UNIVERSITY OF CALIFORNIA

SANTA CRUZ

CHROMOSOME-SCALE GENOMES, RESOLVING THE SISTER PHYLUM TO OTHER ANIMALS, AND NOVEL BIOLUMINESCENT SYSTEMS

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Darrin T. Schultz

September 2021

The Dissertation of Darrin T. Schultz
is approved:

_____
Professor Richard E. Green

_____
Professor Steven H.D. Haddock

_____
Professor David Haussler

_____
Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Chromosome-Scale Genomes, Resolving the Sister Phylum to Other Animals, and Novel Bioluminescent Systems.

by

Darrin T. Schultz

Understanding how animals evolved from unicellular life requires comprehensive analyses that sample all animal phyla. However, some major outstanding questions in evolutionary biology, such as the evolutionary provenance of neurons, developmental pathways, and animal-specific genes, remain unanswered for one reason: it is unclear whether sponges (phylum Porifera) or ctenophores (phylum Ctenophora) as the sister phylum to all other animals. Here, we resolved this question using a chromosome-scale, whole-genome comparative approach. First, we generated chromosome-scale genomes of ctenophore species, and of three species that are unicellular, and outgroups to the Metazoa. By comparing these genomes to other chromosome-scale animal genomes, we identified groupings of genes that have persisted together on the same chromosomes since the common ancestor of the Filozoa, more than one billion years ago. We track the fate of these linked genes in the genomes of extant species, and find irreversible chromosomal fusions-with-mixing that preclude sponges from being the sister phylum to other animals. Thus, ctenophores must be the sister phylum to other animals.

As a parallel effort, we sought to use transcriptomics and genomics to study a trait that is common to many of the marine species that we studied above: bioluminescence. Light-emitting luciferase proteins are a useful tool for visualizing sub-cellular processes in microscopy, and are an interesting case of convergent evolution in over fifty clades. We used transcriptomics, full-length cDNA sequencing, and protein purification techniques to identify a novel luciferase in the polychaete worm *Odontosylllis undecimdonta*. This luciferase appears to be specific to the genus *Odontosyllis*, and there is no evidence for homologs in the transcriptomes of other polychaetes. In addition to the polychaete luminescence, we also identified luminescence in a species of deep-sea sponge. This finding is significant, as reports of sponge luminescence in the past have been dubious. We used a biochemical approach to identify the luminous small molecule, coelenterazine, that is used in the bioluminescence reaction. A metagenomics sequencing approach revealed that the sponge sample contained little to no bacteria, and therefore the luminescence produced by the sponge was likely endogenous. Future efforts will focus on genome sequencing, and identifying the luciferase, to determine if the luciferase is truly encoded in the sponge genome.

Acknowledgements

I would like to thank my committee, Dr. David Haussler, Dr. Richard E. Green, and Dr. Steven H.D Haddock for their guidance and support throughout my graduate education. In addition, I would like to thank Dr. Beth Shapiro and Dr. Ed Green for graciously allowing me to work on ctenophore genomics in their paleogenomics lab. A special thanks goes to Dr. Steven Haddock for his endless support in my pursuit of myriad research questions about deep-sea organisms.

So many other colleagues and mentors have been generous with their time, knowledge, support, and advice, without which I certainly would have failed during my PhD. Thank you, Dr. Warren Francis, Jacob Winnikoff, Dr. Séverine Martini, Dr. Joseph Ryan, Dr. Manabu Bessho-Uehara, Jordan Eizenga, Dr. Erik Thuesen, Dr. Russell Corbett-Detig, Dr. Chris Vollmers, Dr. Nathan Schaeffer, Dr. Edward Rice, Dr. Brendan O'Connell, Dr. Merly Escalona, Dr. Nedda Saremi, Joshua Kapp, Jonas Oppenheimer, Dr. Ilia Yampolsky, Dr. Yuichi Oba, Dr. Alexey Kotlobay, Lynne Christianson, Shannon Johnson, Dr. Claudia Mills, and Dr. Nathan Shaner.

I am fortunate to have had financial support from multiple sources, without which this research would have not been possible. These funding sources include the Packard Foundation, the National Science Foundation, the Institute for International Education, and the UC Santa Cruz Biomolecular Engineering and Bioinformatics department.

The work presented in this thesis required extraordinary expertise and assistance in animal collection by blue-water and in-shore scuba diving, deep-sea-diving robots, deep-sea trawling, and general seamanship. Thank you to the crew, ROV pilots, and captains of the MBARI research vessels *RV Rachel Carson* and *RV Western Flyer* for their help over the 100+ days I spent at sea as a graduate student.

I appreciate the assistance of my colleagues, mentors, and friends in collecting the animals used in these experiments. Thank you to Jacob Winnikoff, Shannon Johnson, Dr. Alejandro Damian-Serrano, Dr. Katie Thomas, and all of my other blue water dive buddies through the years. Thank you to everyone who helped teach me how to trawl for samples, and how to identify the animals swarming in the trawl buckets. Thank you, Dr. George Matsumoto, and Dr. Erik Thuesen.

Special thanks to the team at the Monterey Bay Aquarium for their passion and dedication to not only culturing ctenophores, but in designing new equipment and approaches to completing their life cycle in captivity. Thank you, Wyatt Patry, Thomas Knowles, MacKenzie Bubel, and Cypress Hansen.

Much of my field work was completed while scuba diving in California, Hawaii, and Australia. However, I did not know how to dive, let alone swim, when I started my PhD. So, I would like to thank Cecilia Shin, Steve Clabuesch, and Dave Benet for training me to be a safe and competent diver. Thank you to all of my classmates for their camaraderie as I navigated the life aquatic, and a big thank-you to Micael Nunez for attempting to teach me how to swim.

Lastly, I would like to thank all of the other people in my academic and personal life who provided logistical, emotional, or material support during my time as a PhD candidate. I especially could not have completed my PhD without the support of my parents, Anita Henninger and Todd Schultz, my grandparents Paulette and James Paine, or the continual support of my partner Micael Nunez.

Dedication

This dissertation is dedicated to my family.

You believed in me. You gave me the space and means to pursue my dreams.

Our ancestors emigrated in hopes for a better chance at life,

and now I am the first in our family to complete a PhD.

This dissertation is an homage to your labor and love.

Published Works Statement

Some of the text in this manuscript has been adapted from the published articles listed below. None of the text in this dissertation falls under the copyrights of the journals in which these works are published.

Darrin T. Schultz[†],Warren R. Francis[†], Jakob D. McBroome, Lynne M. Christianson, Steven H.D. Haddock, Richard E. Green. ***A chromosome-scale genome assembly and karyotype of the ctenophore Hormiphora californensis*.** (2021) G3: Genes, Genomes, Genetics

Séverine Martini[†], Darrin T. Schultz[†], Lonny Lundsten, Steven H.D. Haddock. ***Bioluminescence in an Undescribed Species of Carnivorous Sponge (Cladorhizidae) From the Deep Sea*.** Frontiers in Marine Science 7 (2020): 1041.

Darrin T. Schultz[†], Alexey A. Kotlobay[†], Rustam Ziganshin, Artyom Bannikov, Nadezhda M. Markina, Tatiana V. Chepurnyh et al. ***Luciferase of the Japanese syllid polychaete Odontosyllis undecimdonta.*** Biochemical and Biophysical Research Communications 502, no. 3 (2018): 318-323.

[†] - Indicates co-first authorship

Chapter 1

Introduction

> *"We should venture on the study of every kind of animal without distaste; for each and all will reveal to us something natural and something beautiful."*
>
> -Aristotle, Parts of Animals, 350 BC

Taxonomy and systematics are the scientific disciplines of categorizing, and determining the relationships between, all life on Earth. Humans have been documenting attempts to categorize life, and to make inferences from those relationships for more than two thousand years (Aristotle 350 BC). Since Aristotle's first taxonomic classification scheme in 350 BC, humans have recognized "Animalia" as a classification. Since then, with the advent of DNA sequencing, the definition of animal has morphed to mean multicellular organisms, with a single evolutionary common evolutionary ancestor that may have looked something like extant single-celled flagellum-bearing organisms, the choanoflagellates (Wainright *et al.* 1993). This monophyletic group of multicellular organisms is called the metazoans, or the animals in lay terms.

Metazoans can be broken into two general categories: the bilaterians and the non-bilaterians. Bilaterian animals are a monophyletic clade of organisms that are

bilaterally symmetrical. This includes humans, other vertebrates, chordates, echinoderms, annelids, molluscs, arthropods, chaetognaths, rotifers, nematodes, xenacoelomorphs, and more. All of the remaining animals are non-bilaterians. As the name implies, they are not bilaterally symmetrical. This group of animals includes the phyla Cnidaria, Placozoa, Porifera (sponges), and Ctenophora. Some important major differences between the bilaterians and non-bilaterians are that most bilaterians have extensive nervous systems, often with a brain, and three layers of cells in development in addition to germ cells (triploblasty). Non-bilaterians do not have brains, may not have nerves whatsoever (placozoans and sponges), or have diffuse nerve nets (cnidarians and ctenophores), and have only two developmental cell layers (diploblasty - placozoans, cnidarians, and ctenophores, or have no developmental cell layers at all (sponges).

Ctenophore Genomes

Ctenophore genome assemblies have been key to understanding the early evolution of animals. The draft genomes of the ctenophores *Mnemiopsis leidyi* and *Pleurobrachia bachei* showed that many important animal developmental and neuron-specific genes did not evolve until the common ancestor of the bilateria (Ryan *et al.* 2013; Moroz *et al.* 2014). Years after publication, these two ctenophore genomes remain crucial for studying the evolution of gene families and developmental pathways in the ancestor to all animals (Sebé-Pedrós *et al.* 2018; Fernández and Gabaldón 2020; Tikhonenkov *et al.* 2020; Wang *et al.* 2020), and for

studying the evolution of genome regulation within animals (Gaiti *et al.* 2017; Bråte *et al.* 2018).

It remains controversial whether ctenophores or sponges are sister to the rest of animals (Simion *et al.* 2017; Shen *et al.* 2017; Whelan *et al.* 2017; Laumer *et al.* 2019). Therefore, it is unclear on what ancestral evolutionary branch some metazoan characters evolved, such as neurons and the mesoderm. One method that could possibly resolve the phylogenetic position of ctenophores and sponges is comparing whole-chromosomes (Sacerdot *et al.* 2018). However, the *Mnemiopsis leidyi* and *Pleurobrachia bachei* assemblies are not chromosome-scale. Furthermore, karyotypes are not known for *M. leidyi*, *P. bachei*, or any other ctenophore.

