, for which we used the "tiny" setting (allowing 40% similarity between terms) due to the greater number of GO terms. We then plotted the clustered GO terms into treemaps using a *REVIGO*-provided protocol in *R* version 3.5.2 (31).

Results

The tree topology is well-resolved with full support

Our phylogenomic analysis yielded a well-resolved tree (Figure 1) under the best-fit site-heterogenous model implemented in *IQ-TREE* (25). This model approximates the CAT model implemented in *PhyloBayes* (37). Our tree has maximum support for both aLRT

and bootstrapping at all nodes, including Ctenophora as the first branch of Metazoa. When we used the AU test (30) to compare the topology that aligns with our data to a Porifera-first topology, we found overwhelming support for the Ctenophora-first tree ($p = 1$, failed to reject) vs. the Porifera-first tree ($p < 0.001$, reject).

Sponges are well-represented in both taxon sampling and orthogroups

After filtering genomes using *BUSCO* score (18) and transcriptomes using *BUSCO* and *TransRate* scores (17), we retained 114 taxa for use in further analysis including 107 metazoan species (24 sponge species) and 7 outgroups (Table 1). We identified 105,177 orthogroups through *OrthoFinder* (22), and tested to make sure that poriferan species were not being placed into orthogroups at a lower rate than other metazoan species. We used Wilcoxon rank sum tests to quantify the differences in the distributions of number of orthogroups each species had, percentage of genes classified into species-specific orthogroups for each species, and percentage of genes placed into orthogroups (as opposed to remaining unclassified) for each species (Figure 2). The distributions of number of orthogroups and percentage of genes in species-specific orthogroups were not significantly different for sponges compared to other metazoan organisms (number of orthogroups: $p = 0.123$; percent genes in species-specific orthogroups: $p = 1$), indicating that the sponge datasets are performing comparably to other metazoan datasets. The proportion of genes placed into orthogroups was significantly different ($p = 0.0208$), however genes from sponges were placed into orthogroups at a higher rate (85.9% of the time on average) compared to other

metazoan organisms (79.5%), possibly due to the extensive sampling of sponges in our dataset (Table 1).

Stepwise accumulation of metazoan genomic repertoire

Based on Dollo parsimony analysis, gene families are gained and lost throughout the history of Metazoa. A substantial gain of many gene families often accompanies the branching of a major clade, such as at those leading to Choanozoa (4,656), Metazoa (1,912), and Porifera+ParaHoxozoa (13,283). However, the pattern of orthogroups gained shifts depending on the topology of the tree. In the Ctenophora-first tree that is based on our data, the node leading to Porifera+ParaHoxozoa gains a large number of orthogroups (Figure 1,2), but in the constrained Porifera-first tree this node does not exist, and most of those gains are shifted onto the Metazoa node instead (Figure 3). A similar phenomenon happens for the Porifera-first tree, in that all of the orthogroups (958) gained at the node leading to Ctenophora+ParaHoxozoa shift to the Metazoa node in the Ctenophora-first topology, though because it is a much smaller number of orthogroups, the shift is less dramatic. Losses at these nodes are quite minimal and mainly occur on branches leading to individual phyla rather than the backbone of the tree.

Regardless of topology, the ancestral poriferan genome was dismantled by gene family loss

At the ancestral sponge node, our Dollo parsimony analysis showed that sponges gained 1,317 orthogroups and lost 2,765 orthogroups (Figure 2,3). All nodes that we examined were subtended by a minimum of three taxa so that all of our inferences are based on at least three datasets. Even with this restriction, many internal sponge nodes show dramatic losses, such as those leading to Hexactinellida (16246), Homoscleromorpha+Calcarea (12335), Myxospongia (13217), and Haplosclerida (10724) (Table 2). In some cases, these were losses of sponge-specific gene families, but the loss of sponge-specific gene families only represented the majority of losses at two internal poriferan nodes, Poecilosclerida and Haplosclerida2. Both of these nodes are among those closest to the tips of the tree and have many other internal nodes (and therefore chances to gain sponge-specific orthogroups) between them and the ancestral poriferan. Calcarea and Hexactinellida also show substantial gene family gains (2335 and 2210 orthogroups, respectively), though these are still far fewer than the losses at these nodes. Indeed, Demospongiidae is the only internal sponge node at which orthogroup gains outweigh losses (984 gains to 635 losses).

Magnitude of gene family losses at Ctenophora depends on the topology

The node at the origin of Ctenophora gained 2,767 orthogroups and lost 6,180 (Figure 3). We also tested the gains and losses at nodes of interest using a tree that we constrained so that Porifera is the first branch. Under this phylogeny, the gene families that were gained at the Porifera and Ctenophora branches remain consistent with those from the tree that is based on our data, but the number of orthogroup losses at the

Porifera node decreased from 2,765 to 1,854. Of these losses, nearly all (1,808) are shared in common with the orthogroups lost at the Porifera node in the well-supported Ctenophora-first tree, above. The losses at the Ctenophora node increased dramatically from 6,180 to 18,572 (Figure 4), and the majority of these losses (13,284 orthogroups) correspond to orthogroups gained at the Metazoa node under this tree structure.

GO terms that correspond to orthogroup gains and losses at Porifera and Ctenophora are not dependent on topology, and GO terms corresponding to gains and losses generally overlap only partially

Despite different numbers of orthogroups lost at the Porifera and Ctenophora nodes in the different topologies, the numbers of GO terms corresponding to those losses was fairly consistent. In the Ctenophora-first tree, the Porifera node lost 562 GO terms and the Ctenophora node lost 1949. For the Porifera-first tree, the Porifera node lost 510 terms and the Ctenophora node lost 1920. Since the Dollo parsimony approach bases the orthogroups gained at a specific node on orthogroups present in taxa included in that node, the gains found at the Porifera and Ctenophora nodes for the Porifera-first tree are identical to those in the Ctenophora-first tree. Gene ontology (GO) terms for the gains and losses at our focal nodes overlapped somewhat, but never entirely. For the nodes in the Ctenophora-first tree, the losses at Metazoa and Porifera+ParaHoxozoa were very minimal, but overlapped with the gains at those nodes to a significant extent (Figure 5A, B). The gains and losses at Porifera and Ctenophora show more overlapping GO terms overall (182 terms in Porifera and 238 terms in Ctenophora), but

these make up a much smaller proportion of the total losses than in the Metazoa and Porifera+ParaHoxozoa nodes (Figure 5C, D). In the Porifera-first topology the Porifera gains and losses overlapped by 174 terms, and the gains and losses at the Ctenophora node again had 238 overlapping terms. We removed GO terms that overlapped before our analysis of gene ontology for gains and losses at each node.

Poriferans lose gene families associated with multicellularity, morphogenesis, and nervous systems, and gain those related to microbial interactions

We characterized the GO terms associated with orthogroups gained at the Porifera node and found that the orthogroups gained correspond to GO terms related to interactions with microbes, including caveola assembly, endocytosis involved in viral entry to host, and ectoine transport and binding (Figures 6-8), which each have one orthogroup associated with them (Table S1). Conversely, many of the orthogroups that have been lost at the ancestral sponge node are related to developmental-morphogenic processes including apoptotic process, cell morphogenesis, and the mitotic cell cycle, or related to tissue-grade complexity such as extracellular matrix and hyaluronic acid binding. Sponges have also lost orthogroups associated with nervous systems including those involved in vesicle-mediated transport, receptor clustering, and motor activity (Figures 9-11). Of these losses, extracellular matrix has three orthogroups lost, motor activity has five, and each of the others has one or two orthogroups associated with it (Table S1). In the constrained Porifera-first tree, GO terms associated with orthogroups

lost at the Porifera node are strikingly similar to and include many of the same terms from the Ctenophora-first Porifera losses (Figures 12-14).

Few genomic gains at the ctenophore ancestor, but extensive loss of metabolic functionality

At the node representing the origin of Ctenophora, the gains include orthogroups associated with mitotic spindle assembly, clathrin adaptor complex, and RNA transmembrane transporter activity (Figures 15-17). The losses at this node are numerous, and correspond to digestion, brush border assembly, and insulin receptor substrate binding (Figures 18-20). Both the GO term gains and losses highlighted here correspond to either one or two orthogroups each (Table S1). The Ctenophora node lost three times the number of orthogroups lost at that node in the Ctenophora-first tree. Nevertheless the numbers of unique GO terms associated with those losses remain fairly similar (Ctenophora-first topology: 1,949 terms, Porifera-first topology: 1,920 terms), and all GO terms lost at the Ctenophora node in the Porifera-first tree (Figures 21-23) are also found in the losses for the Ctenophora-first topology.

Gains at the ancestral Metazoa node correspond to multicellular processes and cell signaling

Through our analysis of the orthogroups gained along the branch leading to the Metazoa node, we found that these orthogroups are associated with GO terms that have to do with basic processes of multicellular organisms. These include cell

population proliferation, cell adhesion, cell-cell junction, and extracellular space. Orthogroups related to cell communication and signaling are also gained at this node, such as Wnt-protein binding and coreceptor activity (Figures 24-26). While most of these GO terms have only one or two orthogroups that correspond to them, extracellular space has six associated with it, and cell adhesion has 17. The Porifera+ParaHoxozoa node gains orthogroups that have to do with sensory systems such as detection of visible light (which is associated with three different orthogroups) and ion channel regulator activities, and also those that are associated with cellular organization and regulation including aging and regulation of autophagy (Figures 27-29). All orthogroups corresponding to GO terms highlighted here can be found in Table S1.

Discussion

Gene loss can be a significant driver of evolutionary change in a lineage of organisms. We used a phylogenetically informed Dollo parsimony procedure to identify orthogroups that have been gained and lost in the earliest-branching Metazoa clades. We find that at the Porifera node, sponges lose gene families associated with multicellularity, nervous systems, and morphogenetic processes, while gaining many gene families that may facilitate interactions with diverse microbes. Ctenophores lose gene families relating to metabolism and digestion, and gain those that correspond to developmental functions. Gains and losses at these nodes are robust to changes in topology, however we find that the gene repertoire gained at the Metazoa ancestral node shifts dramatically depending on the branching orders of Ctenophora and Porifera.

Sponges have lost gene families that are associated with multicellularity, but gained those that contribute to their holobionts.

Multicellularity in animals is characterized by communication and structure between cells, and coordination of cellular processes such as growth, division, and death (38–40). The traditional view of animal relationships explains the low organismal complexity of extant sponges by invoking sponges as ancestral in nature, hence their placement as the sister to the remaining Metazoa by the proponents of the Porifera-first hypothesis (41–43). In our analyses, we would find support for the Porifera-first hypothesis if we observed limited gene family gain and loss, and the patterns therein would not reflect particular functional relationships to processes associated with multicellularity, development, and morphogenesis. If however, modern sponges have degenerated in complexity, we would expect to find that they have lost gene families that are associated with these functions. Our findings strongly reject the former case and we infer that by the time of their last common ancestor, sponges had already lost much of the genetic potential to construct a tissue-grade organism.

Modern sponges are a unique clade of organisms. While they lack many traits that most other animal groups have, they have a distinctive biology and can respond to their environments in sophisticated ways (44–46). Many sponge lineages have developed rich microbial communities that support their defense (47), immune response (48), and metabolic requirements (49). Maintaining or encouraging the success of these communities could be a strong selective force, either for sponges to lose gene families

that might interfere with microbial interactions, such as elements of a nervous system or the sensory perception of certain chemicals, or to gain gene families that could facilitate more microbial interactions, including ones that enable cells and other particles to be brought into the cell or ones that provide for the binding and transport of microbially produced compounds. Our results reflect exactly these types of changes and show that sponges have altered their genetic repertoire in a way that allows them to be successful hosts to their complex microbial communities.

Ctenophores show losses of gene families related to metabolism, and gains connected to cell cycle regulation

Ctenophores have complex morphologies characterized by rotational symmetry, and many traits or components of traits in common with many other animal lineages, such as nervous systems and complex developmental processes. While the losses at the Ctenophora node are numerous, GO terms associated with metabolic processes dominate, and those associated with development are conspicuously absent. GO terms associated with gains at the Ctenophora node are much more sparse, and have to do with cell growth and communication.

Evolutionary dynamics at the ancestral metazoan node are dependent on lineage branching order

The branch of the tree leading to the ancestral Metazoa node is a pivotal one in animal evolution. Previous studies show that gene families relating to transcription factors,

signaling proteins, and developmental receptors either originated or greatly expanded on this branch (2,14). We find similar gene families represented by the orthogroups gained at the Metazoa and Porifera+ParaHoxozoa nodes, with many orthogroups gained that relate to multicellular development and regulation, and sensory systems and signaling, as we might expect near the origin of Metazoa. However, the pattern of gained orthogroups shifts depending on the topology of the tree. In the Ctenophora-first tree that is favored by our data, the node leading to Metazoa gains a substantial number of orthogroups (1,913), but many of the gains are concentrated on the Porifera+ParaHoxozoa (13,283) (Figure 1,2). In the tree that we constrained so that Porifera branch first, this latter node does not exist, and many of those gains are transferred onto the Metazoa node instead (Figure 3). The shift in gains means that a change in the phylogeny necessitates a change in our hypotheses about the genic complexity of ancient metazoans. If, as our and other analyses suggest, the Ctenophora represent the sister group to the remaining Metazoa (12,19,50,51), then the ancient gain of gene families was likely spread over multiple nodes, both before and after Ctenophora branches from other lineages. However if Porifera branch first (41–43), we infer that many of the gains we observe among these early nodes would instead be concentrated at the origin of Metazoa. Therefore the position of Porifera changes our interpretation of the genome content of the ancestral metazoan.

Gene family losses do not represent missing data in Porifera datasets

If the sponge datasets we used were less complete than other datasets, the losses we observe could be attributed to genes that are missing from the datasets. However, apart from the hexactinellid sponges, all of the datasets have a *BUSCO* score of 80% complete or higher, so it is unlikely that observed losses could be due to missing data alone. For each clade including the Hexactinellida, we identified gene family losses at nodes that are subtended by at least three species, meaning that each of the three taxa would need to be missing the same gene families in order for missing data to show up as a loss in our results. Therefore we conclude that the gene families that we find to be lost at various nodes are not the result of incompleteness in the datasets.

OrthoFinder proves robust to highly divergent sequences and gene sorting mistakes

Highly divergent protein sequences can complicate the process of sorting genes into orthogroups. Genes that are homologous may have diverged far enough that their sequences are dissimilar and difficult to recognize. If sponge genes are more prone to this dissimilarity than other organisms, sponges may be represented in orthogroups in a way that does not reflect the true homology of their sequences. For example, a highly divergent gene may be unclassified, rather than placed into an orthogroup, and this might cause the appearance of a loss of that gene family in that species. We examined the patterns of gene sorting done by *OrthoFinder* (21,22), and compared Porifera to both the outgroups and the rest of animals. We found no indication that sponge genes were misclassified or left out of orthogroups at a disproportionate rate compared with other taxa (Figure 1).

We also compared the GO terms that correspond to the orthogroups gained and lost at our nodes of interest. If *OrthoFinder* incorrectly assigned highly divergent genes to species- or clade-specific orthogroups, rather than to a larger or more inclusive orthogroup to which they really belonged, these orthogroups could have different evolutionary dynamics in our analysis, which could lead to the same GO terms in gains and losses at the same node (52). However, in our analysis, GO terms for gains and losses overlapped only partially at Metazoa, Porifera, and Ctenophora, and losses were very few in number at Porifera+ParaHoxozoa. These overlapping terms were excluded from further analysis, and we conclude that this potential issue is not widespread in our results (Figure 5).