In contrast to the hundreds of published chromosome-scale genome assemblies from vertebrates and other bilterians, there are currently only three from non-bilaterian animals: the freshwater sponge *Ephydatia* (Kenny *et al.* 2020), the cnidarian *Rhopilema* (Li *et al.* 2020; Nong *et al.* 2020), and the cnidarian *Nematostella* (Zimmermann *et al.* 2020). The disparity in the number of bilaterian versus non-bilaterian chromosome-scale assemblies can be partly explained by the difficulties of isolating nucleic acids from non-bilaterians (Dawson *et al.* 1998; Simister *et al.* 2011). Also, non-bilaterians tend to have highly heterozygous genomes (Leffler *et al.* 2012), which complicates standard approaches to genome assembly (Kajitani *et al.* 2014). The assemblies from *Ephydatia*, *Rhopilema*, and *Nematostella*

were possible only due to recent advances in long-read sequencing and the advent of Hi-C data for whole-genome scaffolding (Burton *et al.* 2013; Rice and Green 2019).

Outgroups of Animals and Animal Evolution

The outgroups of the monophyletic animal clade are all unicellular organisms. The immediate outgroup phylum is the Choanoflagellatea (Lang *et al.* 2002). These organisms are solitary, but occasionally agglomerate into larger groups (Kent 1880; Leadbeater 1983). The choanoflagellates and metazoa form a clade called the Choanozoa. The immediate outgroup to the choanoflagellates are the unicellular Filastereans (Paps and Ruiz-Trillo 2010). The Filastereans and the Choanozoa form a clade called the Filozoa. One of the outgroups to the Filozoa are the unicellular Icthyosporeans (Torruella *et al.* 2015). The Icthyosporeans and the Filozoa form a clade called the Holozoa. The outgroup to the Holozoa is the unicellular and multicellular fungi. Together, all of these organisms form a clade called the Opisthokonta.

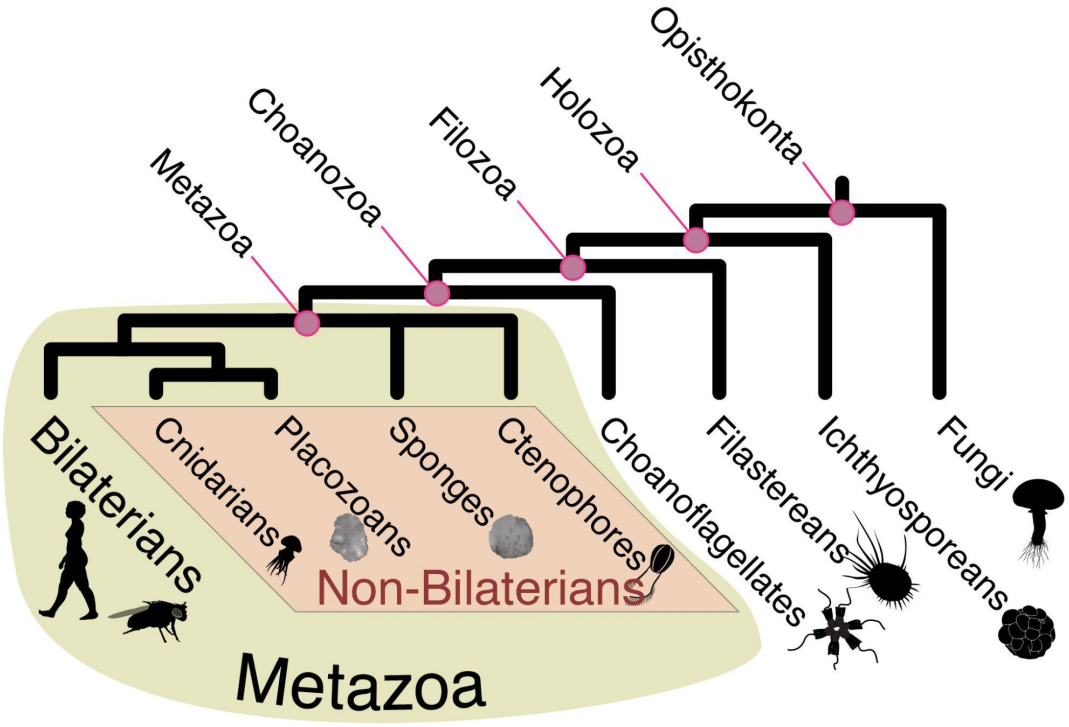

Figure 1.1 - The Animal Tree of Life. The metazoa, also called animals, is a monophyletic clade of multicellular organisms that contains the phyla Ctenophora, Porifera (sponges), Placozoa, Cnidaria, and many phyla within the clade Bilateria.

Because we know the systematics of how multicellular animals are related to other organisms, we know that the common ancestor of the Choanozoa was a unicellular organism that gave rise to all unicellular choanoflagellates and all multicellular animals. This evolutionary transition is the subject of great focus in contemporary biology. There are many studies that have classified which genes were likely animal novelties (Srivastava *et al.* 2010), which multicellularity-associated genes evolved in unicellular organisms (Sebé-Pedrós *et al.* 2017), which transcription

factors common in animals evolved in a unicellular ancestor (Sebé-Pedrós *et al.* 2011), and how genomes likely evolved in animals since the unicellular common ancestor (Suga *et al.* 2013; Sebé-Pedrós *et al.* 2016). It is possible to study the evolutionary transitions from unicellular ancestors to multicellular animals because we know which organisms are the direct outgroups of animals - choanoflagellates, filastereans, and ichthyosporeans (Figure 1.1). It is possible to map characters, or gene gains or losses, onto a phylogenetic tree and draw conclusions based on the topology.

The other transitions that are critically important to understand the evolution of animals happened within the metazoan clade. For example, neurons evolved within the animals, as did triploblasty, the evolution of HOX and other developmental genes, as well as animal-specific transcription factors. We are able to study the evolution of these characters in the bilaterians, the placozoans, and the cnidarians, because we know that these animals form a monophyletic clade (Dunn *et al.* 2008; Philippe *et al.* 2009). However, even after more than a decade of debate using larger and larger phylogenomics datasets, there remains one major outstanding question in animal evolution: Are ctenophores, or are sponges, the sister phylum to all other animals?

Sponges have long been postulated to be the sister phylum to all other animals. This hypothesis grew out of an observation by the educator Henry James Clark in 1866, when he noted the similarities in appearance of choanoflagellates and of flagellum-bearing collar cells in sponges (Clark 1866). Without other definitive evidence to the contrary, this hypothesis was widely accepted until 2008, when a

phylogenomic protein supermatrix analysis showed support for a new hypothesis: that the Ctenophora were the outgroup phylum to all other animals, and that sponges formed a monophyletic clade with cnidarians and bilaterians (Dunn *et al.* 2008). This study has sparked more than a decade of scientific debate over taxon and gene sampling size (Philippe *et al.* 2009; Ryan *et al.* 2013; Simion *et al.* 2017; Whelan *et al.* 2017), gene sample composition (Shen *et al.* 2017), and phylogenetic model suitability (Li *et al.* 2021). In short, it appears that phylogenomics alone, due to its inherent shortcomings, and the short amount of time between the divergence of sponges and ctenophores, will not be able to resolve whether sponges or ctenophores are the sister phylum to all other animals.

However, the answer to this question is too important to leave unresolved. Consider the evolution of neurons in animals. Ctenophores have neurons, but sponges do not. If ctenophores are the sister phylum to animals, then this implies that neurons either evolved independently, once in the ctenophores, and once in the common ancestor of cnidarians and bilaterians with a subsequent loss in placozoans. Or, this implies that neurons evolved once in the ancestor of all animals, but were lost in the sponges and in the placozoans. If sponges are the sister phylum to all other animals, then the evolution of neurons could be explained by a gain in the ancestor to ctenophores, cnidarians, and bilaterians, with a singular loss in the placozoans. In any case, it is impossible to make conclusions about the evolution of neuron-specific genes, or neuron-specific expression patterns, without determining whether sponges or ctenophores are the sister phylum to other animals.

Bioluminescence

Many of the organisms mentioned above, including ctenophores and cnidarians, are bioluminescent. Bioluminescence is the emission of light when an enzyme, usually called a luciferase, oxidizes a substrate called a luciferin. There are thousands of species that are bioluminescent, both on land and in the ocean. Studying bioluminescent animals has not only provided us a lens into the mating (Morin 2019), feeding (Lloyd 1965), and general life history strategies of the animals (Oba and Schultz 2014), but the bioluminescent proteins themselves have become integral research tools to visualize the inner workings of cells (Kaskova *et al.* 2016).

Bioluminescent proteins are generally called luciferases and the small molecule substrates are generally called luciferins. However, these terms are just classifications of function, and luciferases from distantly related groups of organisms are not homologous. For example, luciferases and the luciferins found in beetles are not derived from the same luciferin or ancestral protein found in squid. We know that bioluminescence evolved convergently more than 50 times independently in diverse clades across the tree of life (Haddock *et al.* 2010; Lau and Oakley 2021), however we only know the sequences of a handful of bioluminescent proteins and the structure of even fewer luminous molecules. Discovering more bioluminescent species and studying the proteins and molecules that make them luminesce will help us understand the biology and evolutionary history of those animals, and will provide new molecular tools for scientists and doctors to use in research.

Of all habitats on Earth, the ocean holds the most species that we know to be bioluminescent. Naturally, the ocean also contains the most known bioluminescent species for which we do not know the bioluminescent protein and substrate. Given the small fraction of the ocean that has been explored, there are likely many more species that are bioluminescent, but have not yet been discovered to be bioluminescent.

Chapter 2

A chromosome-scale genome assembly and karyotype of the ctenophore *Hormiphora californensis*

This text is adapted from a published article:

Darrin T. Schultz[†],Warren R. Francis[†], Jakob D. McBroome, Lynne M. Christianson, Steven H.D. Haddock, Richard E. Green. ***A chromosome-scale genome assembly and karyotype of the ctenophore Hormiphora californensis*.** (2021) G3: Genes, Genomes, Genetics

[†] - Indicates co-first authorship

Abstract

Here, we present a karyotype, a chromosome-scale genome assembly, and a genome annotation from the ctenophore *Hormiphora californensis* (Ctenophora: Cydippida: Pleurobrachiidae). The assembly spans 110Mb in 44 scaffolds and 99.47% of the bases are contained in 13 scaffolds. Chromosome micrographs and Hi-C heatmaps support a karyotype of 13 diploid chromosomes. Hi-C data reveal three large heterozygous inversions on chromosome 1, and one heterozygous inversion shares the same gene order found in the genome of the ctenophore *Pleurobrachia bachei*. We find evidence that *Hormiphora californensis* and *Pleurobrachia bachei* share thirteen homologous chromosomes, and the same karyotype of 1n = 13. The manually-curated PacBio Iso-Seq-based genome annotation reveals complex gene structures, including nested genes and trans-spliced leader sequences. This chromosome-scale assembly is a useful resource for ctenophore biology and will aid future studies of metazoan evolution and phylogenetics.