GO terms represent conservative estimates of gene family gains and losses

For organisms like sponges and ctenophores, all GO term analyses must be interpreted carefully, as the organisms in which the terms were originally designed are all highly divergent from these non-bilaterian taxa. The patterns we see here are therefore based on high-level terms that are more likely to be conserved across vast evolutionary distances, rather than more specific terms that could be useful in a finer-scale analysis. We also acknowledge that these lists of GO terms are almost certainly incomplete, as the sequences for a sponge species are unlikely to be annotated as often as those from a human dataset. Despite these inherent limitations, the trends we find in the GO terms remain clear, and point to a degeneration of sponges through their evolutionary history. Further, because of the issues with annotation for these clades, the GO terms we find

gained and lost at these nodes represent conservative inferences, and may not include the extent of the evolutionary dynamics, rather than overreaching the magnitude of change.

Conclusions

Taken together, our findings suggest that ancient organisms near the origin of the sponge lineage possessed more characteristics of organisms with tissue-grade complexity, and that poriferans have subsequently lost many of the necessary gene families for these functions. In their place, sponges have gained gene families that enable them to maintain complex symbioses with diverse microbes. These results do not rely on the branching order of the first metazoan lineages, however the evolutionary dynamics of early animals shift dramatically to be concentrated on the ancestral Metazoa node when mapped onto a constrained Porifera-first tree.

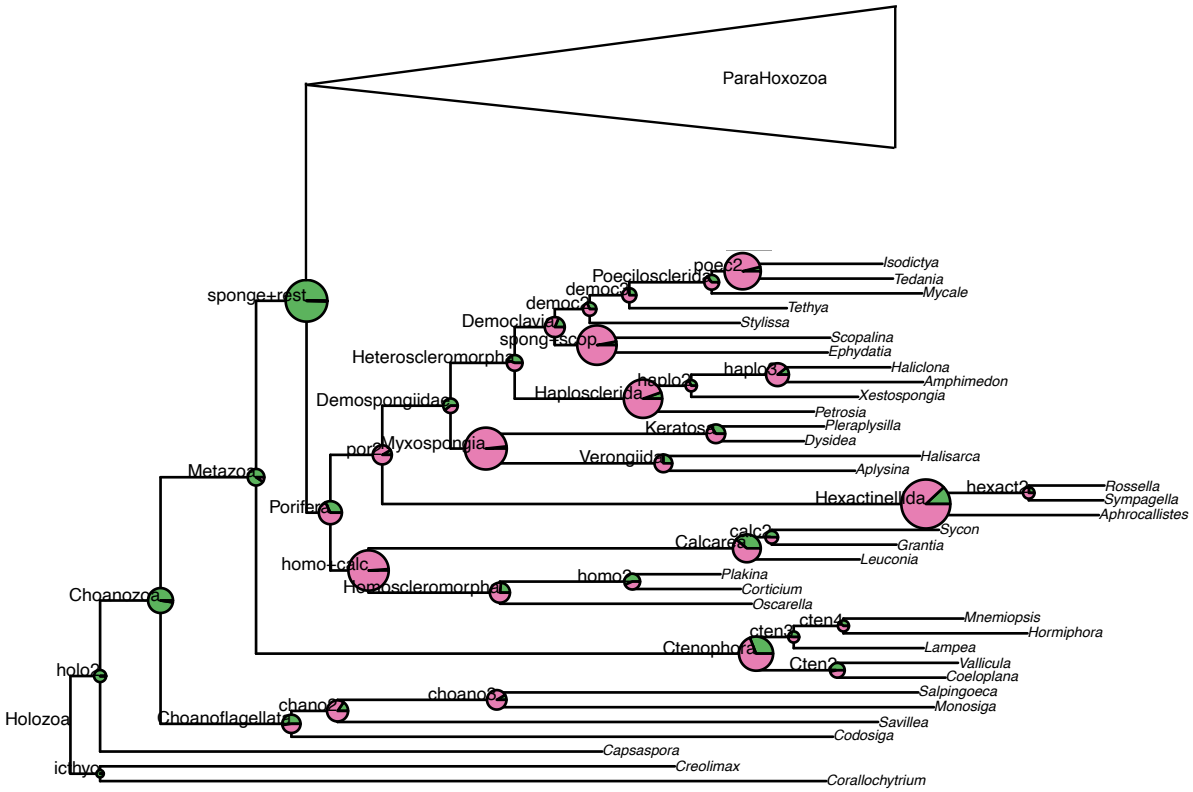

Figure 1: Phylogenomic tree based on our data showing Ctenophora as the first branch of Metazoa. The size of the pie charts on each internal node correspond to the magnitude of change at that node, with green portions representing orthogroups gained, and pink portions representing orthogroups lost, according to Dollo parsimony analysis. The branch leading to ParaHoxozoa has been collapsed for simplicity.

Table 1: All species used in our phylogenetic tree and orthogroup analyses. We required each dataset to have a BUSCO score (with either the Eukaryota database or the Metazoa database) of at least 80% complete and a TransRate score of at least 0.22 to be included in the analysis. The only exceptions are the Hexactinellid sponges, for which no dataset matching these criteria is available.

Species	Phylum	BUSCO score (Eukaryota)	BUSCO score (Metazoa)	TransRate score	Accession or source
Acanthaster planci	Echinodermata	99.60%	98.50%	NA	PRJNA397419, PRJDB3175
Acropora digitifera	Cnidaria	83.80%	80.50%	NA	Compagen
Alcyonium palmatum	Cnidaria	80.90%	81.60%	0.4769	SRR3407216
Amphimedon queenslandica	Porifera	95.00%	93.30%	NA	EnsemblMetazoa
Anolis carolinensis	Chordata	84.20%	84.60%	0.2375	SRR391653
Anthopleura elegantissima	Cnidaria	97.70%	94.70%	0.56126	SRR1645256
Antipathes caribbeana	Cnidaria	87.70%	86.60%	0.5007	SRR3407160
Aphrocallistes vastus	Porifera	56.70%	57.30%	0.37058	SRR1068281
Aplysina aerophoba	Porifera	93.40%	87.70%	0.46766	ERR2560040
Apostichopus japonicus	Echinodermata	86.40%	86.90%		PRJNA37797
Asymmetron lucayanum	Chordata	98.00%	97.30%	0.4738	SRR1138335
Aurelia aurita	Cnidaria	81.90%	80.40%	NA	PRJNA17891
Bathymodiolus platifrons	Mollusca	91.10%	88.30%	0.2556	SRR3866526
Bdellocephala annandalei	Platyhelminthes	87.40%	83.80%	0.2322	DRR014788
Bombus impatiens	Arthropoda	99.70%	99.70%	NA	EnsemblMetazoa
Caenorhabditis elegans	Nematoda	99.70%	89.00%	NA	EnsemblMetazoa
Capitella teleta	Annelida	97.40%	97.30%	NA	EnsemblMetazoa
Capsaspora owczarzaki	Filesterea	95.70%	NA	NA	EnsemblProtists
Chironex fleckeri	Cnidaria	79.20%	80.70%	0.53972	SRR1819888
Ciona intestinalis	Chordata	89.40%	84.50%	NA	Ensembl
Codosiga hollandica	Choanozoa	90.10%	NA	0.4831	SRR6344973
Coeloplana meteoris	Ctenophora	94.10%	85.80%	0.4298	SRR3407215
Corallium rubrum	Cnidaria	95.00%	91.10%	0.44782	SRR1552944
Corallochytrium limacisporum	Ichthyosporea	96.10%	NA	0.6293	SRR1618557
Corticium candelabrum	Porifera	81.90%	80.40%	0.31951	SRR504694

Craspedacusta sowerbyi	Cnidaria	95.10%	93.30%	0.54119	SRR923472
Crassostrea gigas	Mollusca	81.90%	84.60%	NA	EnsemblMetazoa
Creolimax fragrantissima	Ichthyosporea	91.10%	NA	0.4657	SRR1029670
Danaus plexippus	Arthropoda	96.70%	97.50%	NA	EnsemblMetazoa
Danio rerio	Chordata	82.90%	85.00%	NA	Ensembl
Daphnia magna	Arthropoda	93.80%	91.70%	NA	EnsemblMetazoa
Doliolum nationalis	Chordata	89.10%	86.20%	0.5516	SRR6326578
Drosophila melanogaster	Arthropoda	100.00%	99.30%	NA	EnsemblMetazoa
Dysidea avara	Porifera	95.70%	86.90%	0.42687	ERR2560071
Echinorhynchus gadi	Acanthocephala	87.40%	73.30%	0.3839	SRR2131254
Ephydatia muelleri	Porifera	93.10%	87.30%	0.5437	SRR1041944
Eptatretus burgeri	Chordata	90.50%	89.80%	NA	Ensembl
Eudiplozoon nipponicum	Platyhelminthes	84.50%	77.20%	0.5455	SRR5816789
Gallus gallus	Chordata	89.70%	86.90%	NA	Ensembl
Glossoscolex paulistus	Annelida	93.40%	93.50%	0.513	SRR1519963
Golfingia vulgaris	Annelida/Sipuncula	93.00%	94.80%	0.4124	SRR1797875
Gorgonia ventalina	Cnidaria	96.70%	91.60%	0.57707	SRR935083
Grantia compressa	Porifera	93.40%	88.80%	0.5383	SRR3417193
Haliclona amboinensis	Porifera	81.80%	75.10%	0.58572	SRR1630907
Halisarca dujardini	Porifera	87.80%	81.60%	0.54496	ERR1143553
Helobdella robusta	Annelida	96.40%	93.30%	NA	EnsemblMetazoa
Homo sapiens	Chordata	100.00%	100.00%	NA	Ensembl
Hormiphora californensis	Ctenophora	96.10%	85.80%	0.4064	SRR1992642
Hydra vulgaris	Cnidaria	95.70%	91.40%	NA	PRJNA31231
Hydractinia symbiolongicarpus	Cnidaria	93.70%	92.10%	0.41473	SRR1796511
Hypsibius dujardini	Tardigrada	88.50%	79.90%	0.4589	SRR1739983
Isodictya sp	Porifera	93.40%	86.00%	0.53914	SRR6202911
Lampea pancerina	Ctenophora	94.10%	84.90%	0.4282	SRR3407163
Lepeophtheirus salmonis	Arthropoda	87.80%	81.30%	NA	PRJNA15531
Lepisosteus oculatus	Chordata	82.90%	84.80%	NA	Ensembl
Lucernaria quadricornis	Cnidaria	91.40%	88.80%	NA	ERR2248383
Leuconia nivea	Porifera	88.10%	79.20%	0.52762	SRR3417190

Limulus polyphemus	Arthropoda	92.00%	93.80%	NA	PRJNA238073
Lingula anatina	Brachiopoda	95.40%	96.60%	NA	EnsemblMetazoa
Lottia gigantea	Mollusca	97.30%	96.10%	NA	EnsemblMetazoa
Meara stichopi	Xenacoelomorpha	91.80%	86.30%	0.3047	SRR2681155
Membranipora membranacea	Bryozoa	90.70%	89.80%	0.4662	SRR2131259
Mnemiopsis leidyi	Ctenophora	87.40%	80.30%	NA	EnsemblMetazoa
Monodelphis domestica	Chordata	79.50%	81.50%	NA	Ensembl
Monosiga brevicollis	Choanozoa	82.60%	NA	NA	EnsemblProtists
Mus musculus	Chordata	91.40%	91.50%	NA	Ensembl
Mycale grandis	Porifera	93.70%	87.90%	0.4831	SRR3339394
Nanomia bijuga	Cnidaria	89.70%	86.40%	0.27426	SRR871527
Nasonia vitripennis	Arthropoda	92.10%	93.80%	NA	EnsemblMetazoa
Nematostella vectensis	Cnidaria	94.80%	93.70%	NA	EnsemblMetazoa
Nemopilema nomurai	Cnidaria	96.30%	93.50%	NA	PRJNA415234
Neomenia megatrapezata	Mollusca	80.20%	79.40%	0.2763	SRR331899
Occasjapyx japonicus	Arthropoda	83.80%	85.00%	0.4966	SRR1182465
Octopus bimaculoides	Mollusca	90.40%	92.70%	NA	EnsemblMetazoa
Oscarella carmela (or pearsei)	Porifera	94.70%	87.50%	0.51527	SRR1042012
Palythoa variabilis	Cnidaria	94.10%	87.40%	0.57525	SRR1952746
Parasagitta elegans	Chaetognatha	90.40%	88.20%	0.3935	SRR7754742
Pelagia noctiluca	Cnidaria	95.40%	92.90%	0.534	SRR3407257
Petrosia ficiformis	Porifera	81.20%	77.20%	0.355	SRR504688
Physalia physalis	Cnidaria	93.70%	87.50%	0.41561	SRR871528
Plakina jani	Porifera	83.50%	81.80%	0.5902	SRR3417194
Pleraplysilla spinifera	Porifera	82.60%	77.10%	0.56608	SRR3417588
Plumapathes pennacea	Cnidaria	91.80%	89.50%	0.519	SRR3407161
Pocillopora damicornis	Cnidaria	90.40%	90.60%	NA	PRJNA506040
Polypodium hydriforme	Cnidaria	94.70%	86.20%	0.393	SRR1336770
Praesagittifera naikaiensis	Xenacoelomorpha	82.80%	76.30%	NA	ftp://parrot.genomics.cn/gigadb/pub/10.5524/100001_101000/100564/
Priapulis caudatus	Priapulida	91.00%	92.00%	NA	PRJNA303167, PRJNA20497

Pristionchus pacificus	Nematoda	83.50%	67.70%	NA	EnsemblMetazoa
Proasellus beticus	Arthropoda	89.10%	86.90%	0.4523	ERR1433113
Prothalotia lehmanni	Mollusca	82.20%	83.90%	0.4514	SRR1505133
Pteraster tesselatus	Echinodermata	97.10%	93.80%	0.4726	SRR2846094
Renilla muelleri	Cnidaria	88.80%	84.80%	NA	ReefGenomics
Rhabdopleura sp.	Hemichordata	78.50%	81.10%	0.28	SRR1806842
Rhodactis indosinensis	Cnidaria	97.70%	94.40%	0.47323	SRR3201278
Rossella fibulata	Porifera	64.40%	55.50%	0.17366	SRR1915835
Saccoglossus kowalevskii	Hemichordata	92.10%	91.30%	NA	PRJNA42857, PRJNA12887
Salpingoeca rosetta	Choanozoa	88.80%	NA	NA	EnsemblProtists
Savillea parva	Choanozoa	85.10%	NA	0.5562	SRR6344983
Scopalina sp CDV2016	Porifera	94.10%	87.50%	0.52681	SRR3708901
Stegodyphus mimosarum	Arthropoda	81.20%	85.80%	NA	EnsemblMetazoa
Stenostomum sthenum	Platyhelminthes	97.70%	88.80%	0.5774	SRR1801788
Strigamia maritima	Arthropoda	92.10%	91.60%	NA	EnsemblMetazoa
Stylissa carteri	Porifera	83.10%	81.10%	0.33212	SRR1738069
Sycon ciliatum	Porifera	84.50%	78.40%	NA	Compagen
Sympagella nux	Porifera	84.50%	77.60%	0.14738	SRR1916581
Tedania anhelans	Porifera	88.80%	85.80%	0.54022	SRR3708911
Terebratalia transversa	Brachiopoda	90.40%	86.80%	0.4785	SRR2564755
Tethya wilhelma	Porifera	90.70%	88.60%	0.52739	SRR4255675
Tribolium castaneum	Arthropoda	98.40%	98.30%	NA	EnsemblMetazoa
Trichoplax adhaerans	Placozoa	97.00%	91.20%	NA	EnsemblMetazoa
Vallicula multiformis	Ctenophora	91.10%	82.40%	0.4713	SRR3407164
Xenopus tropicalis	Chordata	80.50%	81.70%	NA	Ensembl
Xenoturbella bocki	Xenacoelomorpha	88.20%	85.20%	0.3662	SRR2681987
Xestospongia testudinaria	Porifera	86.10%	83.50%	0.38335	SRR1738073