Introduction

Hormiphora californensis is a globular 2 cm ctenophore abundant in the temperate Pacific Ocean with several attractive features for experimental work (Matthews and Vosshall 2020). This species is readily cultured in aquaria (Patry *et al.* 2019), has a life cycle as short as two weeks, produces hundreds to thousands of eggs per spawning event, and has easily-observed embryonic development (Freeman 1977). In addition, there are established CRISPR-Cas9 genome-editing methods for other ctenophore species that may be adaptable to *Hormiphora* (Presnell and Browne 2019). The genus *Hormiphora* is in the same family, the Pleurobrachiidae, as the ctenophore *Pleurobrachia bachei*. Given these useful traits, and the availability of the *P. bachei* genome, we selected *Hormiphora californensis* for chromosome-scale genome assembly.

Here, we report a karyotype, chromosome-scale genome assembly, and a manually-curated genome annotation of *Hormiphora californensis* individual Hc1. Using Hi-C data from Hc1 and Hc3, we present evidence for three heterozygous inversions that span 73% of one *Hormiphora* chromosome. We find that there are several inversion breakpoints in common between *Hormiphora* and *Pleurobrachia*. We estimate the indel and SNP heterozygosity of *H. californensis*. We use Iso-Seq based annotation to resolve hundreds of complex nested intronic genes, and find that trans-spliced leaders are common in ctenophore mRNAs. The *H. californensis* genome assembly, annotation, and sequencing data will be a valuable resource for comparative genomics and evolutionary studies.

Materials and Methods

Sample Collection

We sampled two *H. californensis* individuals, Hc1 and Hc2, wild-caught from the Monterey Bay, (Figure 2.1). We also sampled a third individual, called Hc3, from the seventh generation of a lab-reared culture at the Monterey Bay Aquarium. The aquarium culture's provenance was a single broadcast spawning event from ten individuals wild-caught in the Monterey Bay. Hc1 and Hc2 samples were collected with the *ROV Ventana* aboard the Monterey Bay Aquarium Research Institute's *R/V Rachel Carson*, and from a Tucker trawl aboard MBARI's *R/V Western Flyer*. Samples were flash frozen in liquid nitrogen after allowing the gut to clear. Samples were collected under the State of California Department of Fish and Wildlife collecting permit SC-4029. Additional details are included in Table 2.1 and Figure A1. An individual collected by Tucker trawl, Hc1, was selected as the sole source for DNA and RNA sequencing for the genome assembly and annotation.

Karyotyping

We prepared *H. californensis* chromosomes from embryos to produce a karyotype. To collect embryos we placed Tucker trawl-collected *H. californensis* individuals in 200 mL of 12°C filtered seawater, adapted the animals to darkness for 4 hours, then induced spawning with light (Patry *et al.* 2019). The embryos were concentrated into 10mL of seawater using a 40µm mesh Fisherbrand Sterile Cell Strainer, then were incubated at 12°C for six hours to allow development to

approximately the 64-cell stage. The embryos were fixed using a protocol for chromosome spread preparation for *Nematostella vectensis* (Guo *et al.* 2018). The slides of chromosome preparations were stained using DAPI, mounted with Fluoromount-G, then stored at 4°C until imaging. Micrographs of chromosome spreads were collected with a 100x objective and 1.5x diopter on a Leica DM5500 B microscope with a DAPI excitation light and filter at the UC Santa Cruz Life Sciences Microscopy Center.

Data Preparation

In total, we constructed 13 Illumina and PacBio DNA and RNA sequencing libraries. Eleven of these libraries were from the individual used for genome assembly and annotation, Hc1. The remaining two libraries were one Illumina WGS of individual Hc2, and a Hi-C library of individual Hc3. Briefly, from Hc1 we collected 247x coverage of PacBio WGS CLR reads, 573x coverage of Illumina WGS reads, 1956x coverage of Chicago and Hi-C reads, 28 Gbp of Illumina RNA-seq reads, and 2.5 million Iso-Seq transcripts. The mean read length for both the PacBio Sequel I CLR and the PacBio Sequel II Iso-Seq data was 2.7kb (Figure A2). The Iso-Seq read length distribution roughly matched the size distribution of the input RNA (Figure A3). Sequencing was performed at the University of California Davis (UCD) DNA Technologies Core, Fulgent Genetics, MedGenome Inc., or at the University of Utah. The raw data are available on NCBI under BioProject PRJNA576068. Details for each library are available in Table 2.2. Reads were

trimmed and prepared for genome assembly and genome annotation. For details, see the supplementary section *Sequencing data preparation*.

De Novo Transcriptome Assembly

The trimmed Hc1 Illumina RNA-seq data were assembled using the Trinity v2.5.1 (Grabherr *et al.* 2011) with the parameter `--SS_lib_type RF`. Transcripts that contained adapters or vector contamination in the NCBI contamination database were removed. The assembly is available on the NCBI Transcriptome Shotgun Assembly archive, accession GHXS00000000.

Mitochondrial Genome and Phylogeny

We assembled the mitochondrial genomes of *H. californensis* individuals Hc1 and Hc2 using PacBio and Illumina reads. To determine the phylogenetic position of *H. californensis* individuals Hc1 and Hc2 we constructed an 18S tree, a COX1 nucleotide tree, and a multi-locus mitochondrial protein tree. See the supplementary materials sections *Mitochondrial Genome Assembly and Annotation*, and *Phylogeny construction*.

Genome Size Estimation

K-mers were counted from trimmed Illumina WGS H. californensis reads and from publicly-available P. bachei WGS libraries (SRR116669 and SRR116670) (Moroz *et al.* 2014) using jellyfish v2.2.10 (Marçais and Kingsford 2011) with

options `-C -s 1000000000 -k 21`. Genome sizes of both species were estimated using the k-mer count histograms on the GenomeScope2 server (Ranallo-Benavidez *et al.* 2020).

Genome Assembly

The genome was assembled using wtdbg2 v2.4 (Ruan and Li 2019). The assembly was then polished with arrow v2.2 (github.com/PacificBiosciences/gcpp), then with pilon v1.22 (Walker *et al.* 2014). Haplotigs were removed with Purge Haplotigs v1.0.4 (Roach *et al.* 2018). Dovetail Genomics HiRise vAug2019 was used to scaffold the haplotig-purged assembly with the trimmed Chicago and Hi-C reads. Scaffolds with a mean coverage of less than 100, or having greater than 50% GC, were removed from the assembly using BlobTools v1.1.1 (Laetsch and Blaxter 2017). Assembly gaps were closed with LRGapcloser (commit 156381a). The assembly was polished with pilon. See the *Genome Assembly* section of the supplementary methods for additional details.

Genome Quality Assessment

We calculated the final assembly statistics such as the number of scaffolds, contigs, and the N50, using the program fasta_stats included with the Meraculous assembler (Chapman *et al.* 2011). We also assessed the completeness of the assembly by calculating the percent of PacBio Sequel subreads and full-length non-chimeric (FLNC) transcripts that mapped to the assembly. We also used a custom python script

to calculate the percent of bases of each read type that mapped to the assembly. We performed a self-to-self genome alignment using LASTZ v1.04.03 (Harris 2007) to check for erroneously duplicated regions. To check for uncollapsed haplotypes or regions with many indels we used samtools mpileup v1.7 (Li *et al.* 2009) and chep commit 60c4312 (github.com/conchoecia/chep).

Characterizing Chromosomal Inversions

We generated a Hi-C heatmap to check for genome misassemblies. For details, see the *Hi-C heatmap generation* supplementary section. We noticed three strong off-diagonal bowtie-shaped Hi-C hotspots on Chromosome 1. If this type of signal arises from a misassembly, then the misassembly can be corrected by inverting the bowtie-shaped region of the Hi-C matrix. Heterozygous inversions are not correctable by the same process (Corbett-Detig *et al.* 2019; Chida *et al.* 2020). We used PretextView to combinatorially invert sections of the Hi-C matrix to attempt to remove the off-diagonal signal.

Genome Variant Calling and Phasing

To find diploid variants in the genome, we mapped PacBio CLR and Illumina WGS reads to the genome with minimap2 v2.17 (Li 2017) and BWA-MEM v0.7.17 (Li 2013), called variants using freebayes v1.3.2-38 and gnu parallel v20161222 (Tange and Others 2011), then filtered the VCF to only include diploid calls. We then phased the variants using Picard v2.25.1 (“Picard Toolkit” 2016) and HapCUT2

v1.3.1 (Edge *et al.* 2017). See the section *Genome Variant Calling and Phasing* in the supplementary methods for parameters.

Genome Annotation

We annotated the genome by manually combining transcript models generated from several datasets. The transcript sets were generated with PacBio Iso-Seq and Illumina RNA-seq reads as input for BRAKER v2.14 (Hoff *et al.* 2019), AUGUSTUS v3.3.3 (Stanke *et al.* 2004), and GeneMark-ES/ET v4.65 (Hoff *et al.* 2016). The PacBio Iso-Seq data were used as input for PacBio Cupcake tools v8.0 (github.com/Magdoll/cDNA_Cupcake), StringTie v2.0.4 (Pertea *et al.* 2015), and Pinfish commit b6f3c06 (github.com/nanoporetech/pinfish). See the *Genome Annotation* and *Transcript phasing* sections of the supplementary materials for details on how each program was run.

Genome annotation consisted of three rounds. In annotation round 1 we reviewed the genome in IGV or JBrowse and manually verified the StringTie transcripts that were generated with the PacBio Iso-Seq data. Specifically, we identified if the StringTie transcripts were fused, correct, or fragmented by comparing them to the full-length, non-chimeric (FLNC) PacBio Iso-Seq reads. If the mapping pattern of PacBio Iso-seq reads suggested that a StringTie transcript was a fusion between two or more adjacent transcripts, then we replaced the fused StringTie transcript with Pinfish, AUGUSTUS, or GeneMark-ES/ET transcripts. The replacement transcripts were selected if they matched the gene structure of the PacBio

Iso-Seq reads that mapped to the same locus. If a StringTie transcript was only a partial gene, also evidenced by the PacBio Iso-Seq reads, then the partial StringTie transcript was replaced with a correct Pinfish, AUGUSTUS, or GeneMark-ES/ET transcript. StringTie transcripts that did not match a transcript observed in the PacBio Iso-Seq data were removed. If Iso-Seq reads were mapped to a locus, but the locus had no representative StringTie transcript, then a matching Pinfish, AUGUSTUS, or GeneMark-ES/ET transcript was added to the annotation. StringTie transcripts that were grouped together by StringTie, but actually represented multiple genes with mutually exclusive exons, were split into multiple genes. At this stage the annotation contained genes and transcripts representing the complement of PacBio Iso-Seq data derived from the adult Hc1.

In annotation round 2, AUGUSTUS gene models generated from hints that included Illumina RNA-seq reads were added to the annotation if they did not overlap with the transcripts from round 1.