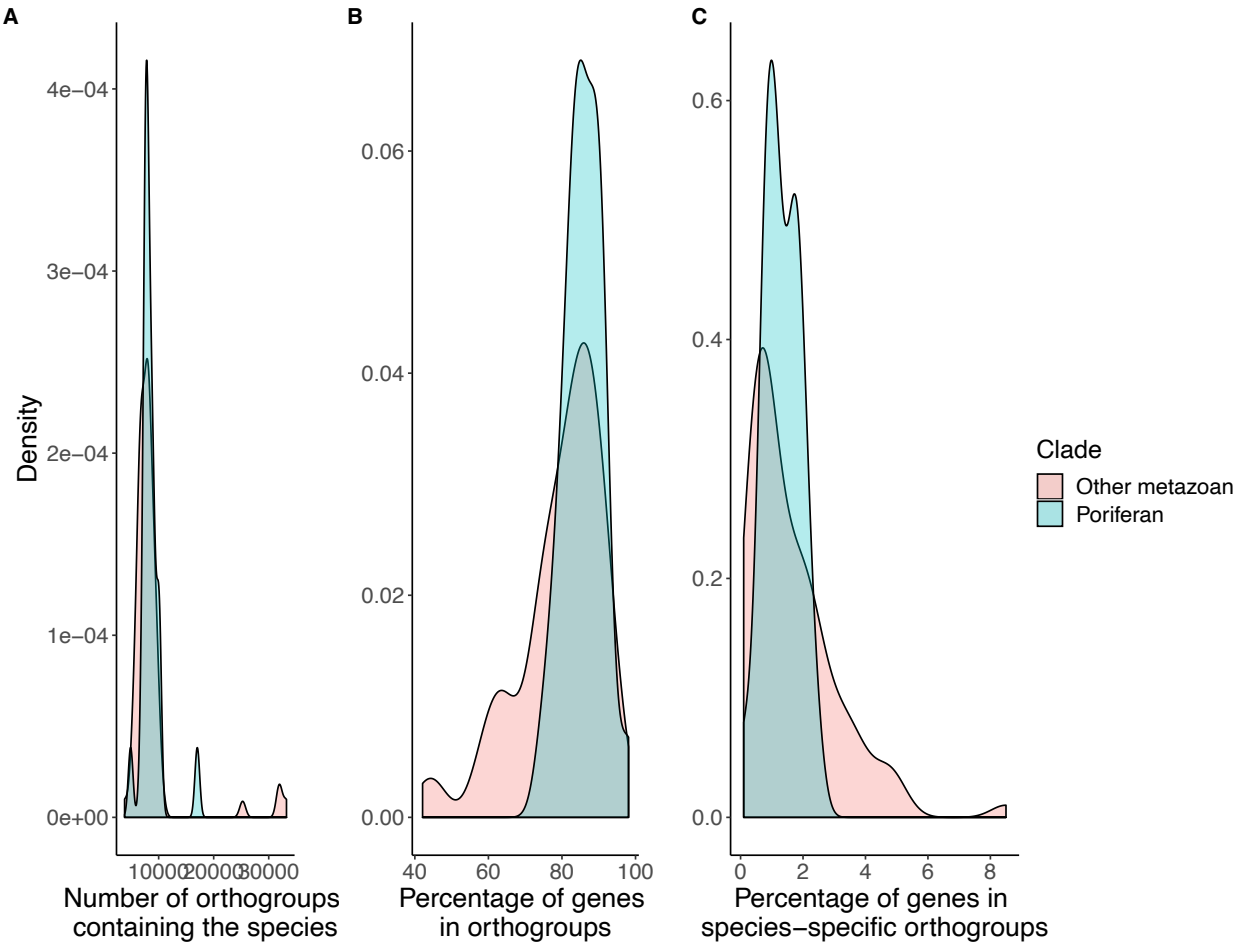

Figure 2: Genes of sponge species are placed into orthogroups in similar proportions to other metazoans. A: Density plot of the number of orthogroups containing each species; distributions are not significantly different (P = 0.123). B: Density plot of the percentage of genes from each species that were placed into orthogroups; distributions are significantly different, with sponge species having a higher percentage of genes placed into orthogroups (P = 0.0208). C: Density plot of the percentage of genes that were placed into species-specific orthogroups; distributions are not significantly different (P = 1).

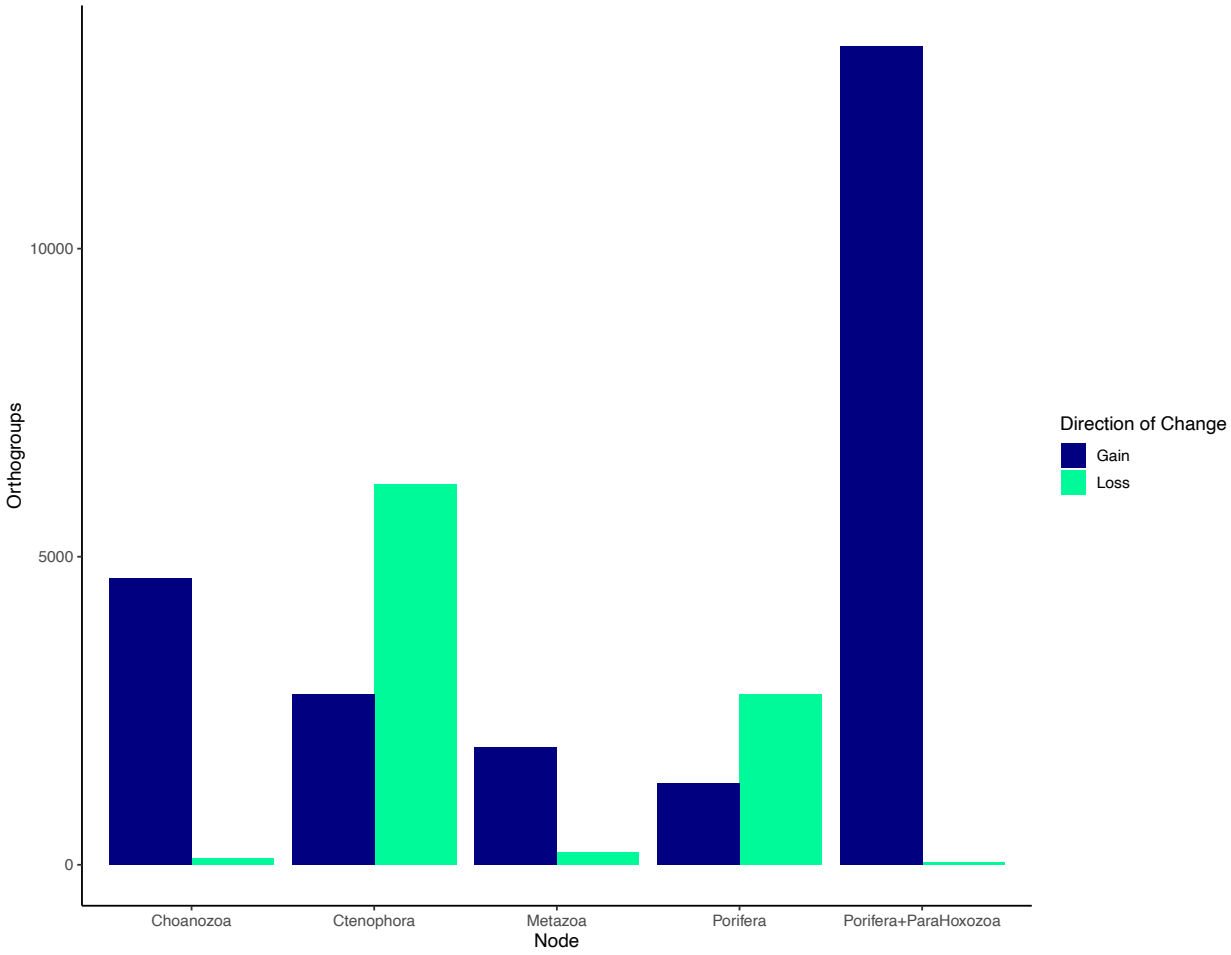

Figure 3: Orthogroup gains and losses for nodes of interest in the holozoan tree that is based on our data, with Ctenophora as the first branch of Metazoa. In this topology, the Ctenophora and Porifera nodes lose substantial numbers of orthogroups (6,180 and 2,765, respectively), and many of the gene family gains in early metazoan evolution occur on the Porifera+ParaHoxozoa node (13,283).

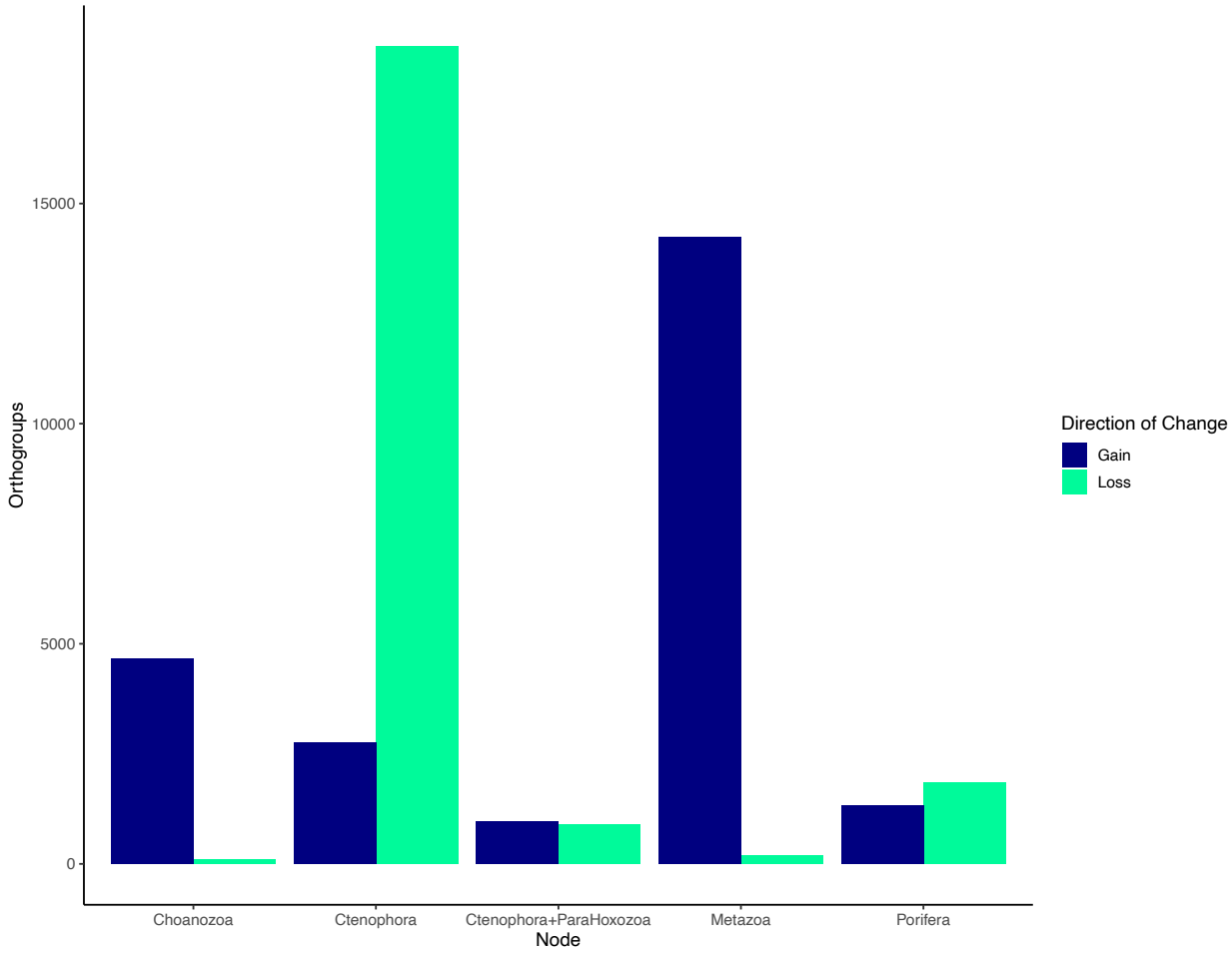

Figure 4: Orthogroup gains and losses for nodes of interest in the holozoan tree that is constrained so that Porifera is the first branch of Metazoa. In this topology, the Ctenophora node loses many more orthogroups than in the Ctenophora-first tree (18,572), and the Porifera node loses fewer (1,854). Nearly all of the gains that occur on the Porifera+ParaHoxozoa node in the Ctenophora-first tree are shifted to the Metazoa node in this topology, which shows 14,238 orthogroups gained.

Table 2: Gene family losses at specific nodes within Porifera are mainly orthogroups acquired before the ancestral Porifera node. In only two cases (Poecilosclerida and Haplosclerida2) do sponge-specific orthogroups form the majority of losses at an internal sponge node.

Internal poriferan node	Number of orthogroups gained	Number of orthogroups lost	Number of sponge-specific orthogroups lost
Homoscleromorpha+Calcarea	144	12335	0
Homoscleromorpha	722	2374	270
Calcarea	2335	3635	943
Hexactinellida	2210	16246	1169
Hexactinellida+Demospongiidae	237	2572	0
Myxospongia	230	13217	976
Demospongiidae	984	635	61
Heteroscleromorpha	841	1060	173
Haplosclerida	617	10724	700
Haplosclerida2	314	662	355
Democlavia	570	2492	914
Democlavia2	433	941	362
Democlavia3	387	1135	501
Poecilosclerida	600	1066	576

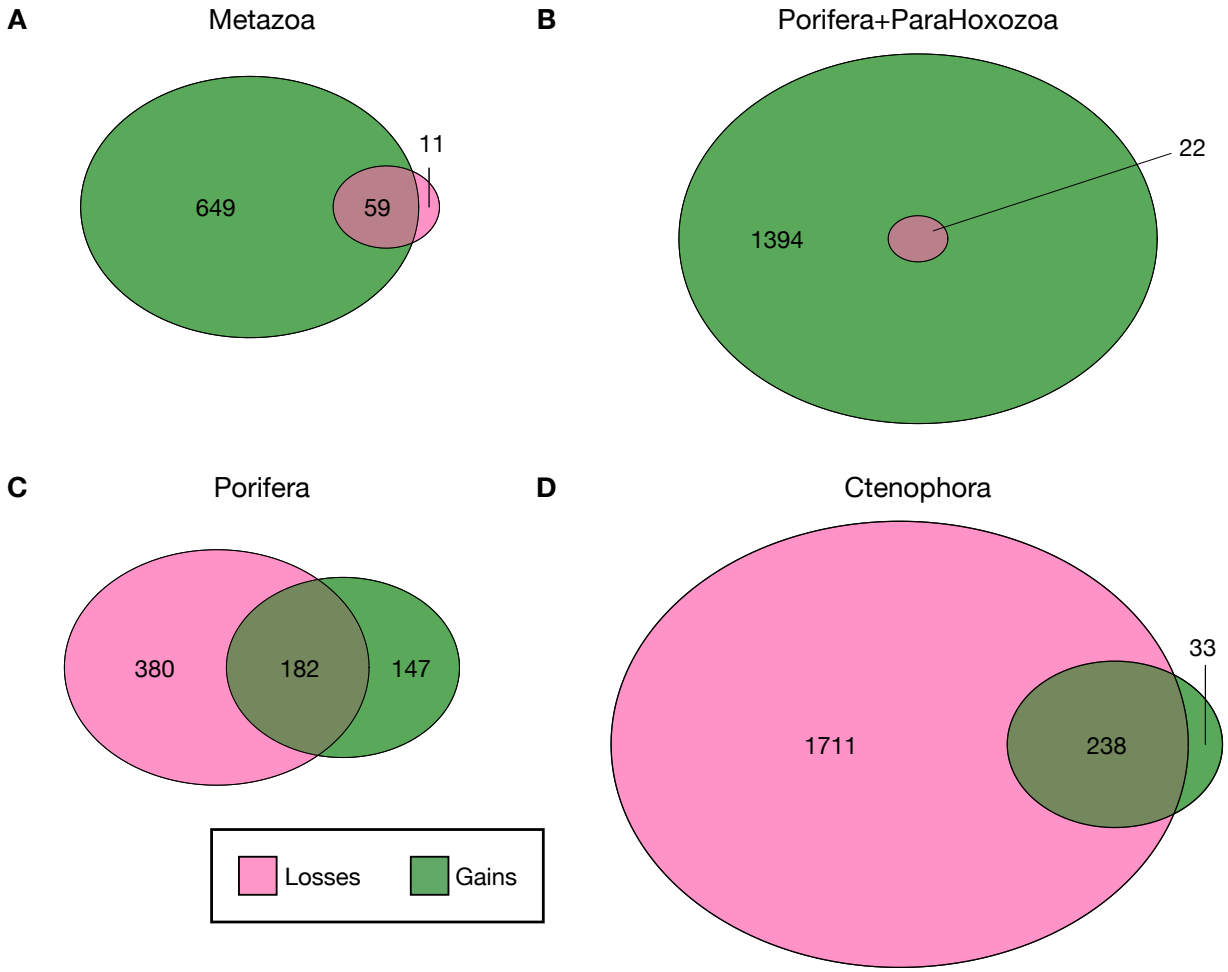

Figure 5: We compared GO terms derived from orthogroups that were gained and lost at important nodes at the start of the Metazoa tree. All gains and losses shown are from the Ctenophora-first topology. In each case, a portion of GO terms from the gains and losses overlapped, and these we excluded from further GO terms analysis. A: Numbers of GO terms for orthogroups gained and lost at the Metazoa node. We further analyzed only the orthogroups gained. B: Numbers of GO terms gained and lost at the Porifera+ParaHoxozoa node. At this node, all GO terms associated with losses were also found amongst the gains. We further analyzed only the orthogroups gained. C: Numbers of GO terms for orthogroups gained and lost at the Porifera node. We further analyzed both orthogroups gained and lost. D: Numbers of GO terms for orthogroups gained and lost at the Ctenophora node. We further analyzed both orthogroups gained and lost.