In annotation round 3, we sought to find life-stage-specific and tissue-specific transcripts in the *H. californensis* genome that may not have been present in the RNA sample from the adult Hc1. Gene models were generated by mapping *P. bachei* transcripts to the *H. californensis* genome. The resulting gene models were removed if they did not contain an ORF in the *H. californensis* genome, or if they overlapped with *H. californensis* annotation round 1 or round 2 genes. Gene models were only included in the annotation if their ORF's protein product had a blastp hit to publicly-available ctenophore transcriptomes with an e-value of less than 1e-10.

For each transcript in the annotation we generated haplotype-resolved protein sequences. See the *Genome Annotation* and *Transcript phasing* sections of the supplementary materials for more details.

Annotation Completeness Assessment

We used gVolante and BUSCO Eukaryota v3 to calculate the BUSCO score of the protein models from our annotation, the *de novo* transcriptome, and the genome assembly (Simão *et al.* 2015; Nishimura *et al.* 2017).

TAD Calling and Boundary Analysis

We called topologically associating domains (TADs) using the HOMER Hi-C analysis pipeline (Heinz *et al.* 2018). The TADs were called with 1kb/4kb bin resolutions and 10kb/40kb windows. We masked regions around Hi-C heatmap irregularities such as off-diagonal signal that appeared to be due to inversions or misassemblies. This signal confounds the discovery of TADs in well-assembled genome regions. We calculated TAD separation scores with HiCExplorer v3.6 (Ramírez *et al.* 2018) using Cooler v0.8.10 (Abdennur and Mirny 2020).

We applied HOMER's *de novo* motif discovery pipeline to 1.5kb regions on either side of each TAD boundary (Heinz *et al.* 2018). For motif discovery, we selected background regions that exhibited minimal local change in TAD separation score, as these regions least resemble TAD boundaries.

We noticed that TADs tended to occur near gene boundaries. To test the significance of this observation we performed a permutation test. We first measured the median distance between TAD boundaries and the nearest gene. The background distribution was calculated by 1000 permutations of randomly placing TADs across the genome using the same size distribution as our observed TADs, then measuring the distance to the nearest gene.

Identification of Nested Intronic Genes

Nested intronic genes were identified using chep gff_to_intron_bed.py (github.com/conchoecia/chep), allowing for a 15% overlap with the host exons at both the 5′ and 3′ ends of the nested transcript. We excluded the longest 0.5% of introns from the analysis to avoid counting the introns from trans-spliced splice leaders.

Repeats and Centromeres

We used Tandem Repeats Finder v March 13, 2006 (Benson, 1999) and EDTA v1.8.3 (Ou *et al.* 2019) to identify repeats and transposable elements.

Whole-Genome Heterozygosity Estimation

The single nucleotide heterozygosity of Hc1 was estimated by only using sites that had exactly 178x Illumina WGS read mapping depth. This depth, 178x, was the mode of the mapping depth for the whole genome, and thus represents sites at which

reads from both haplotypes mapped. We implemented this method, first described in Saremi et al. 2019, in a purpose-built software package called chep (github.com/conchoecia/chep).We also measured the heterozygosity of the Hc1 exonic, intronic, and intergenic regions on individual chromosomes using chep.

We measured the heterozygosity of Hc1, Hc2, and *P. bachei* individual SAMN00216730 by counting 21-mers with jellyfish v2.2.10 (Marçais and Kingsford 2011) then using the resulting spectrum in GenomeScope 2 (Ranallo-Benavidez *et al.* 2020), by using vcftools' `--het` option (v0.1.17), and by using angsd realSFS (v0.921) (Danecek *et al.* 2011; Korneliussen *et al.* 2014). We were not able to measure the heterozygosity of *M. leidyi* because the sequencing libraries were derived from multiple individuals.

Analysis of the HiC-scaffolded Pleurobrachia genome

The *Pleurobrachia bachei* genome assembly was recently scaffolded using Hi-C data (Hoencamp *et al.* 2021). The new, scaffolded assembly did not include a genome annotation. We identified the protein positions in the *P. bachei* genome assembly using tblastn and the previously published *P. bachei* protein sequences (Moroz *et al.* 2014). We also looked for orthologous scaffolds *H. californensis* and *P. bachei* genomes by plotting the protein coordinates of reciprocal best blastp hits between the proteins of the two genomes. For comparative analysis, we generated a *P. bachei* Hi-C heatmap by mapping the *P. bachei* Hi-C reads, SRR13364273

(Hoencamp *et al.* 2021), to the new assembly using the same protocol as for *H. californensis*. See the *Hi-C heatmap generation* supplementary section for details.

Microsynteny Between Ctenophores

The *H. californensis* proteins were queried against the *M. leidyi* and *P. bachei* proteins using diamond blastp v0.9.24 (Buchfink *et al.* 2015). The gene positions and the diamond blastp table were then used to identify collinear blocks of genes using the purpose-built Python script microsynteny.py (github.com/wrf/genomeGTFtools). We required a minimum of 3 consecutive genes, allowed for up to 5 intervening genes, and allowed a maximum distance of 30kb to the next gene.

Results and Discussion

Genome Sequencing and Assembly

To determine the ploidy of *H. californensis* and to estimate its genome size, we computed k-mer spectra from *H. californensis* and *P. bachei* WGS libraries. Libraries from both species had two major k-mer peaks. The lower-coverage peak was larger than the higher-coverage peak in both species. This pattern is consistent with the k-mer spectra of other highly heterozygous diploid organisms (Ranallo-Benavidez *et al.* 2020). From the k-mer spectrum, the predicted 1C genome size of *H. californensis* was 96-98 Mb (Figure A4), which is close to the predicted *P. bachei* genome size, 97.5 Mb (Figure A5).

We aimed to generate a chromosome-scale reference genome for *H. californensis* in which each chromosome is represented by a composite sequence obtained by combining both haplotypes. This genome sequence was assembled from PacBio long reads, polished with Illumina short paired-end reads, and scaffolded with in vitro and in vivo chromatin conformation capture reads from a single individual, Hc1 (Methods). The *H. californensis* genome assembly totaled 110.6 Mb in 44 scaffolds and 351 contigs. Half of the sequence is present in scaffolds longer than 8.5 Mb, with 2.76 gaps per Mb within scaffolds. The 13 longest scaffolds comprise 99.47% of the assembly, ranging in size from 10.3 Mb to 6.4 Mb. These 13 long scaffolds match the microscopy-based karyotype of n=13, detailed below. The remaining 31 scaffolds were each shorter than 50kb and represent short unplaced sequences. We found no detectable contamination from marine bacteria or gut

contents based on the blobtools results (Figure A6). The genome dotplot made with D-Genies (Cabanettes and Klopp 2018) did not reveal erroneously duplicated assembly regions (Figure A7). 95.32% of the PacBio Sequel subreads (Table A2), and 99.02% of PacBio Sequel II Iso-Seq full-length non-chimeric (FLNC) transcripts mapped to the 13 largest scaffolds.

We note that our 110 Mb *H. californensis* assembly is substantially shorter than the published *P. bachei* assembly (156 Mb) despite similar size estimates based on k-mer spectra. Our analysis of the *P. bachei* assembly, included in the supplemental text, suggests that over half of the reported *P. bachei* scaffolds are unmerged haplotypes.

Variant Calling and Phasing

We called variants using freebayes after mapping Hc1 PacBio CLR and Illumina WGS reads to the *H. californensis* reference genome. After filtering we identified 2.24 million heterozygous single nucleotide or indel variants. These variants were phased using PacBio CLR, Chicago, and Hi-C reads, resulting in phased blocks of variants that spanned more than 99% of the length of each chromosome-length scaffold. Of the 2.24 million diploid variants, 1.75 million (77.9%) were in the chromosome-scale phased variant blocks. The high density of phased variants, one for every 63 bp of genome assembly, suggests that the *H. californensis* data may be a useful benchmarking candidate for phased, or diploid, genome assemblers.

Mitochondrial Genome and Phylogeny

We assembled and annotated the mitochondrial genomes of Hc1 and Hc2 (Figure A8), two individuals from the same Monterey Bay population, and collected three years apart. The mitochondrial genomes (mtDNA) from Hc1 and Hc2 were 99.6% identical. The *H. californensis* mtDNA is 71.5% identical to the mtDNA of the closely-related *P. bachei*, and 80% identical to *P. bachei* when only considering coding regions. The *H. californensis* mitochondrial genome has a 1.8 kb insertion relative to *P. bachei*, between COX2 and 16S. The percent identity between the *H. californensis* and *P. bachei* mtDNA confirms they are distinct species, despite their similar morphology. Phylogenetic analysis of *P. bachei* and *H. californensis* mtDNA is also consistent with these being distinct species (Figure A9).

Despite the distinct mtDNA of *H. californensis* and *P. bachei,* phylogenetic trees based on 18S rRNA and COX1 show that *H. californensis* falls within the *Pleurobrachia* clade. Furthermore, other *Hormiphora* species are sister to *Pleurobrachia*. These results suggest that the genus *Hormiphora* as currently defined may be polyphyletic. Future taxonomy work should consider reassigning *Hormiphora californensis* to the genus *Pleurobrachia*.

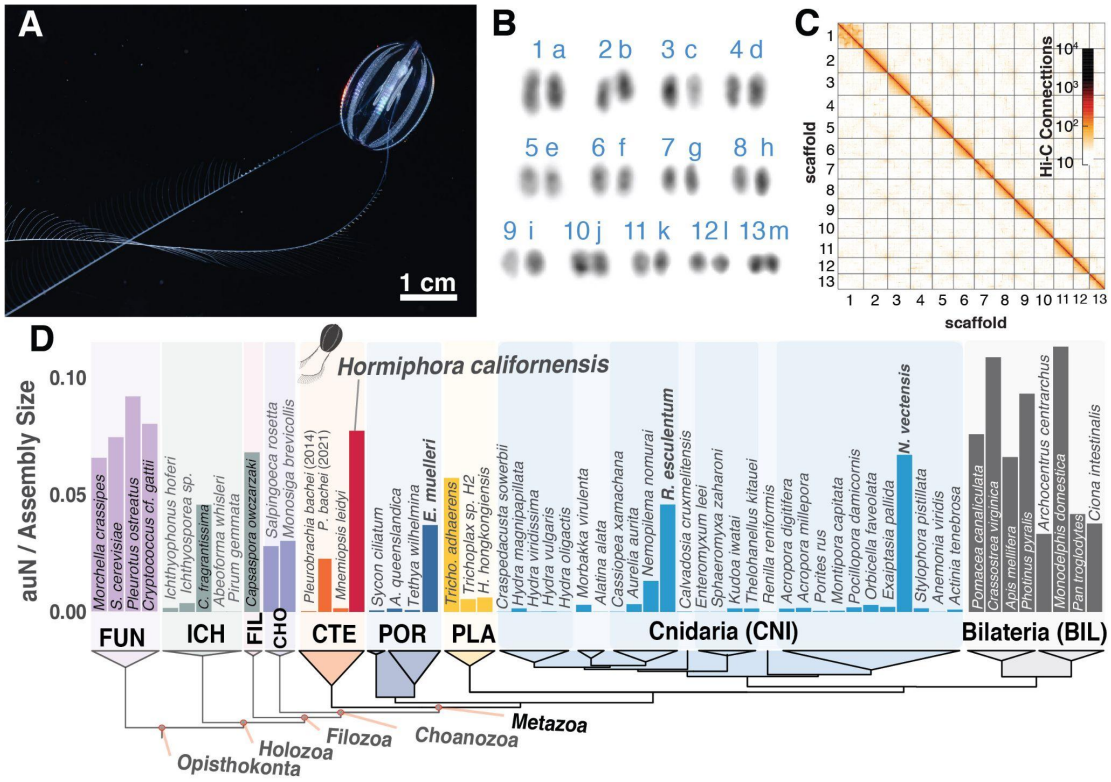

Figure 2.1 - Karyotype and genome assembly quality. (A) *H. californensis* with its tentacles extended during feeding. (B) One karyotype image (rearranged and color-inverted) of a *H. californensis* chromosome preparation that contained 13 chromosome pairs. (C) The *H. californensis* Hi-C map, showing thirteen chromosome-scale scaffolds. (D) Genome contiguity normalized by genome size (auN / Assembly Size, per (Li 2020) from all available non-bilaterian holozoan genomes and select fungal and bilaterian genomes, with tree topology based on (Whelan *et al.* 2017). Bolded species names are non-bilaterians with chromosome-scale genome assemblies. Terms- FUN Fungi, ICH Ichthyosporea, FIL - Filasteria, CHO - choanoflagellates, CTE - ctenophores, POR - Porifera, PLA - placozoans, CNI - cnidarians , BIL - bilaterians.

BioSample Accession	Species	Sample Name	Depth	Collection Temp	Collection Method	Data types	Collection Date yyyymmdd	Collection Latitude DMS	Collection Longitude DMS	Mitochondria Accession
SAMN12924379	*Hormiphora californensis*	Hc1	0-520m	6°C-13°C	Tucker Trawl	DNA, RNA	20161213	36° 41' 42" N	122° 5' 22" W	MN544300
SAMN12924380	*Hormiphora californensis*	Hc2	230m	8.9°C	ROV Ventana	DNA	20131111	36° 41' 48" N	122° 2' 36" W	MN544301
SAMN16124402	*Hormiphora californensis*	Hc3	0m	12°C	F7 from Monterey Bay Aquarium	DNA	20200605	36° 41' 48" N	122° 2' 36" W	NA

Table 2.1 - *Hormiphora californensis* sample collection details.

Karyotype

The karyotype has not been previously described for any ctenophore species. We used microscopy of DAPI-stained chromosome spreads to determine that the *H. californensis* genome is composed of n = 13 chromosome pairs (Figure 2.1, Figure A10). Four of the images correspond to a 2n of 26, and the remaining images are within 3 chromosomes of 2n of 26. This count, 13 pairs, is consistent with the 13 multi-megabase scaffolds in the *de novo* genome assembly presented here.

Individual	SRR	Data Type	Total GB	Number of reads (or pairs) - Millions	Physical coverage
Hc1	SRR10237148, SRR10237149, SRR10237137	PacBio WGS CLR	27.4	9.7	247.7
	SRR10237134	10X Chromium	22.3	74.3	201.4
	SRR10237129, SRR10237130, SRR10237131, SRR10237132, SRR10237133	Illumina WGS - NEB Next UII	36.1	120.2	325.8
	SRR10237128	Chicago - DpnII	10.7	35.6	96.6
	SRR10237146, SRR10237147	Chicago - MluCI	20.9	69.8	189.2
	SRR10237145, SRR13784183	HiC - DpnII - rep 1	50.0	166.5	451.3
	SRR10237144, SRR13784182	HiC - DpnII - rep 2	68.1	226.8	614.8
	SRR10237139, SRR13784181	HiC - MluCI - rep 1	38.1	127.1	344.5
	SRR10237138, SRR13784180	HiC - MluCI - rep 2	28.8	95.9	260.0
	SRR10237136	TruSeq RNA Library Prep Kit v2 (stranded)	28.8	96.0	NA
	SRR10403849, SRR10403581	Iso-Seq Express	6.0	2.5	NA
Hc2	SRR10237135	Illumina TruSeq Nano DNA	12.8	64.0	115.7
Hc3	SRR12632403, SRR13784179	HiC - DpnII	70.2	233.9	633.8

Table 2.2 - A summary of the *H. californensis* SRAs sequenced for this study.

Each row is a single library. For individual Hc1, the total physical coverage of DNA WGS reads is 573.5x, and the coverage for Hi-C and Chicago data is 1956.4x. More detailed information can be found in Data S1.

Genome Annotation

We manually annotated the genome using gene models generated with Hc1 Iso-Seq reads, Illumina RNA-seq data, and *P. bachei* transcripts. The long Iso-Seq reads capture, in many cases, complete cDNA sequences and represent transcripts from a single haplotype. These features allowed us to produce haplotype-specific transcripts and proteins.

Our approach to annotating the *H. californensis* genome identified 14,265 protein-coding genes, of which 13,235 are supported by Iso-Seq reads (Table A2). The BUSCO complete plus fragmented score was 96% (303 Eukaryotic genes - Table A3).We found that 96% of the *Pleurobrachia* proteins with orthologs in other ctenophores also have an ortholog in the *H. californensis* annotation. We note that due to the haplotype redundancy of the *P. bachei* assembly, many annotated *P. bachei* genes are reported in allelic copies, which therefore overestimates the gene count of this species (Supplementary Materials).

Additionally, the *H. californensis* genome contains 619 protein-coding genes that have orthologs in other ctenophore transcriptomes, but do not appear in either the *M. leidyi* or *P. bachei* genomes (Table S2). Of those 619 genes, 122 had blastp hits to nr, and included genes with a wide variety of functions such as DNA-binding proteins, calmodulins, histones, proteases, and more. Among these 122 genes we did not find any evidence for the presence of the neural and mesoderm-component genes reported to be missing from ctenophores (Ryan *et al.* 2013).

We found 1729 cases where two or more neighboring *P. bachei* gene models, and 1200 cases where *M. leidyi* gene models, appear to be fragments of a larger gene based on orthology with *H. californensis*. For example, the pecanex gene (2096 amino acids in *H. californensis*) appears to be split into four proteins in the *M. leidyi* annotation (Figure A11).

97.7% of the eukaryotic BUSCOs were complete or partial in the translated Iso-Seq FLNC data, and 99.0% were complete or partial in the Illumina RNA-seq *de novo* transcriptome. Because these values are higher than the 96% complete or partial BUSCOs from the genome annotation, it is possible that the genome annotation does not capture the full complement of *H. californensis* genes. Future annotation iterations will benefit from Iso-Seq sequencing of different tissues and developmental stages.

Tandem Repeats Finder (Benson, 1999) identified 14 Mb (13%) of the genome as repeats, none of which were identifiable as centromeric. Thus, we are unable to annotate or further describe centromeres in these genomes.

Topologically Associating Domains and 3D Genome Structure

Genome analyses using Hi-C data have shown that in many species, chromatin is organized in segments of close proximity that are known as topologically associating domains (TADs) (Lieberman-Aiden *et al.* 2009).

We used proximity ligation data to identify and characterize TADs in *H. californensis* and found evidence that the *H. californensis* genome contains small TADs with a median length of 60kb. The *H. californensis* TADs are significantly smaller than human TADs (median length 1.15 Mb) (McArthur and Capra 2021). Despite the fact that the *Ephydatia* and *Drosophila* genomes are comparable in size to the *H. californensis* genome, the mean *H. californensis* TAD length is half the length of the TADs in those two species (Hou *et al.* 2012; Kenny *et al.* 2020). We found that TAD boundaries tend to occur in the non-coding DNA bordering genes ($p=0.001$, permutation). The mean distance from a TAD boundary to a gene is 7.8kb.

We used HOMER to search, de novo, for DNA motifs in the sequences flanking the outside of TADs. In these sequences flanking TADs we found enriched motifs, most of which resembled the RNA polymerase II-binding motifs of known transcription factors. The six most-enriched motifs were homeodomain and MYB-related transcription factor binding sites, which are conserved in eukaryotes. Homeodomain and MYB-related transcription factor genes, as well as RNA polymerase II, were present in the *H. californensis* genome annotation.

Analysis of the HiC-scaffolded Pleurobrachia genome

We assessed the quality of the *P. bachei* genome assembly that was recently scaffolded using Hi-C data (Hoencamp *et al.* 2021). This assembly was generated by linking together contigs and scaffolds from the original *P. bachei* assembly (Moroz *et al.* 2014). The authors reported that they found 13 or more putative chromosomal scaffolds, but did not provide further description or analysis of the assembly.

The scaffolded *P. bachei* genome contains 157.1 Mbp in 20,121 scaffolds and 39,072 contigs. The 13 largest scaffolds contain 81.5 Mb of sequence, and appear to be chromosome-scale in the Hi-C map. Those scaffolds contain 69.4 Mb of contigs, and 12.2 Mb of stretches of Ns. Given that the predicted 1C size of the *P. bachei* genome is 96.6 Mbp (see results above), the assembly size of the contigs in the 13 largest scaffolds relative to the predicted size is 72%. The 81.5 Mbp in the 13 largest scaffolds is 51.9% of the total assembly size (Figure A12).

To further investigate the completeness and correctness of the 13 chromosome-scale *P. bachei* scaffolds, we performed synteny comparisons with the *H. californiensis* genome assembly. The Moroz *et al.* (2014) and Hoencamp *et al*. (2021) *P. bachei* genomes do not include genome annotations, although Moroz *et al.* (2014) included 19,002 *P. bachei* protein models from the genome sequence. Using those 19,002 proteins we identified 9,714 genes on the 13 largest *P. bachei* scaffolds. This is 68.2% of the total number of genes found on the 13 largest *Hormiphora* scaffolds. We performed a reciprocal-best blastp search between the *P. bachei* and *H. californensis* proteins and plotted their coordinates in 2-dimensions to visualize

regions of macrosynteny, also called an Oxford dot plot. This plot revealed that each of the 13 largest *P. bachei* and *H. californensis* scaffolds predominantly had reciprocal best blastp hits on only one scaffold of the other species (Figure A13). We did not find evidence for chromosomal fusions or fissions between the karyotype of the two animals.