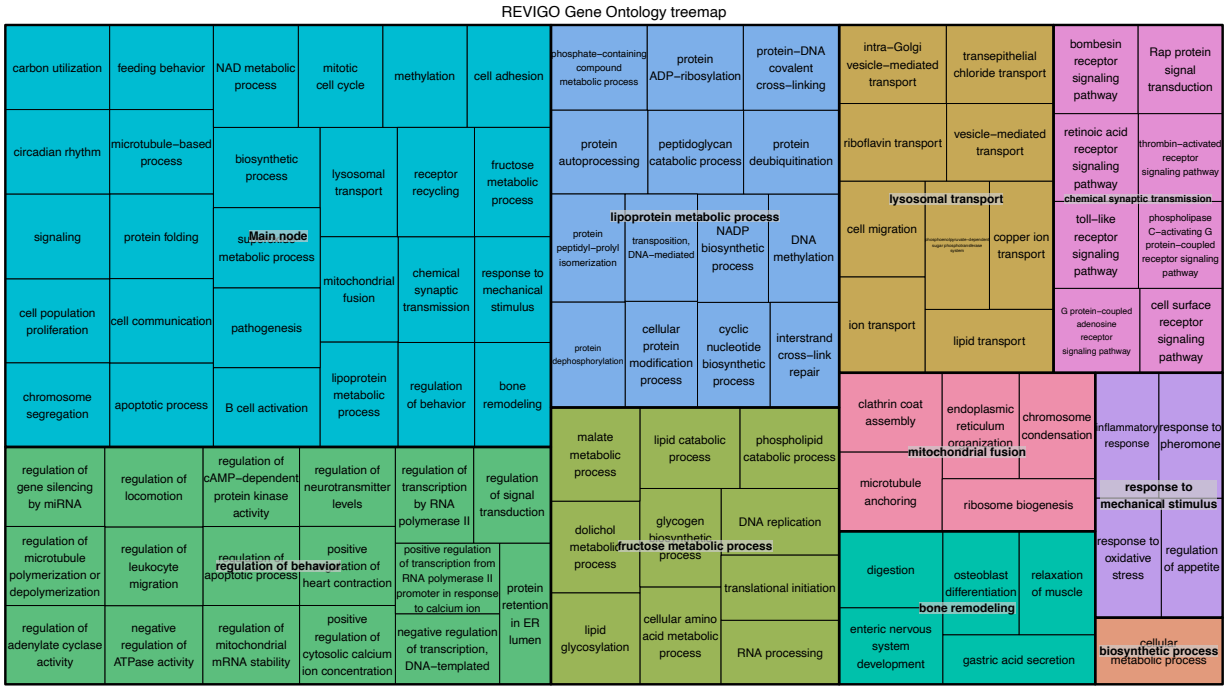

Figure 6: Treemap from Revigo analysis showing GO terms in the biological process category for gains at the Metazoa node in the Ctenophora-first topology.

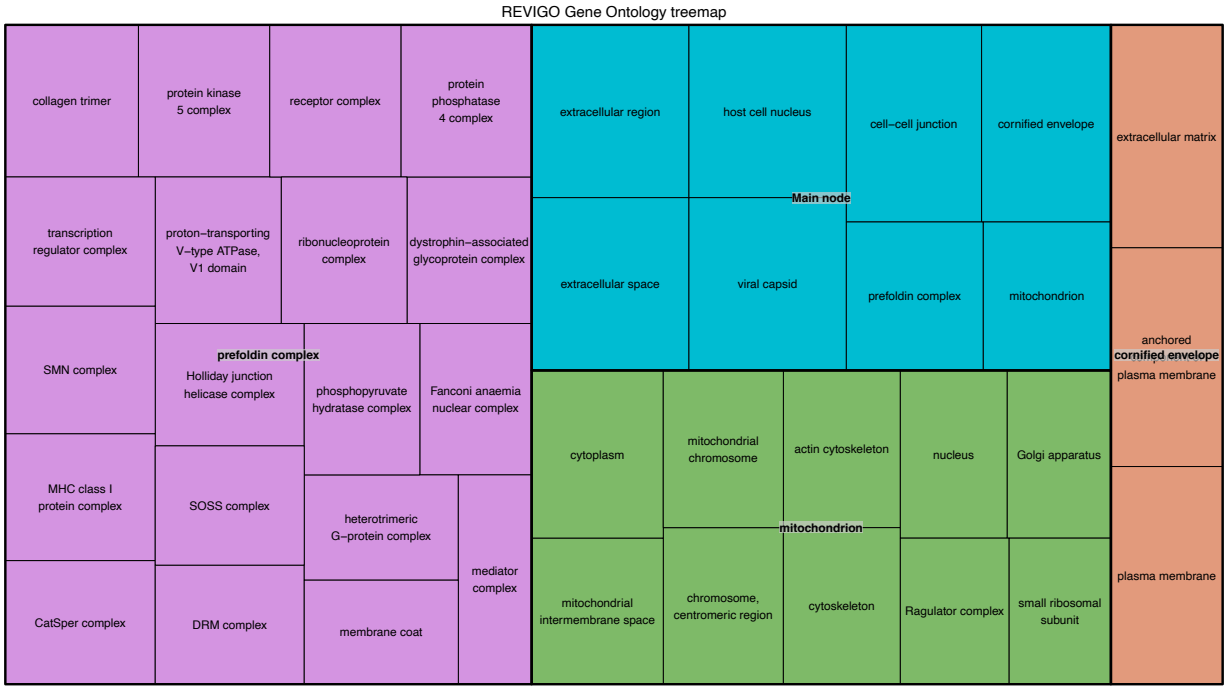

Figure 7: Treemap from Revigo analysis showing GO terms in the cellular component category for gains at the Metazoa node in the Ctenophora-first topology.

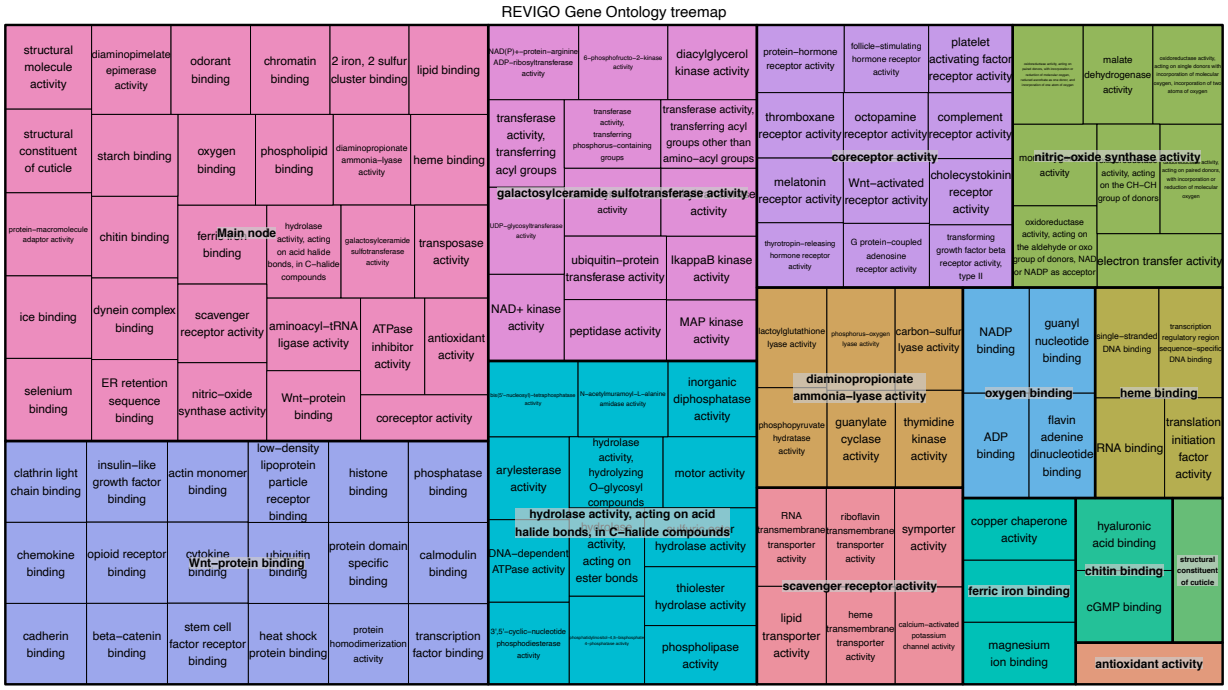

Figure 8: Treemap from Revigo analysis showing GO terms in the molecular function category for gains at the Metazoa node in the Ctenophora-first topology.

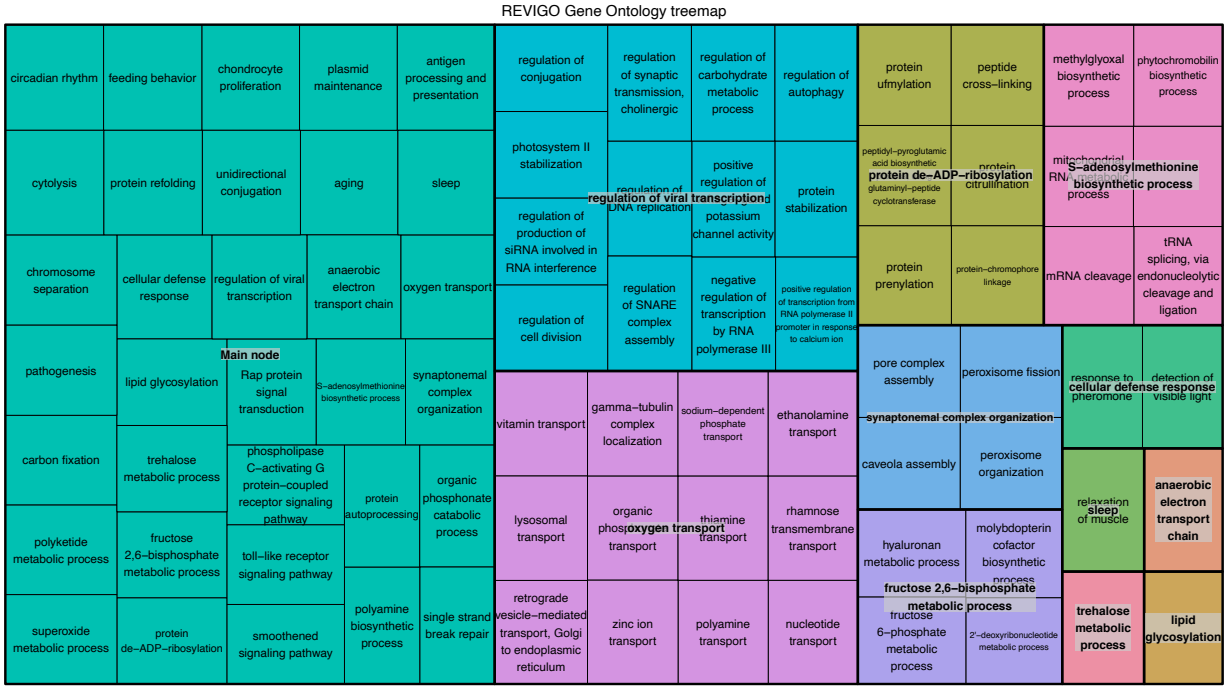

Figure 9: Treemap from Revigo analysis showing GO terms in the biological process category for gains at the Porifera+ParaHoxozoa node in the Ctenophora-first topology.

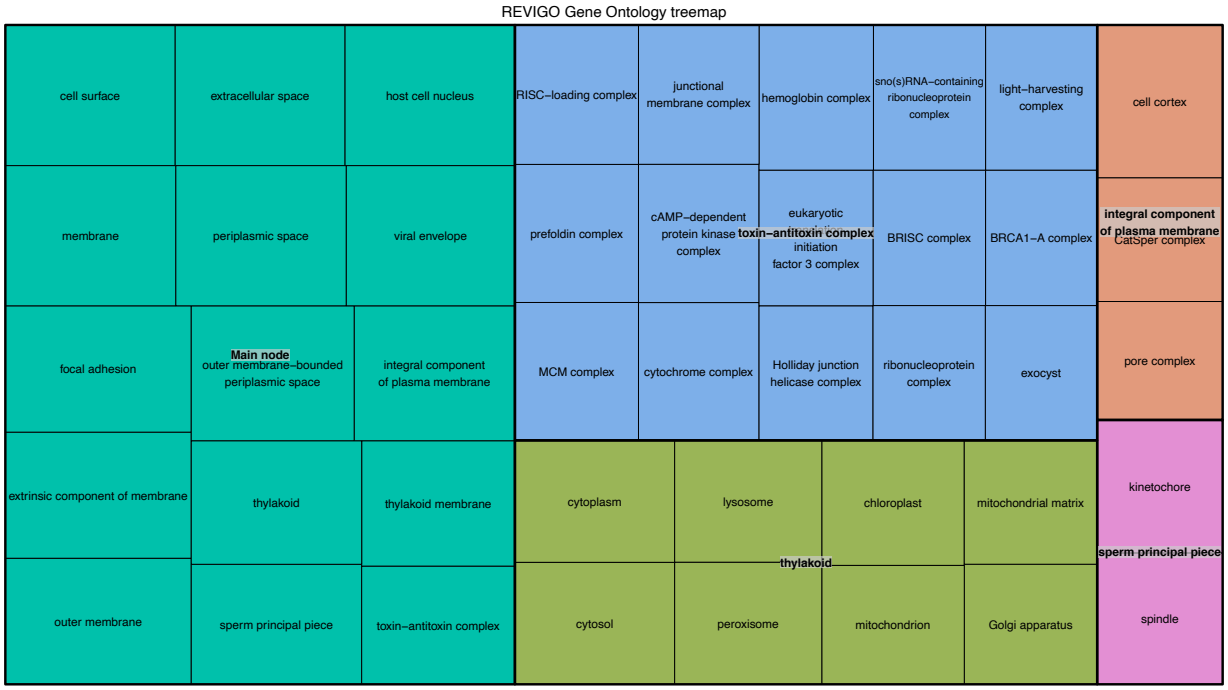

Figure 10: Treemap from Revigo analysis showing GO terms in the cellular component category for gains at the Porifera+ParaHoxozoa node in the Ctenophora-first topology.

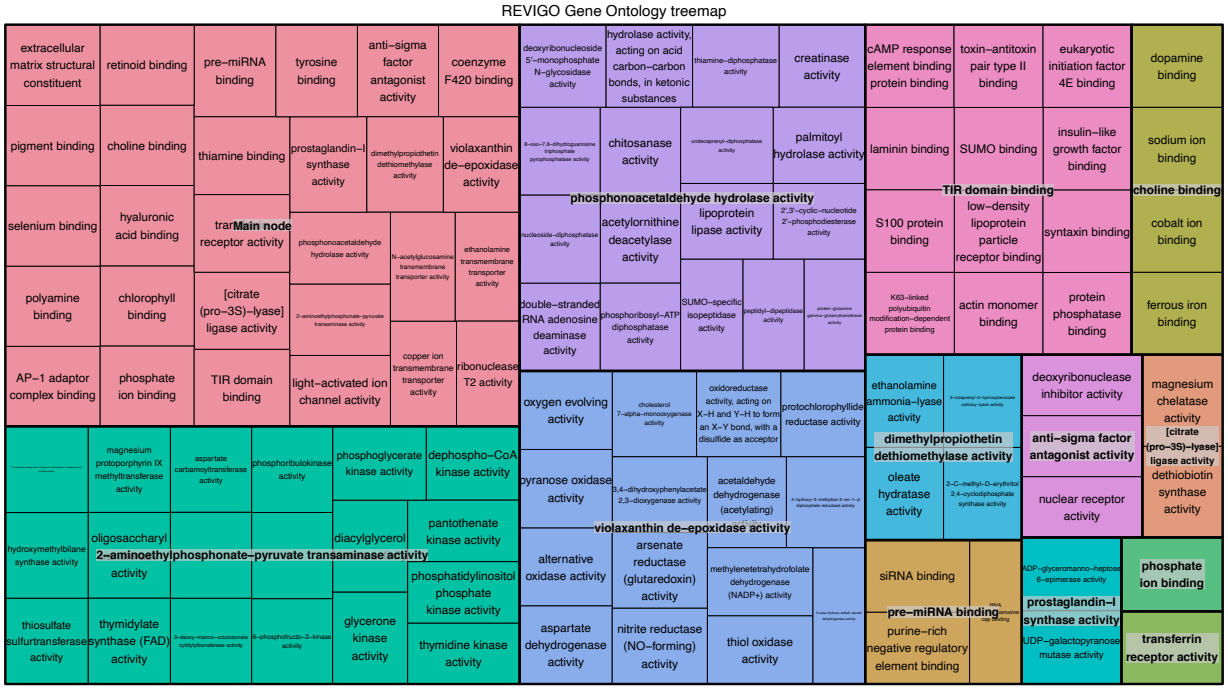

Figure 11: Treemap from Revigo analysis showing GO terms in the molecular function category for gains at the Porifera+ParaHoxozoa node in the Ctenophora-first topology.

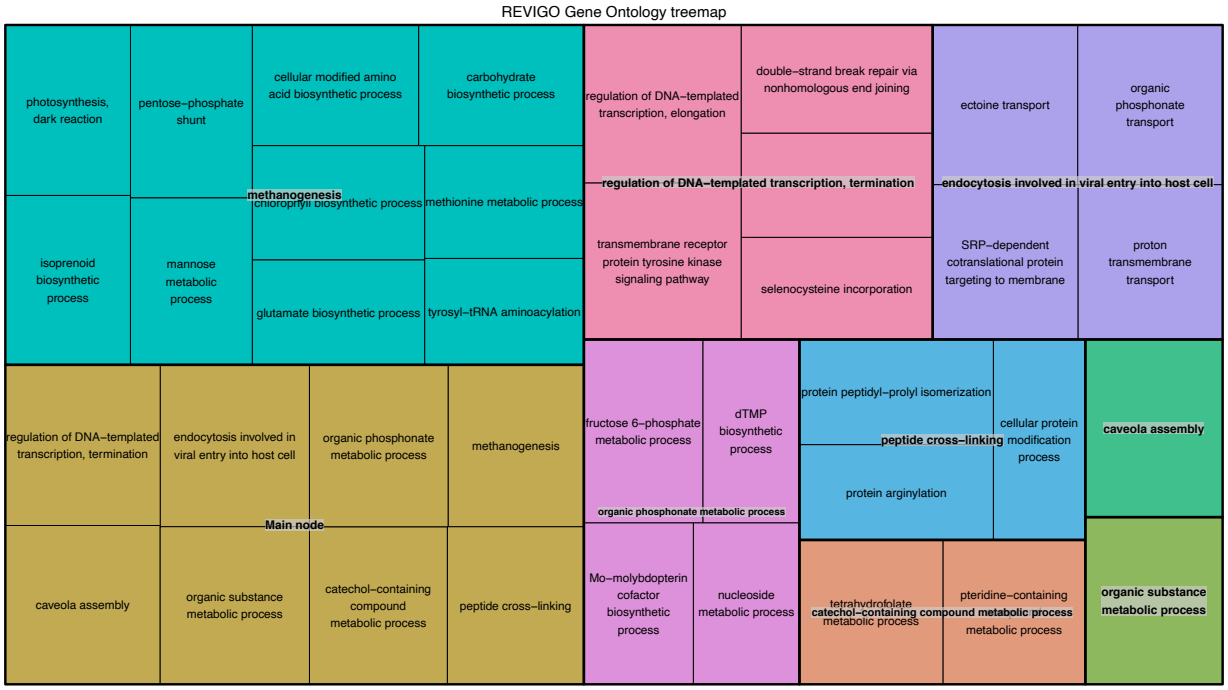

Figure 12: Treemap from Revigo analysis showing GO terms in the biological process category for gains at the Porifera node in the Ctenophora-first topology.

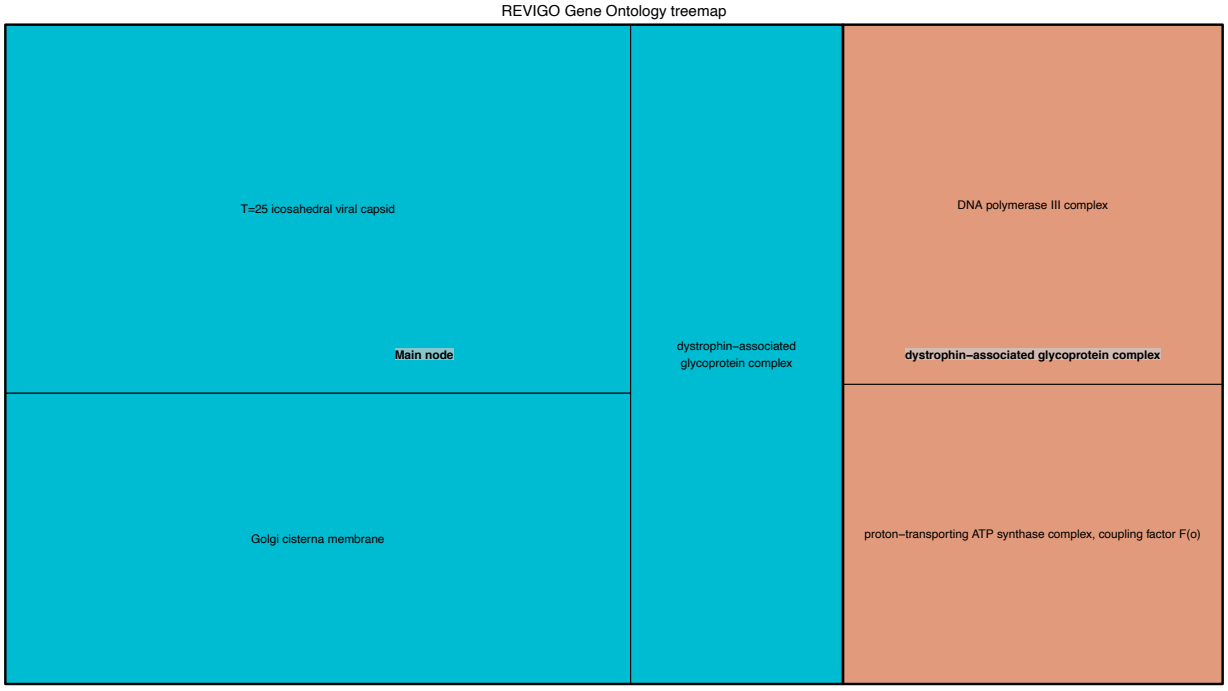

Figure 13: Treemap from Revigo analysis showing GO terms in the cellular component category for gains at the Porifera node in the Ctenophora-first topology.

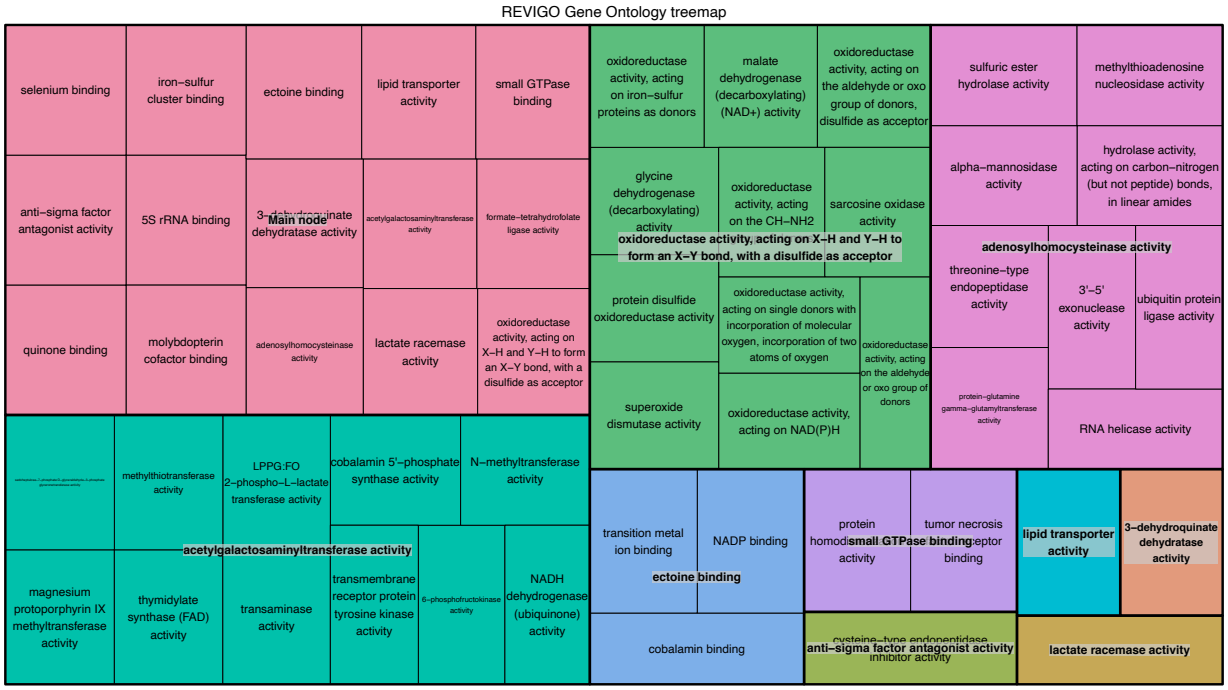

Figure 14: Treemap from Revigo analysis showing GO terms in the molecular function category for gains at the Porifera node in the Ctenophora-first topology.

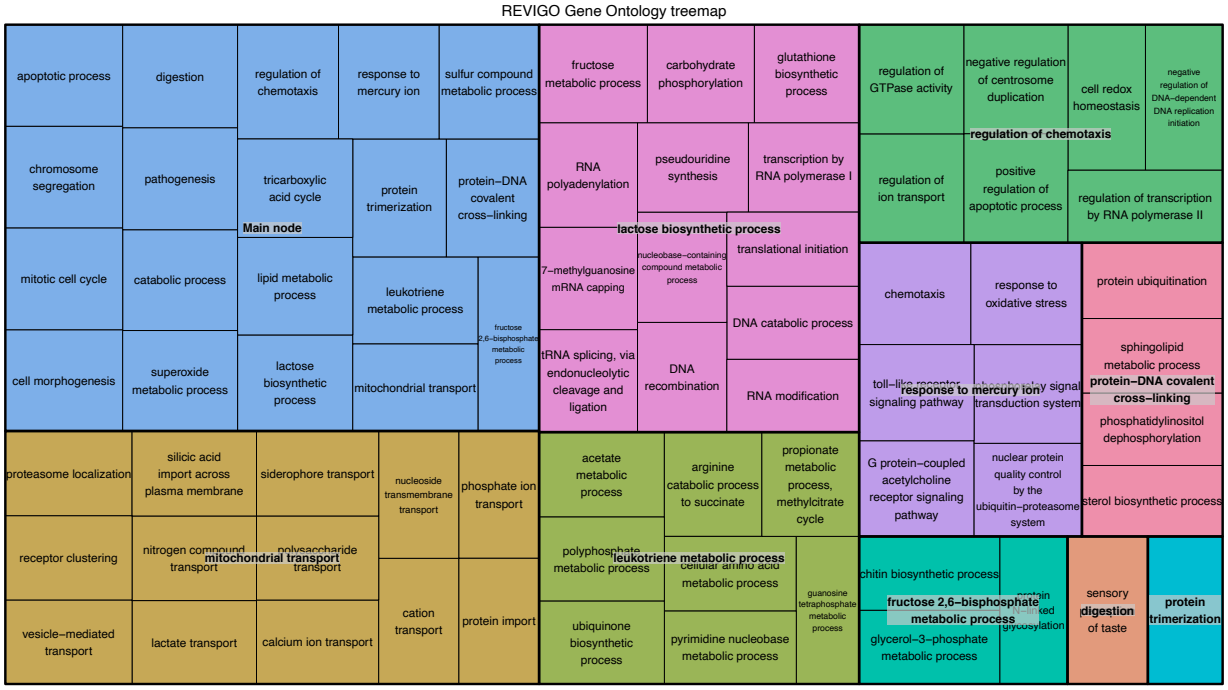

Figure 15: Treemap from Revigo analysis showing GO terms in the biological process category for losses at the Porifera node in the Ctenophora-first topology.

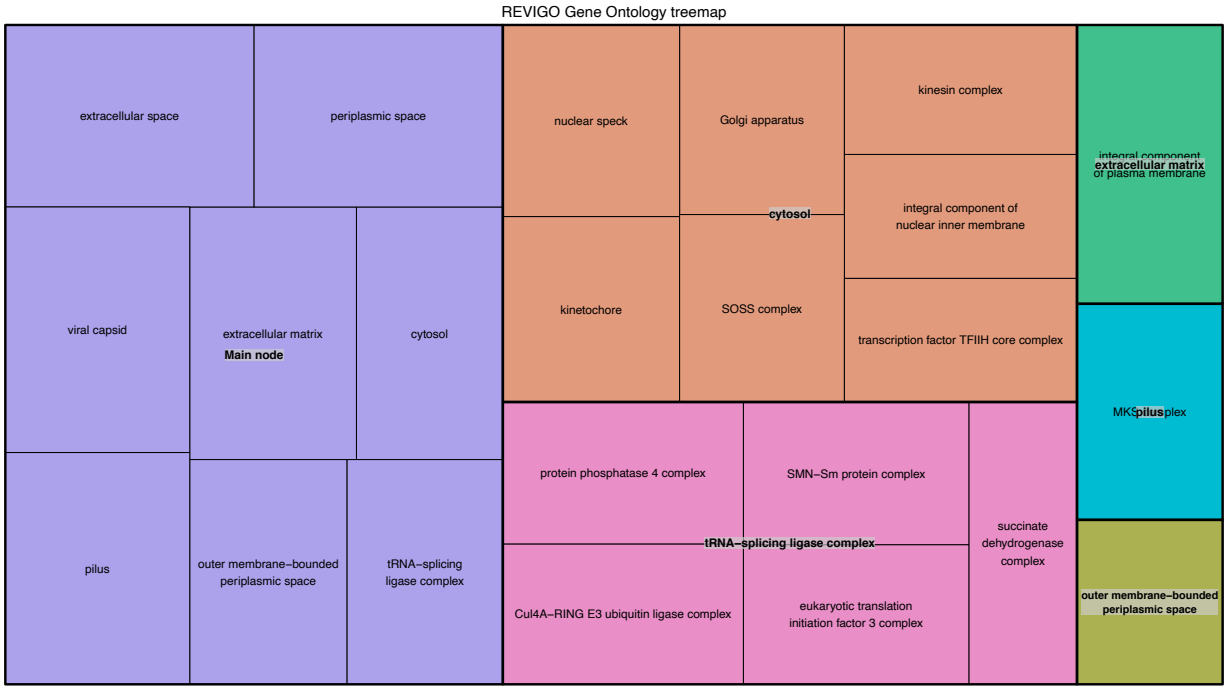

Figure 16: Treemap from Revigo analysis showing GO terms in the cellular component category for losses at the Porifera node in the Ctenophora-first topology.