In summary, we find that the scaffolded *P. bachei* assembly contains 13 scaffolds that correspond to the 13 chromosomes of *H. californiensis*. However, as a large fraction of the *P. bachei* genome is not represented in these 13 scaffolds as measured by overall assembled genome sequence or gene content, the *P. bachei* genome would likely be greatly improved using a contemporary long-read contig assembly approach, followed by chromosome-scale scaffolding.

Heterozygous Chromosome Inversions and Microsynteny

Chromosome 1 of *H. californensis* individual Hc1 contains three large heterozygous chromosomal inversions (Figure 2.2A, Figure A14). Each inversion is approximately 2 Mbp, or 20% of chromosome 1. Together, these putative inversions span 73% of the length of chromosome 1. These are unlikely to be assembly errors, since inverting the Hi-C heatmap around the errors does not remove the off-diagonal signal. All six breaks of between-haplotype synteny appear to occur between genes, and outside of TAD boundaries. The Hi-C matrix from individual Hc3 does not have off-diagonal hotspots (Figure 2.2C), suggesting that both haplotypes of Hc3 chromosome 1 match the genome assembly sequence. Large heterozygous inversions can prevent recombination over large chromosomal regions (Kirkpatrick 2010;

McBroome *et al.* 2020), therefore these two haplotypes of *H. californensis* chromosome 1 may be segregating independently. Large heterozygous inversions between the haplotypes in one individual are not prevalent in vertebrate species, but have been observed before in the genomes of other invertebrates, such as in the mosquito *Anopheles gambiae* (Corbett-Detig *et al.* 2019).

We examined whether the inversion breakpoints in *H. californensis* chromosome 1 also occurred in the genome of the closely-related *P. bachei.* Three of these breakpoints, including an exact match to Hc1 heterozygous inversion 2, were found to occur in the *P. bachei* genome. Hc1 inversion 2, in which *H. californensis* gene 355 on chromosome 1 lies next to *H. californensis* gene 864 (sequentially numbered) on the alternate haplotype, reflects the gene order in the *P. bachei* genome on scaffold AVPN01000013.1 (Figure 2.2B). The *H. californensis* inversion breakpoint at position 5.20 Mb is also a point of synteny mismatch in *P. bachei*, in which the gene on one side of the *P. bachei* synteny break matches a synteny breakpoint in *H. californensis* (*H. californensis* c1.g741). However, the gene on the opposite side of the synteny break (*H. californensis* c1.g424) does not match any of the gene intervals from inversions 1, 2, or 3 from *H. californensis*. These results suggest that chromosomal inversions may not only exist between different ctenophore species, but also may be prevalent within a single population of one species.

Gene colinearity analyses suggest that *H. californensis* and *M. leidyi* only share limited gene microsynteny. The largest identifiable blocks of gene colinearity only contained four genes in common. Given the extensive gene rearrangements seen

between the closely-related species *H. californensis* and *P. bachei*, it is not surprising to find the lack of gene colinearity between the distantly-related *H. californensis* and *M. leidyi*.

The largest colinear block was over 5.8 Mbp of *H.californensis-P.bachei* chromosome 5, encompassing 964 genes in *H. californensis*. However, most other chromosomes were significantly rearranged between the two species. There were no chromosomes in which the gene order appeared completely conserved between the two species.

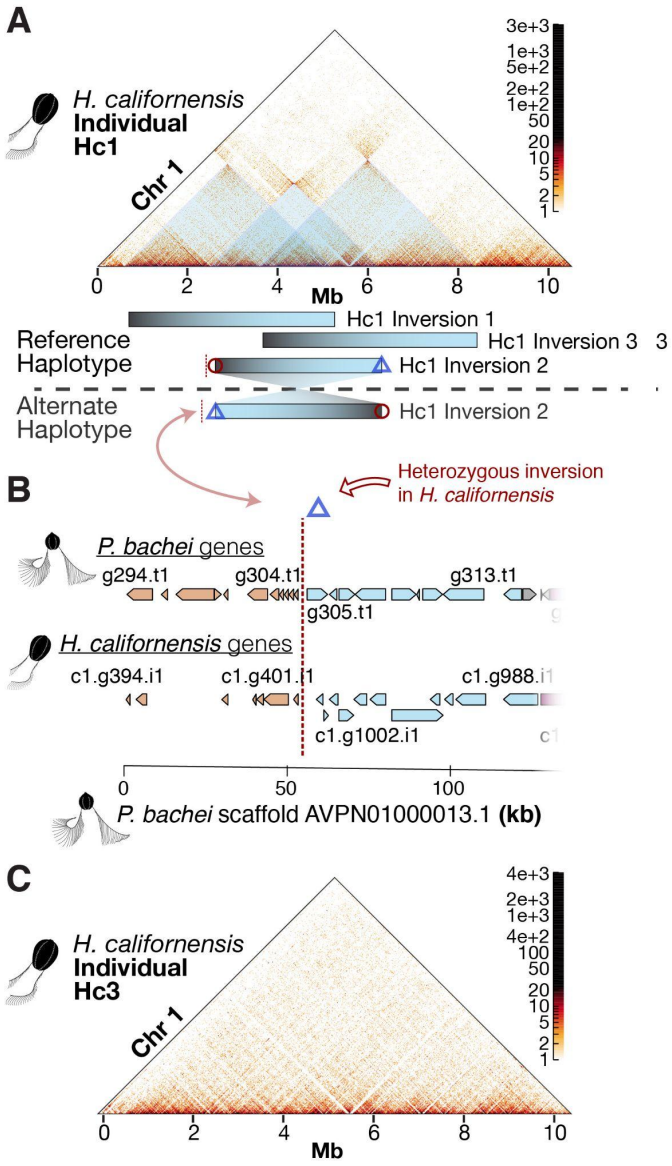

Figure 2.2 - Heterozygous inversions on *H. californensis* Chromosome 1. (A) Three heterozygous overlapping inversions are present on chromosome 1 in the Hc1 genome. Black/blue bars show the spans of each heterozygous inversion. (B) The alternative haplotype of Hc1 inversion 2, indicated by the close proximity of the blue diamond and red line, has the same gene order as the *P. bachei genome*. The x-axis is genome coordinates of *P. bachei* scaffold AVPN01000013.1. Orange *H. californensis* and *P. bachei* genes to the left of the vertical red dotted line are orthologous and have microsynteny. Blue genes to the right of the vertical red dotted line are orthologous and have microsynteny. (C) Hi-C map of *H. californensis* individual Hc3 shows concordance with Hc chromosome 1, but no off-diagonal Hi-C evidence for heterozygous inversions.

Heterozygosity

We measured the heterozygosity of the intronic, exonic, and intergenic regions of *H. californensis* and six other metazoan species (Figure 2.3, Figure A15, Table A4, Table A5) using a method that avoids mis-estimation due to genome assembly errors or inaccurate heterozygous site calls in a VCF file (Saremi *et al.* 2019). *H. californensis* had a high combined single nucleotide and indel heterozygosity rate —approximately 3.2% overall, and a per-chromosome rate of between 2.4% to 4.7%. The overall single nucleotide heterozygosity was 2%. These analyses also revealed that both *H. californensis* and the sponge *Tethya wilhelma* had high SNP heterozygosity in exons, but depressed SNP heterozygosity in both intergenic and intronic regions (Figure 2.3D,E). This pattern is contrary to other species, where heterozygosity of the introns and intergenic regions is higher than in the exons. This pattern in our data is likely due to short Illumina reads from one allele not mapping to regions with high combined SNP and indel heterozygosity (Figure 2.3A), therefore placing an artificial ceiling on the measurable heterozygosity. Using long, accurate reads such as PacBio HiFi data, or measuring heterozygosity with a diploid genome assembly, should overcome these shortcomings.

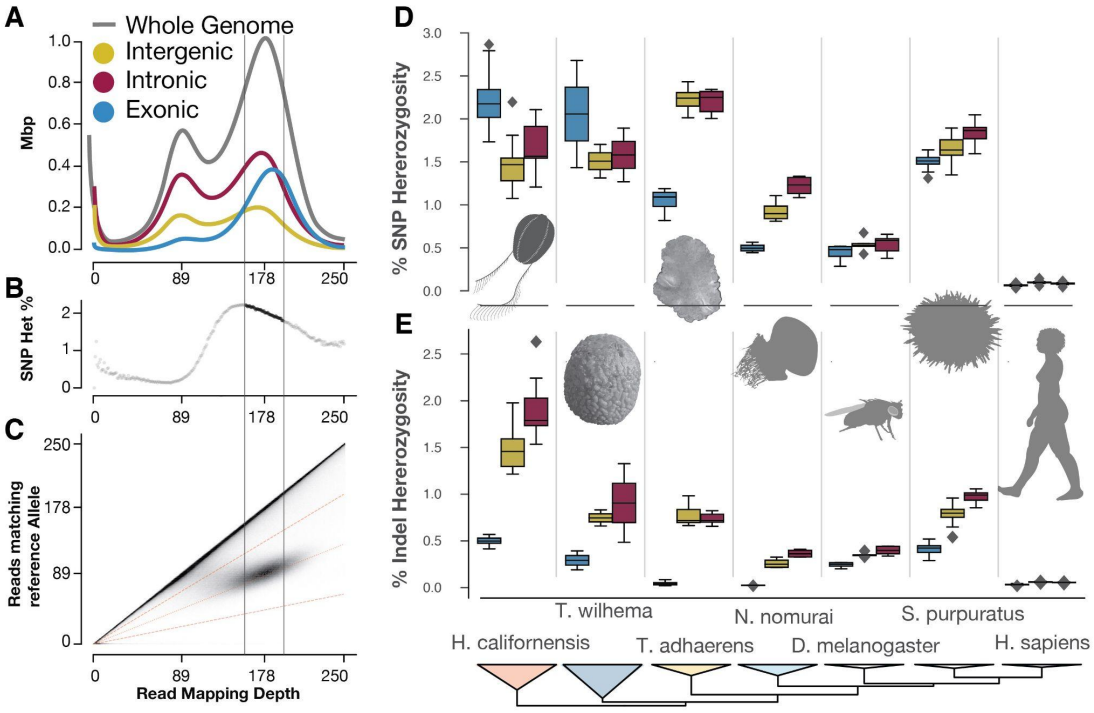

Figure 2.3 - The *Hormiphora* genome is highly heterozygous and contains many indels in non-coding regions. (A) The histogram of the number of bases in a genome assembly (y-axis) with a specific mapping depth (x-axis) using Illumina WGS reads. The mode of the mapping depth of the Illumina WGS reads was 178x, so bases with close to 178x mapping coverage have reads from both haplotypes. Therefore, bases with a mapping coverage around 89x have reads mapped from a single haplotype. (B) The whole-genome heterozygosity calculated at each read mapping depth. The heterozygosity value for each mapping depth (x-axis) is the number of sites where only one half of the mapped reads match the reference allele, divided by the total number of sites at that depth. (C) The 2D histogram showing that bi-allelic sites are centered around 178x read mapping depth. Also, the 2D histogram is one way to visualize the data input for panel B. (D-E) Box-and-whiskers plots show the

distribution of SNP (D) or indel (E) heterozygosity in each chromosome-size scaffold. In *H. californensis*, the intergenic and intronic regions have a reduced SNP heterozygosity, but an elevated rate of indels. This pattern differs from the correlation between SNP and indel heterozygosity found in species with lower overall heterozygosity. (End of Figure 2.3 caption)