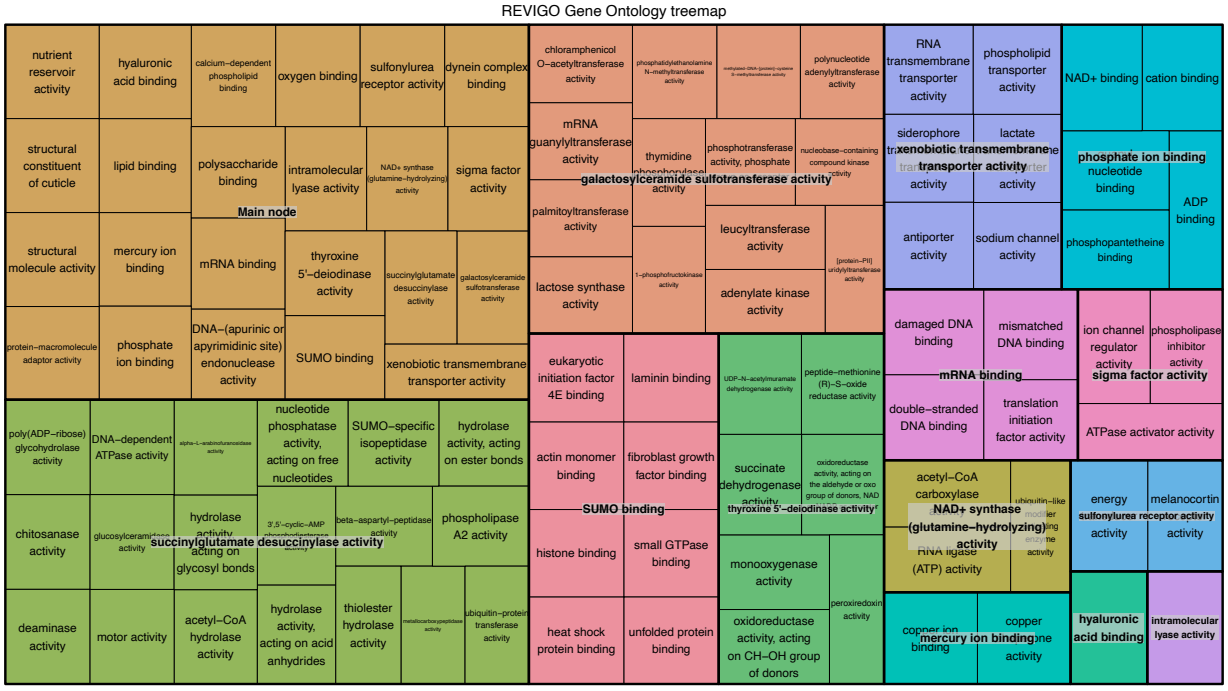

Figure 17: Treemap from Revigo analysis showing GO terms in the molecular function category for losses at the Porifera node in the Ctenophora-first topology.

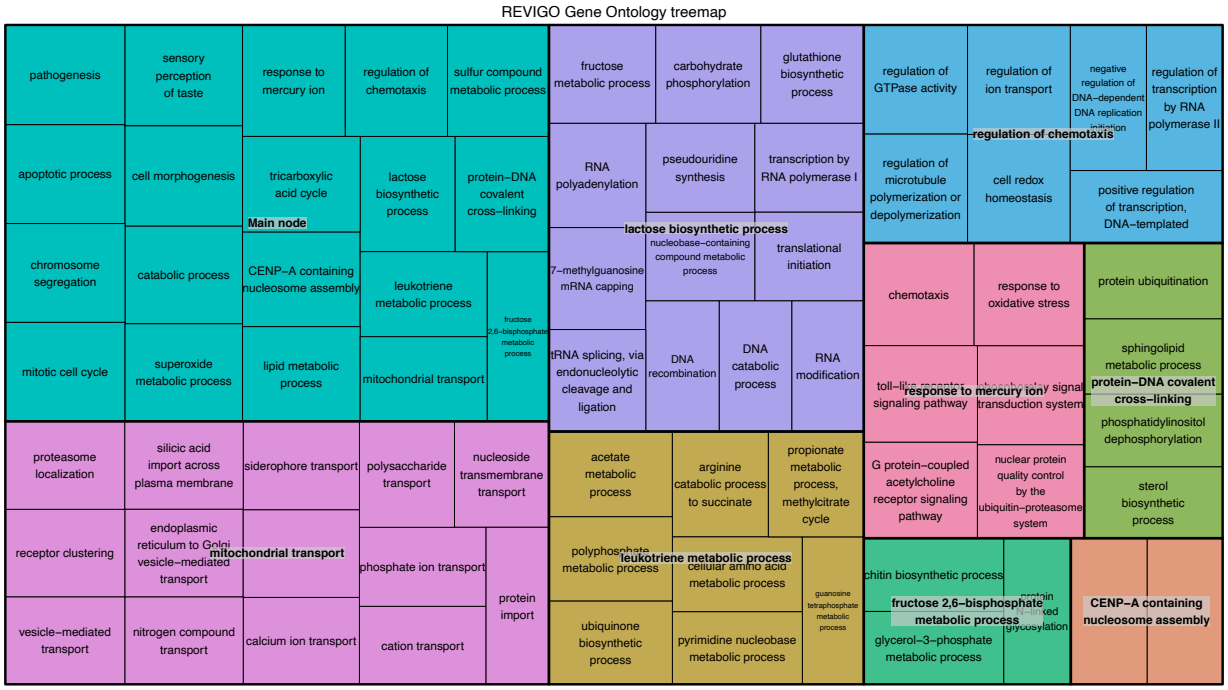

Figure 18: Treemap from Revigo analysis showing GO terms in the biological process category for losses at the Porifera node in the Porifera-first topology.

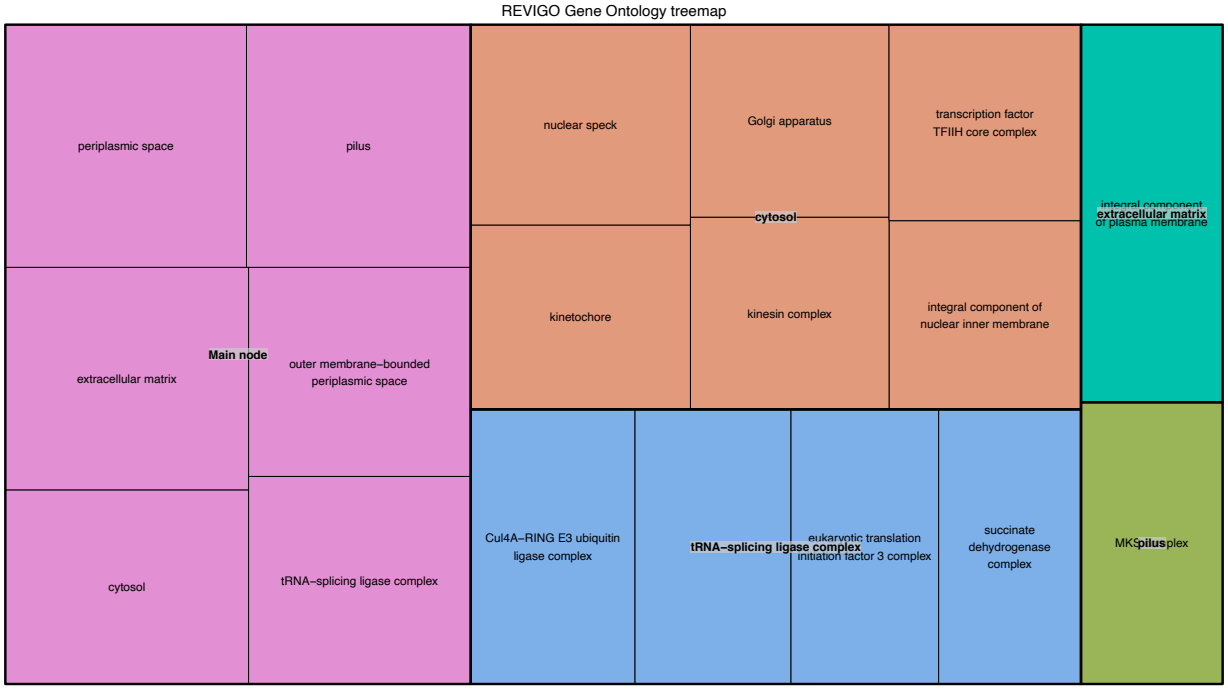

Figure 19: Treemap from Revigo analysis showing GO terms in the cellular component category for losses at the Porifera node in the Porifera -first topology.

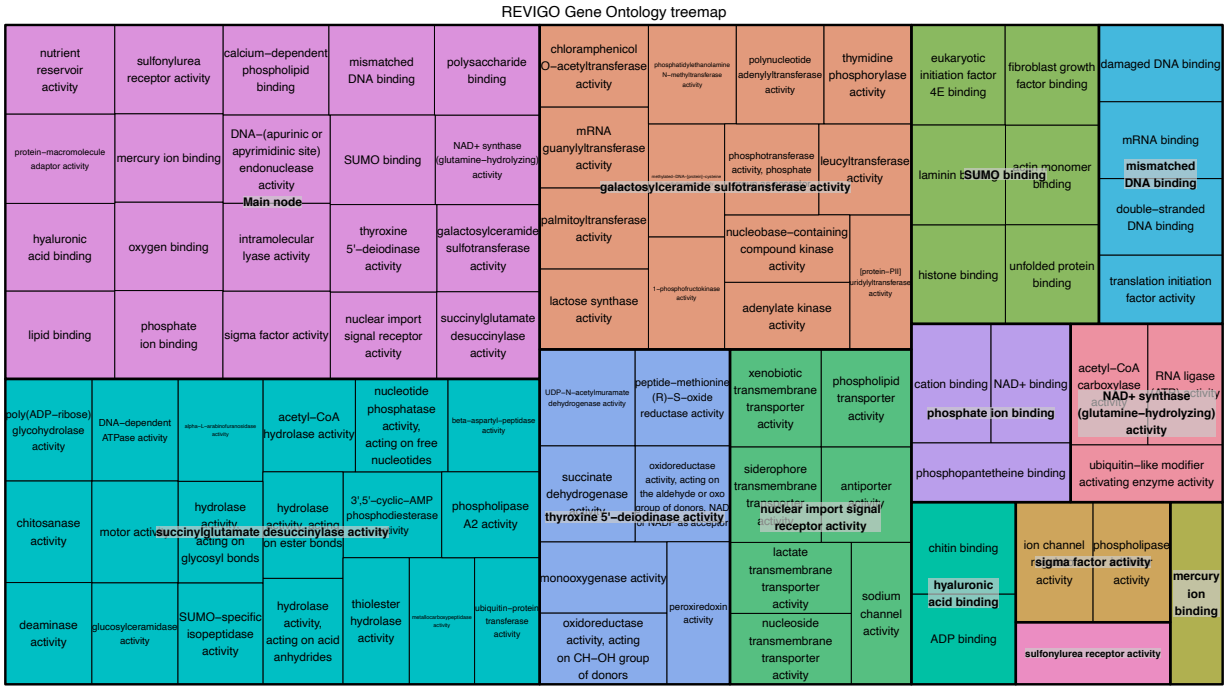

Figure 20: Treemap from Revigo analysis showing GO terms in the molecular function category for losses at the Porifera node in the Porifera -first topology.

Figure 21: Treemap from Revigo analysis showing GO terms in the biological process category for gains at the Ctenophora node in the Ctenophora-first topology.

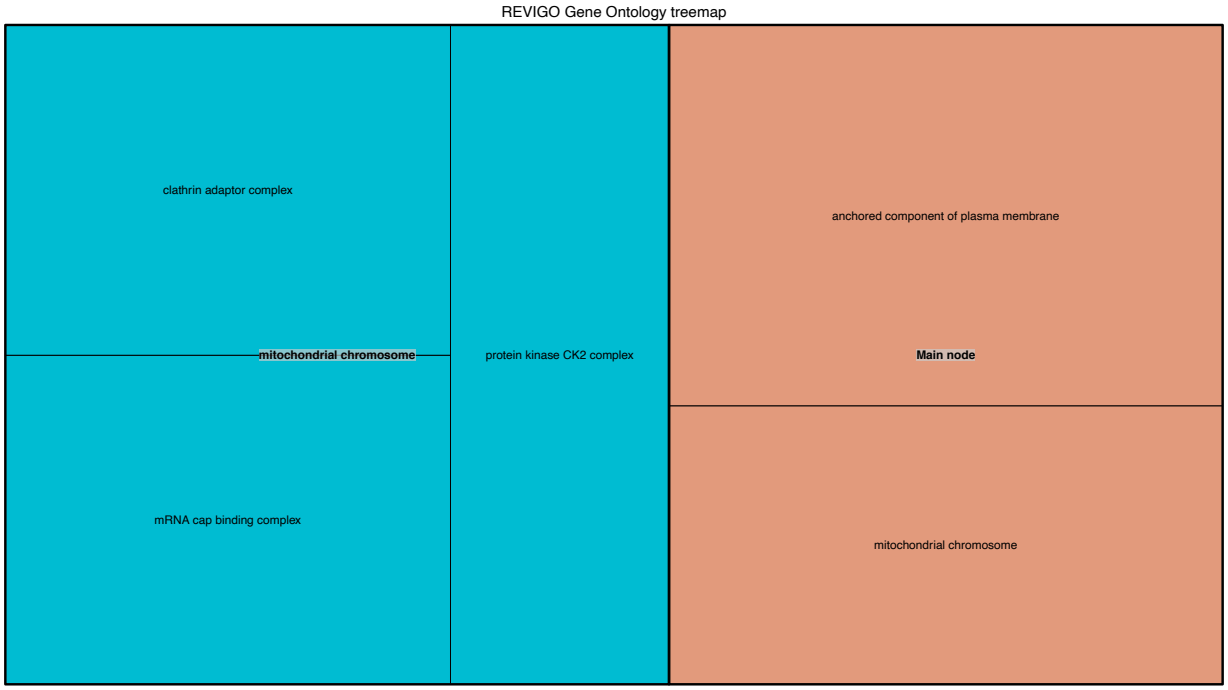

Figure 22: Treemap from Revigo analysis showing GO terms in the cellular component category for gains at the Ctenophora node in the Ctenophora-first topology.

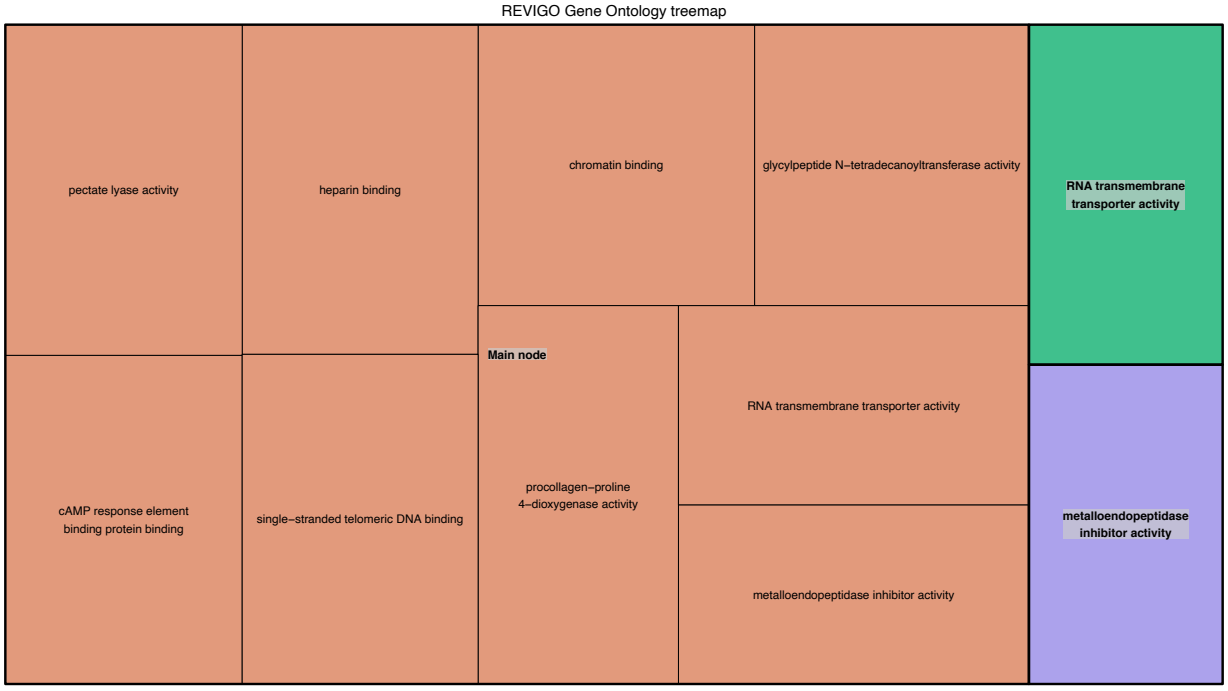

Figure 23: Treemap from Revigo analysis showing GO terms in the molecular function category for gains at the Ctenophora node in the Ctenophora-first topology.

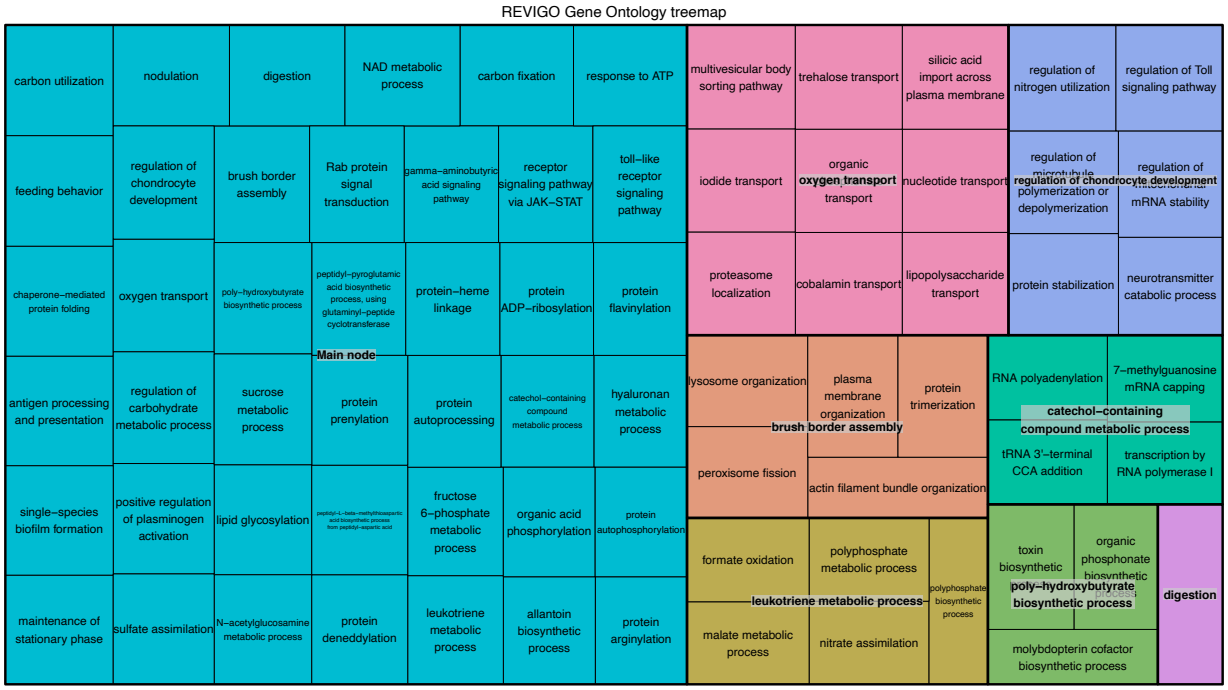

Figure 24: Treemap from Revigo analysis showing GO terms in the biological process category for losses at the Ctenophora node in the Ctenophora-first topology.

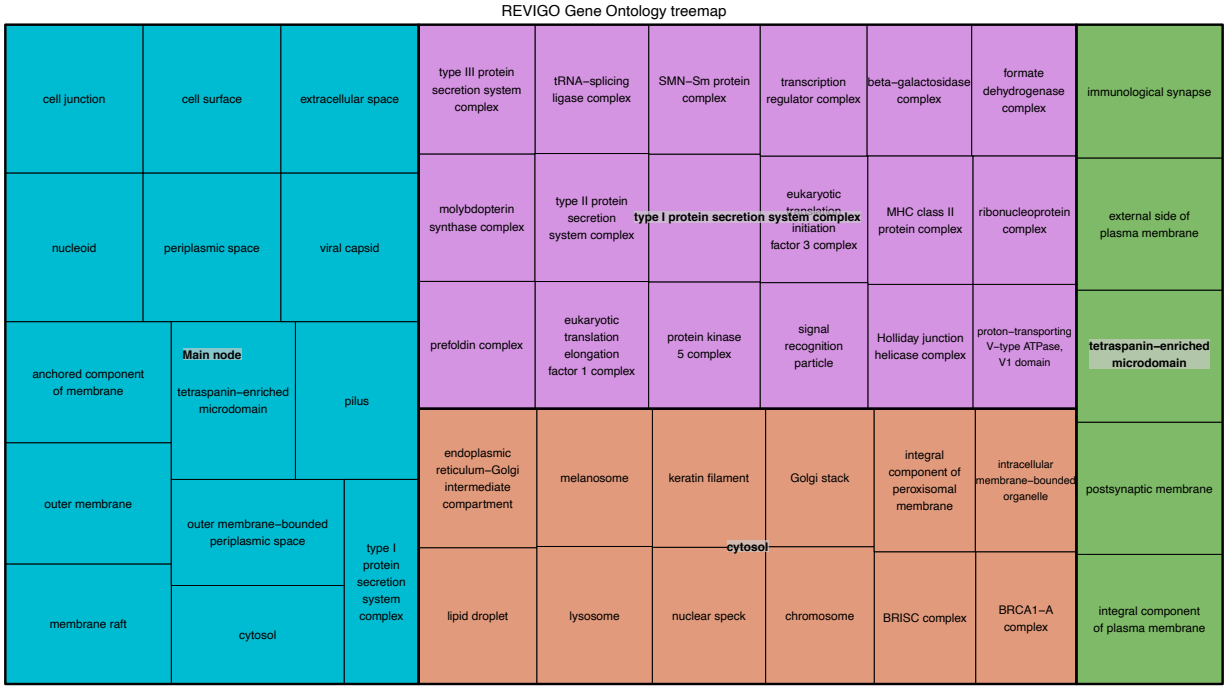

Figure 25: Treemap from Revigo analysis showing GO terms in the cellular component category for losses at the Ctenophora node in the Ctenophora-first topology.

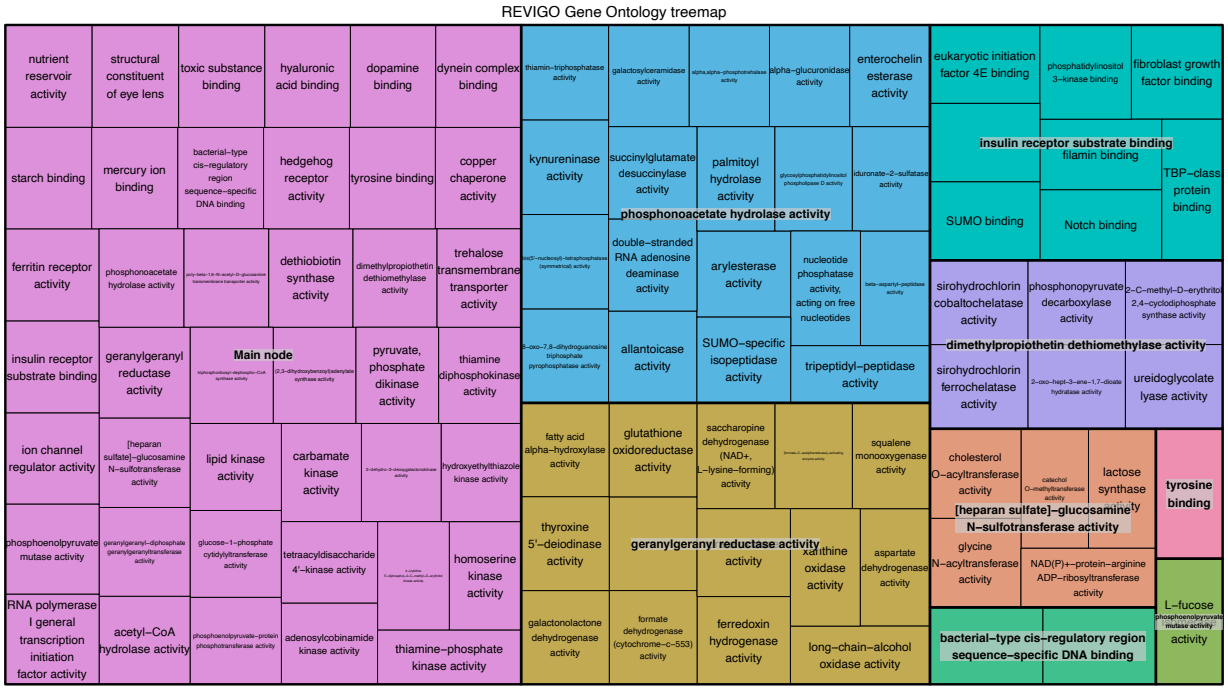

Figure 26: Treemap from Revigo analysis showing GO terms in the molecular function category for losses at the Ctenophora node in the Ctenophora-first topology.

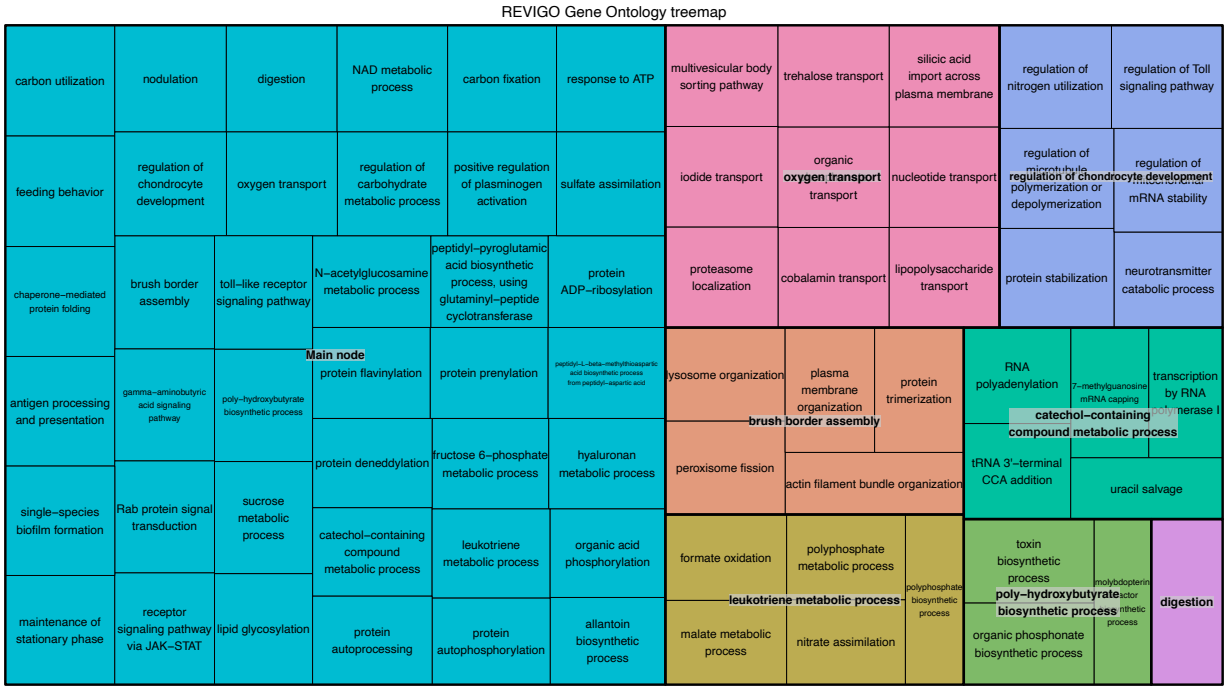

Figure 27: Treemap from Revigo analysis showing GO terms in the biological process category for losses at the Ctenophora node in the Porifera-first topology.

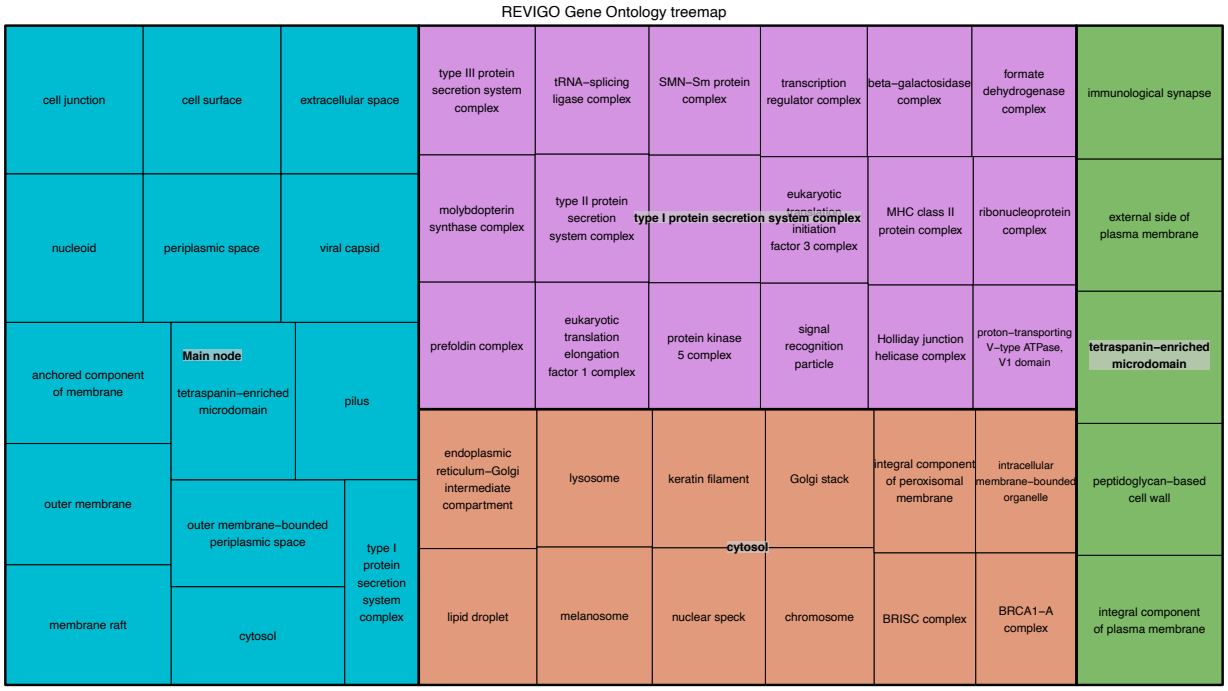

Figure 28: Treemap from Revigo analysis showing GO terms in the cellular component category for losses at the Ctenophora node in the Porifera -first topology.