Ubiquity of trans-spliced leader sequences

A 2010 study of ctenophore and cnidarian ESTs showed that these phyla have extensive 5′-capping trans-splicing (Derelle *et al.* 2010). However, this study lacked genomes to examine the origins of the leader sequence (Derelle *et al.* 2010). One prevalent feature of *H. californensis* genome organization was gene clusters sharing a 5′ exon, but otherwise having mutually exclusive exons. The shared first exons were between 35-48bp long, and were all >90% identical to 5′GAGTTTCAAACTTTTCAACACTACTTTAAACAAATTAATTTGAG 3′. We identified 718 of these leader sequences in the *H. californensis* genome. The leader sequence was found on 56% of our IsoSeq reads. This appears to be the result of trans-splicing of a leader sequence (Boroni *et al.* 2018). The Iso-Seq reads lacking the leader sequence may be sheared at the 5′ end, as is common in full-length cDNA library preparation. Thus, 56% represents a lower bound for the true percentage of *H. californensis* mRNAs with trans-spliced leaders. The shared exons we identified in the genome may be result of the leader sequences on Iso-Seq reads being artifactually

mapped to the nearest spliced-leader locus 5′ of the transcript in the genome assembly.

In *Mnemiopsis leidyi*, although it was not reported previously, we found several examples of gene clusters with shared first exons using a *de novo M. leidyi* transcriptome (SRX993241) mapped to the *M. leidyi* genome. The leader motif from Derelle et al. 2010 was also identified in the *M. leidyi* genome 491 times. Both the *M. leidyi* and *H. californensis* leader sequences end in a TGAG motif, part of a mostly-conserved 5′AATTTGAG 3′ motif. Over half of the annotated transcripts in *H. californensis* begin with AG.

Nested Intronic Genes In the Metazoa

Using 1,058 eukaryotic genomes, including all genomes available on NCBI RefSeq, we quantified the percent of exonic basepairs that are from nested intronic (NI) genes — genes whose transcripts are within the boundaries of a single intron of another gene (Figure 2.4). In the *H. californensis* genome we found 1,654 genes hosting one or more NI genes. There were 2,357 NI genes inside the 1,654 host genes (Supplementary Data). We estimated that *H. californensis* has 12.24% of exonic bases in NI genes, similar to the rate found in primate and some arthropod genomes. From the 2,357 NI genes we identified 484 doubly-nested genes, which are nested intronic genes within another nested intronic gene. We found that 1,109 NI genes are flanked by transposable elements (TEs) on at least one side of the gene, and 176 NI genes are flanked on both sides by TEs. Many NI genes are also parallel with the host gene,

necessitating a complex transcription or splicing system to separately process the two genes. Parallel NI genes have been observed before in other taxa, such as human (Yu *et al.* 2005) and fly (Henikoff *et al.* 1986).

We observed that NI genes are largely absent from the genomes of protists, fungi, and other non-metazoan opisthokonts such as the choanoflagellates. This observation could be due to genome annotation errors in those clades, or NI genes may have undergone genomic expansions in the metazoan last common ancestor (LCA), and in the plant LCA. We also found that smaller animal genomes tend to have a higher percent of exonic bases in nested intronic genes (Figure 2.4C). Given that nested intronic genes introduce additional ways to control transcription, such as antisense transcription competition (Yu *et al.* 2005), the punctuated appearance of these genes in the metazoa is possibly one of the complex transcriptional control mechanisms that evolved in the ancestor to all animals. High-quality genome assemblies and annotations of outgroup species to the metazoa will be necessary to determine if nested genes are a feature of metazoan genomes.

Summary

We describe a chromosome-scale genome assembly of the ctenophore *Hormiphora californensis*. The assembly consists of 13 chromosome-scale scaffolds that comprise 99.47% of the assembly and 31 additional small scaffolds. The number of chromosome-scale scaffolds in the genome is concordant with microscopic karyotyping. We found evidence that the karyotype of *P. bachei* genome is also 1n =

13, and that *P. bachei* and *H. californensis* chromosomes are homologous, but highly rearranged between the two species. Our manual genome annotation using Iso-Seq data reveals many large transcripts that appear to be fragmented in previously-published ctenophore genome annotations. In addition, we find genomic evidence in support of the putative trans-spliced leader sequences that were first discovered in ctenophore EST datasets over 10 years ago. We characterize nested intronic genes. Analyses of SNP and indel heterozygosity in *Hormiphora* show that the genome is approximately 3.2% heterozygous, with an exceptionally high indel heterozygosity ranging between 1%-2.25% in non-coding genomic regions. Lastly, the Hi-C data show that large heterozygous chromosomal inversions are present in single ctenophore populations, and are a distinguishing feature between ctenophore species. Future studies of ctenophore genomes would benefit from focusing on haplotype-resolved assemblies to clarify remaining questions about genome structure and between-haplotype nucleotide diversity. This high-quality annotation will benefit future ctenophore and metazoan phylogenomic and evolutionary studies, as well as future ctenophore genome assemblies.

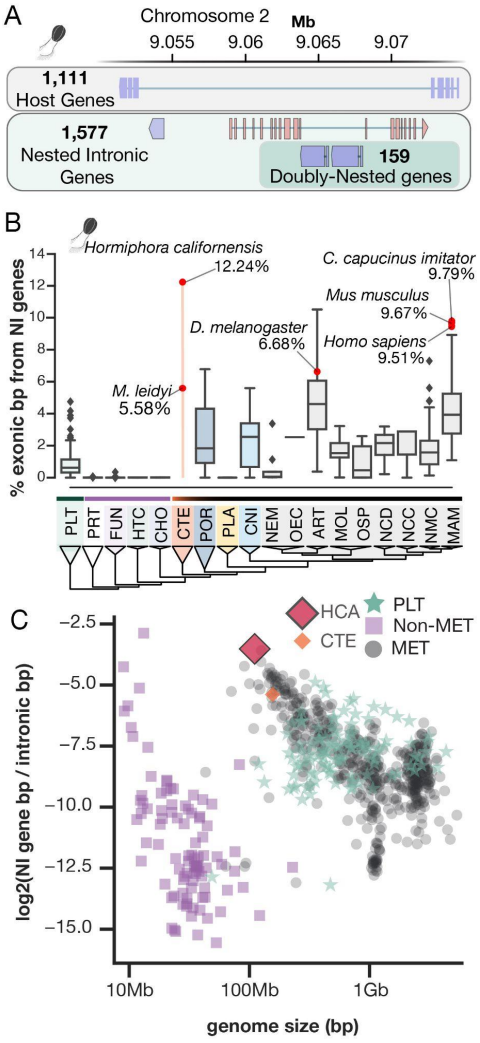

Figure 2.4 - An analysis of nested intronic genes in 1058 eukaryote genomes. (A) Nested intronic (NI) genes are genes that are contained within the introns of other genes. In this example from the *H. californensis* genome, four genes are nested within a single host gene. Two of the genes are doubly-nested — they are within the intron of another nested intronic gene. Blue genes are coded on the antisense strand relative to the reference sequence and red genes are coded on the sense strand. (B) The percent of all exonic bp that are in NI genes. Protists and non-metazoans have a negligible proportion compared to animals. Abbreviations: ART - panarthropoda, CHO - choanoflagellates, CNI - cnidarians, CTE - ctenophores, FUN - fungi, HCA - *Hormiphora californensis*, HTC - holozoa through choanoflagellates, MAM - mammals, MET - Metazoa, MOL - molluscs, NCD - non-chordate deuterostomes, NCC - non-craniata chordates, NEM - nematodes, NMC - non-mammalian chordates, OEC - other ecdysozoa, OSP - other spiralians, PLA - placozoans, PLT - plants, POR - Porifera, PRT - other protists. (C) As animal and protist genomes become smaller, a higher proportion of the exonic bases are in NI genes.

Acknowledgements

We thank the crew, ROV pilots, and ship captains of the R/V Rachel Carson and R/V Western Flyer for assistance in collecting animals. We would also like to thank Daniel Rokhsar, Karolin Luger, Joseph Ryan, Manabu Bessho-Uehara, and May Roberts for critical feedback on the manuscript. We also thank Wyatt Patry at the Monterey Bay Aquarium for providing *Hormiphora* tentacles. We thank Ben Abrams of the UCSC Life Sciences Microscopy Center for his suggestions for producing chromosome images.

Funding

This work was supported by the David and Lucile Packard Foundation, the Monterey Bay Aquarium Research Institute, the University of California Biomolecular Engineering and Bioinformatics department; NSF DEB-1542679 to SHDH; United States National Science Foundation GRFP DGE 1339067 to DTS.

Author Contributions

D.T.S., W.R.F., R.E.G., and S.H.D.H. conceived and designed the study. D.T.S. and L.M.C. collected sequencing data. D.T.S assembled the genome and transcriptomes. D.T.S and W.R.F. annotated the genome. D.T.S., W.R.F., and J.D.M. performed analyses on the data. D.T.S, W.R.F., J.D.M., and R.E.G. wrote the manuscript. All authors contributed to the revision and review of the manuscript.

Data and Materials Availability

All sequencing reads, transcriptomes, and the genome assembly are available for download via NCBI BioProject PRJNA576068. Hc1 and Hc2 mitochondrial sequences are available through NCBI accessions MN544300 and MN544301. Additional data are available through G3 figshare. Custom programs used in this publication, and annotation guidelines, are available at github.com/conchoecia/hormiphora and Zenodo DOI: 10.5281/zenodo.4074309. No tissue from these samples is available, as it was consumed during library preparation.

Chapter 3

Deeply conserved syntenies show that ctenophores are sister to other metazoa

Abstract

One of the last major outstanding phylogenetic questions in early Metazoan evolution is determining the sister phylum to all other animals. Answering this question is critical to understanding the emergence of animal traits, such as neurons, developmental cell types, and animal-specific genes and regulatory pathways. Phylogenetic studies have identified ctenophores or sponges as the phylum sister to the rest of animals (Dunn *et al.* 2008; Philippe *et al.* 2009), but have not been able to resolve which is the sister phylum with certainty (Li *et al.* 2021). Comparing chromosome-scale genome assemblies, and specifically tracking the fate of ancestral gene linkages (Simakov *et al.* 2020), is a promising technique that may provide phylogenetically-diagnostic information in resolving the ctenophore-sister or sponge-sister question. However, a lack of chromosome-scale genomes for non-bilaterian animals and unicellular outgroups to the Metazoa has made this approach intractable.