Figure 29: Treemap from Revigo analysis showing GO terms in the molecular function category for losses at the Ctenophora node in the Porifera -first topology.

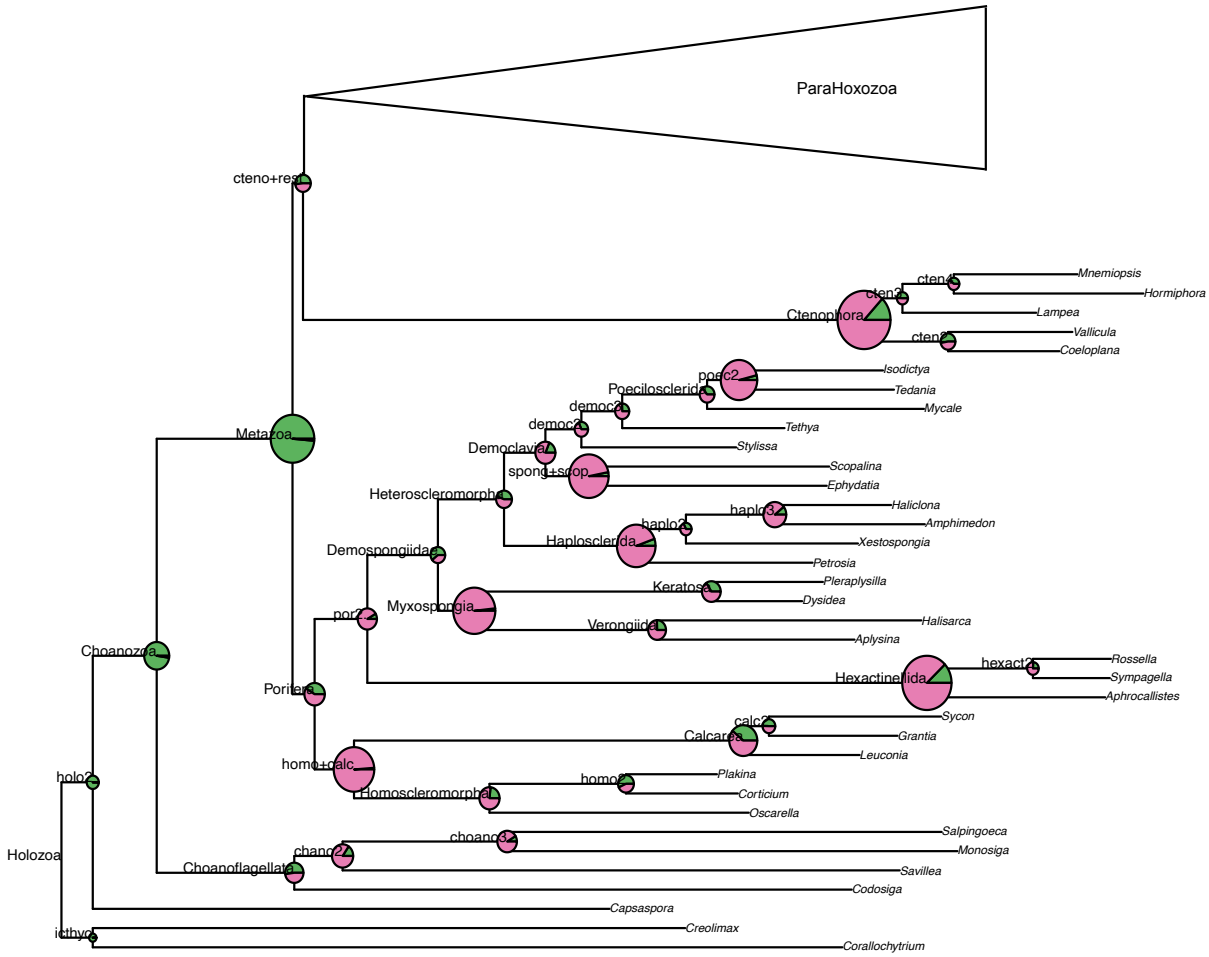

Figure S1: Phylogenomic tree constrained so that Porifera is the first branch of Metazoa. The size of the pie charts on each internal node correspond to the magnitude of change at that node, with green portions representing orthogroups gained, and pink portions representing orthogroups lost, according to Dollo parsimony analysis. The branch leading to ParaHoxozoa has been collapsed for simplicity.

Table S1: Gene Ontology terms highlighted in the text for each node of interest, with orthogroup identities and numbers associated with each term.

Node and direction of change	Gene Ontology term	Associated Orthogroups
Metazoa gain	cell adhesion	OG0003495, OG0000821, OG0002452, OG0008099, OG0006772, OG0000696, OG0063416, OG0045604, OG0003668, OG0000009, OG0029462, OG0009221, OG0069489, OG0018438, OG0000345, OG0000205, OG0008154
	cell-cell junction	OG0001805
	extracellular space	OG0000110, OG0002109, OG0004814, OG0032101, OG0000051, OG0000203
	cell population proliferation	OG0000016
	Wnt protein binding	OG0000067
	coreceptor activity	OG0005525, OG0000009
Porifera+ParaHoxozoa gain	detection of visible light	OG0009705, OG0042173, OG0013200
	regulation of autophagy	OG0048545
	aging	OG0002497
	ion channel regulator activity	OG0008591
Porifera gain	caveola assembly	OG0010034
	endocytosis involved in viral entry into host cell	OG0091630
	ectoine transport	OG0046989
	ectoine binding	OG0046989
Porifera loss	mitotic cell cycle	OG0007776
	apoptotic process	OG0020094
	cell morphogenesis	OG0017665
	extracellular matrix	OG0078217, OG0034281, OG0023295
	hyaluronic acid binding	OG0015041
	vesicle mediated transport	OG0025571, OG0012351
	receptor clustering	OG0032215
	motor activity	OG0026955, OG0046220, OG0028521, OG0034180, OG0026770
Ctenophora gain	mitotic spindle assembly	OG0026271
	clathrin adaptor complex	OG0037883
	RNA transmembrane transporter activity	OG0037656, OG0046401
Ctenophora loss	digestion	OG0006549
	brush border assembly	OG0005896
	insulin receptor substrate binding	OG0001466

References

1. Babonis LS, Martindale MQ. Old Cell, new trick? Cnidocytes as a model for the evolution of novelty. Integr Comp Biol. 2014;54(4):714–22.

2. Paps J, Holland PWH. Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. Nat Commun [Internet]. 2018;9(1):1–8. Available from: http://dx.doi.org/10.1038/s41467-018-04136-5

3. Hoballah ME, Gu T, Stuurman J, Broger L, Barone M, Mandel T, et al. Single Gene – Mediated Shift in Pollinator Attraction in Petunia. 2007;19(March):779–90.

4. Drouin G, Godin J, Pagé B. The Genetics of Vitamin C Loss in Vertebrates. 2011;371–8.

5. White JK, Gerdin AK, Karp NA, Ryder E, Buljan M, Bussell JN, et al. Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. Cell. 2013;154(2):452.

6. Blomen VA, Májek P, Jae LT, Bigenzahn JW, Nieuwenhuis J, Staring J, et al. Gene essentiality and synthetic lethality in haploid human cells. Science (80-). 2015;350(6264):1092–6.

7. Félix MA, Barkoulas M. Pervasive robustness in biological systems. Nat Rev Genet [Internet]. 2015;16(8):483–96. Available from: http://dx.doi.org/10.1038/nrg3949

8. Papp B, Notebaart RA, Pál C. Systems-biology approaches for predicting genomic evolution. Nat Rev Genet [Internet]. 2011;12(9):591–602. Available from: http://dx.doi.org/10.1038/nrg3033

9. Albalat R, Cañestro C. Evolution by gene loss. Nat Rev Genet [Internet]. 2016;17:379–91. Available from: http://dx.doi.org/10.1038/nrg.2016.39

10. Jeffrey WR. Regressive evolution in Astyanax cavefish. Annu Rev Genet. 2008;23(1):1–7.

11. Yamamoto Y, Byerly MS, Jackman WR, Jeffery WR. Pleiotropic functions of embryonic sonic hedgehog expression link jaw and taste bud amplification with eye loss during cavefish evolution. Dev Biol [Internet]. 2009;330(1):200–11. Available from: http://dx.doi.org/10.1016/j.ydbio.2009.03.003

12. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, et al. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature. 2008;452(7188):745–9.

13. Ryan JF, Pang K, Schnitzler CE, Nguyen A-D, Moreland RT, Simmons DK, et al. The Genome of the Ctenophore Mnemiopsis leidyi and Its Implications for Cell Type Evolution. Science (80-) [Internet]. 2013;342(6164):1242592–1242592. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1242592

14. Richter DJ, Fozouni P, Eisen MB, King N. Gene family innovation, conservation and loss on the animal stem lineage. Elife [Internet]. 2018;7:1–43. Available from: https://elifesciences.org/articles/34226

15. Fernandez R, Gabaldon T. Gene gain and loss across the metazoan tree of life. Nat Ecol Evol. 2020;4(4):524–33.

16. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658–9.

17. MacManes MD. The Oyster River Protocol : a multi-assembler and kmer approach for de novo transcriptome assembly. PeerJ. 2018;6(e5428):1–18.

18. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. TransRate: reference free quality assessment of de-novo transcriptome assemblies. Genome Res. 2016;26.

19. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.

20. Borowiec ML, Lee EK, Chiu JC, Plachetzki DC. Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. BMC Genomics [Internet]. 2015;16(2015):987. Available from: http://dx.doi.org/10.1186/s12864-015-2146-4

21. Dunn C, Smith S, Ryan J. Gblockswrapper [Internet]. Bitbucket; 2009. Available from: https://bitbucket.org/caseywdunn/labcode/src/master/scripts_phylogenomics_21Feb2009/Gblockswrapper

22. Emms DM, Kelly S. OrthoFinder : solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol [Internet]. 2015;16(157):1–14. Available from: http://dx.doi.org/10.1186/s13059-015-0721-2

23. Emms DM, Kelly S. OrthoFinder : phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(238):1–14.

24. Kocot KM, Citarella MR, Moroz LL, Halanych KM. PhyloTreePruner: A phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. Evol Bioinforma. 2013;2013(9):429–35.

25. Kayal E, Bentlage B, Sabrina Pankey M, Ohdera AH, Medina M, Plachetzki DC, et al. Phylogenomics provides a robust topology of the major cnidarian lineages and insights on the origins of key organismal traits. BMC Evol Biol. 2018;18(1):1–18.

26. Nguyen L, Schmidt HA, Haeseler A Von, Minh BQ. IQ-TREE : A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol. 2014;32(1):268–74.

27. Friedrich M, Tautz D. Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. Nature. 1995;376(6536):165–7.

28. Scholtz G, Mittmann B, Gerberding M. The pattern of Distal-less expression in the mouthparts of crustaceans, myriapods and insects: New evidence for a gnathobasic mandible and the common origin of Mandibulata. Int J Dev Biol. 1998;42(6):801–10.

29. Riutort M, Álvarez-Presas M, Lázaro E, Solà E, Paps J. Evolutionary history of the Tricladida and the platyhelminthes: An up-to-date phylogenetic and systematic account. Int J Dev Biol. 2012;56(1–3):5–17.

30. Giribet G, Edgecombe GD. The Phylogeny and Evolutionary History of Arthropods. Curr Biol [Internet]. 2019;29(12):R592–602. Available from: https://doi.org/10.1016/j.cub.2019.04.057

31. Shimodaira H. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 2002;51(3):492–508.

32. R Core Team. R: a language and environment for statistical computing [Internet]. Vienna, Austria; 2018. Available from: https://www.r-project.org/

33. Ryan JF. Alien Index: identify potential non-animal transcripts or horizontally transferred genes in animal transcriptomes. 2014.

34. Plachetzki DC, Pankey MS, MacManes MD, Lesser MP, Walker CW. The Genome of the Softshell Clam Mya arenaria and the Evolution of Apoptosis. Genome Biol Evol. 2020;12(10):1681–93.

35. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460–1.

36. Jones P, Binns D, Chang H, Fraser M, Li W, Mcanulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236–40.

37. Supek F, Bošnjak M, Škunca N, Šmuc T. Revigo summarizes and visualizes long lists of gene ontology terms. PLoS One. 2011;6(7).

38. Lartillot N, Lepage T, Blanquart S. PhyloBayes 3: A Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics. 2009;25(17):2286–8.

39. King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, et al. The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. Nature [Internet]. 2008;451(7180):783–8. Available from: http://dx.doi.org/10.1038/nature06617

40. Suga H, Chen Z, Mendoza A De, Sebe-Pedros A, Brown MW, Kramer E, et al. The Capsaspora genome reveals a complex unicellular prehistory of animals. Nat Commun. 2013;4(2325):1–9.

41. Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier MEA, Mitros T, et al. The

Amphimedon queenslandica genome and the evolution of animal complexity. Nature [Internet]. 2010;466(7307):720–6. Available from: http://dx.doi.org/10.1038/nature09201

42. Simion P, Phillippe H, Baurain D, Jager M, Richter DJ, Di Franco A, et al. A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. Curr Biol. 2017;27:1–10.

43. Pisani D, Pett W, Dohrmann M, Feuda R, Rota-Stabelli O, Philippe H, et al. Genomic data do not support comb jellies as the sister group to all other animals. Proc Natl Acad Sci [Internet]. 2015;112(50):15402–7. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1518127112

44. Feuda R, Dohrmann M, Pett W, Philippe H, Rota-Stabelli O, Lartillot N, et al. Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals. Curr Biol. 2017;27(24):3864-3870.e4.

45. Dunn CW, Leys SP, Haddock SHD. The hidden biology of sponges and ctenophores. Trends Ecol Evol [Internet]. 2015;30(5):282–91. Available from: http://dx.doi.org/10.1016/j.tree.2015.03.003

46. Leys SP, Meech RW. Physiology of coordination in sponges. Can J Zool. 2006;84(2):288–306.

47. Ryan JF, Chiodin M. Where is my mind? How sponges and placozoans may have lost neural cell types. Philos Trans R Soc B Biol Sci. 2015;370(1684).

48. Flórez L V., Biedermann PHW, Engl T, Kaltenpoth M. Defensive symbioses of animals with prokaryotic and eukaryotic microorganisms. Nat Prod Rep. 2015;32(7):904–36.

49. Pascelli C, Laffy PW, Botté E, Kupresanin M, Rattei T, Lurgi M, et al. Viral ecogenomics across the Porifera. Microbiome. 2020;8(1):1–22.

50. Webster NS, Thomas T. The sponge hologenome. Am Soc Microbiol. 2016;7(2):1–14.

51. Whelan N V., Kocot KM, Moroz LL, Halanych KM. Error, signal, and the placement of Ctenophora sister to all other animals. Proc Natl Acad Sci [Internet]. 2015;112(18):5773–8. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1503453112

52. Whelan N V, Kocot KM, Moroz TP, Mukherjee K, Williams P, Paulay G, et al. Ctenophore relationships and their placement as the sister group to all other animals. Nat Ecol Evol [Internet]. 2017;1–10. Available from: http://dx.doi.org/10.1038/s41559-017-0331-3

53. Natsidis P, Kapli P, Schiffer PH, Telford MJ. Systematic errors in orthology inference and their effects on evolutionary analyses. iScience [Internet]. 2021;24(2):102110. Available from: https://doi.org/10.1016/j.isci.2021.102110