Here, we scaffold to chromosome-scale the genomes of three unicellular Holozoan species that are outgroups to animals. These species are *Creolimax fragrantissima* (Opisthokonta; Holozoa; Ichthyosporea; Ichthyophonida), *Capsaspora*

owczarzaki (Opisthokonta; Holozoa; Filozoa; Filasterea; Ministeriida), and *Salpingoeca fragrantissima* (Opisthokonta; Holozoa; Choanoflagellatea; Choanoflagellida). Together with our chromosome-scale ctenophore genome assemblies from *Hormiphora californensis* and *Bolinopsis infundibulum*, and the chromosome-scale genomes of the sponge *Ephydatia muelleri* and several cnidarian and bilaterian genomes, we identify twenty three linkage groups of genes traceable to the ancestor of the Filazoa. By tracing the irreversible fusion-with-mixing of ten these linkage groups, we find that ctenophores are the sister phylum to a monophyletic clade containing sponges, placozoans, cnidarians, and bilaterians.

Introduction

Determining the phylogenetic position of sponges and ctenophores is paramount to understanding the evolution of animals. The evolutionary provenance of neurons, the animal development programme, and the evolution of cell types are unclear without being able to root the animal tree of life. Phylogenomic studies leveraging large protein supermatrices have repeatedly found that sponges or ctenophores are the sister clade to bilaterians, cnidarians, and placozoans (Dunn *et al.* 2008; Philippe *et al.* 2009; Simion *et al.* 2017; Whelan *et al.* 2017). However, the same protein supermatrices can produce trees that support either ctenophores or sponges as the sister phylum to the rest of the Metazoa depending on which phylogenetic model is used (Li *et al.* 2021). These hypotheses are also referred to as the ctenophore-sister (or ctenophore-early) and sponge-sister (or sponge-early) hypotheses. It is unclear whether phylogenomics will be able to resolve these

hypotheses is the most likely. Furthermore, fossils cannot be used to resolve this question, as ctenophore fossils are non-existent or dubious in ctenophores, and sponge fossils from the Precambrian, the period in which ctenophores and sponges diverged, are exceedingly rare (Antcliffe *et al.* 2014).

One unexplored approach for determining which phylum is the sister to other animals is in comparing chromosomal macrosynteny, or finding megabase-scale homologous chromosome regions between species. It is possible to find regions of macrosynteny by plotting orthologous pairs of proteins between two species on an x-y axis corresponding to genome coordinates. This is called an Oxford dot plot. In these plots, regions of macrosynteny are detectable between chromosomes that have many pairs of homologous proteins. Comparing blocks of macrosynteny between many species enables the determination of ancestral karyotypes, or ancestral groups of linked genes (Putnam *et al.* 2008). It is possible to use these ancestral groups of genes to define how extant species' chromosomes formed (Simakov *et al.* 2020).

There are three fundamental mechanisms that can change a chromosome's sequence and composition. Chromosomes can split into multiple pieces, fuse with other partial or complete chromosomes, or undergo internal inversions in which the gene order is changed. The availability of chromosome-scale genome assemblies of Metazoans has shown that genes tend to stay on the same homologous chromosomes, even after hundreds of millions of years of divergence (Lv *et al.* 2011). In other words, animal chromosomes mostly change by internal inversions that only change

the gene order. However, chromosomes do not often split into multiple smaller chromosomes or fuse with other chromosomes.

One combination of events can cause two groups of genes on separate chromosomes to irreversibly fuse with one another. When pieces of chromosomes A and B fuse, then undergo extensive inversions, the genes from A and B are interlaced. It is incalculably improbable for the new A+B chromosome to undergo a series of inversions, then a fission, that would restore the original separate A and B chromosomes. Because the reversal of this character is highly unlikely, it is possible to use these events to polarize the evolutionary relationship between a known outgroup species and two or more ingroup species. For example, consider the scenario in which a the known outgroup species has gene groups A and B on separate chromosomes, and a known ingroup has the same groups of gene A and B on separate chromosomes, but a third clade of organisms has groups of genes A and B on single chromosomes. If we did not know the root of the phylogenetic relationships between all of these species, we could determine whether the A and B-split state, or the A and B-fused state was ancestral. However, since we know that the outgroup has gene groups A and B on separate chromosomes, and that chromosomal fusion-then-mixing is irreversibile, we can conclude that gene groups A+B mixed on a single chromosome is a derived state. It is important to use chromosome-scale genomes for such analyses, and not genomes with fragmented genomes, in order to properly characterize fusions and fissions.

It has not been possible to use the simple mechanism of fusion-then-mixing to characterize early animal evolution because of the lack of chromosome-scale genomes from non-bilaterian animals and the closest unicellular outgroups to animals. The chromosome-scale genomes of two cnidarian species, *Rhopilema esculentum* and *Nematostella vectensis (Li et al. 2020; Nong et al. 2020)*, along with the chromosome-scale genome of one sponge, *Ephydatia muelleri* (Kenny *et al.* 2020), were published in 2020. The genome assembly of the placozoan, *Trichoplax adhaerens*, appears to have some chromosome-scale scaffolds (Srivastava *et al.* 2008), but it is unclear without chromatin-capture proximity-ligation (Hi-C) sequencing data. The only published chromosome-scale genome assembly and annotation of a ctenophore is that of *Hormiphora californensis* (Schultz *et al.* 2021). There are several published genome assemblies of species that are unicellular outgroups to animals, including the choanoflagellate *Salpingoeca rosetta (Fairclough et al. 2013)*, the filasterian *Capsaspora owczarzaki (Suga et al. 2013)*, and the icthyosporean *Creolimax fragrantissima* (de Mendoza *et al.* 2015). While the genomes are very contiguous relative to the genome assembly size, it is unclear if these genomes are chromosome scale due to the lack of Hi-C data.

Here, we scaffold four genomes to chromosome-scale, identify ancestrally-linked groups of genes common to the Filozoa, and find phylogenetically diagnostic fusion-with-mixing events to determine which phylum is sister to the rest of the Metazoa. We first produced chromosome-scale genome assemblies of the ctenophore *Bolinopsis infundibulum,* and of the unicellular animal outgroup species

Salpingoeca rosetta, Capsaspora owczarzaki, and *Creolimax fragrantissima* using new chromatin capture data (Hi-C) data and the already-published assemblies. Then, we compared the chromosome-scale genomes of the unicellular outgroups to published chromosome-scale non-bilaterian genomes and identified 14 linkage groups of genes that existed in the ancestor to the Filazoa (filasterians, choanoflagellates, and animals). We first identified the linkage groups by performing a 4-way reciprocal best blast hit analysis on the proteins of *Capsaspora*, *Hormiphora*, *Rhopilema*, and *Ephydatia*. Using Hidden Markov Models, we expanded this geneset to analyze the present and location of these genes in five more chromosome-scale genomes. We found that several of the FLGs correspond with regions of macrosynteny between the *Capsaspora* genome and metazoan genomes. We evaluated the hypotheses that can explain the fate of these linkage groups and found 5 irreversible fusion-then-mixing events that link sponges, cnidarians, placozoans, and bilaterians into a monophyletic clade to the exclusion of ctenophores. Our analyses show that ctenophores are the sister phylum to the rest of animals.

Main Text

Hi-C library preparation

With the goal of scaffolding the published genome assemblies to chromosome-scale, we prepared two Hi-C libraries per species from *Salpingoeca rosetta* (ATCC® PRA-366™), *Capsaspora owczarzaki* (ATCC® 30864™), and *Creolimax fragrantissima* (ATCC® PRA-284™). The American Type Culture Collection samples that we used are the same accessions that were sequenced for the published genome assemblies for those three species (Suga *et al.* 2013; Fairclough *et al.* 2013; de Mendoza *et al.* 2015). Each library was prepared using 0.125 mL of cell culture stock with the protocol described in (Adams *et al.* 2020)). We also prepared Hi-C libraries from a single F3 *Bolinopsis infundibulum* ctenophore reared at the Monterey Bay Aquarium, from a source population collected in the Monterey Bay, California. For each species we prepared one library with the enzyme DpnII and one library with MluCI. For *B. infundibulum* we also prepared a Hi-C library with the FatI enzyme. These libraries were sequenced on a NovaSeq 6000 with 2x150 cycles to a depth of over 500x for each species (Table 1). Assessing the quality of the libraries showed that between 10% and 44% of the individual reads had linker sequences of GATCGATC or AATTAATT that indicate a captured *in vitro* ligation event.

Bolinopsis infundibulum genome sequencing and assembly

We isolated DNA from the *Bolinopsis infundibulum* ctenophore using a previously published protocol (Schultz *et al.* 2021). One PacBio CLR library was constructed with the DNA at the Brigham Young University Sequencing Center, and the library was sequenced on two 15-hour SMRT cells. We constructed two *B. infundibulum* Illumina whole-genome shotgun libraries at the University of California Santa Cruz using the NEBNext Ultra II FS DNA Library Prep Kit. The Illumina WGS libraries were sequenced at MedGenome, Inc.

The genome was assembled and scaffolded using the published protocol that was used to assemble the *Hormiphora californensis* genome (Schultz *et al.* 2021). The assembly, scaffolding, and polishing was all performed using reads from the single individual collected from the Monterey Bay Aquarium. Scaffolds that appeared to be bacterial contamination were removed. However, unplaced scaffolds were not removed using Purge Haplotigs (Roach *et al.* 2018) as was performed during assembly of the *H. californensis* genome.

The final *Bolinopsis infundibulum* genome assembly was 277.5 Mb in 774 scaffolds in 985 contigs. 92.99% of the bases of the assembly were contained in 13 scaffolds and 224 contigs. 50% of the assembly was in scaffolds of at least 20.1 Mbp in length. The high number of scaffolds is likely due to the lack of a haplotig purging step. Hi-C maps were constructed as previously described (Schultz *et al.* 2021). The Hi-C map showed that 13 largest scaffolds inthe *B. infundibulum* genome were consistent with chromosome-scale scaffolds of other species (Figure 3.1).

Species	Data Type	Total Gb	Number of reads (or pairs)	% Linker	Read Depth Coverage
Salpingoeca rosetta ATCC® PRA-366™	HiC - DpnII	12.0 Gb	40.1 M	44.11 %	217 x
	HiC - MluCI	15.9 Gb	53.1 M	17.28 %	288 x
Capsaspora owczarzaki ATCC® 30864™	HiC - DpnII	22.6 Gb	75.4 M	40.2 %	821 x
	HiC - MluCI	6.3 Gb	20.8 M	16.69 %	226 x
Creolimax fragrantissima ATCC® PRA-284™	HiC - DpnII	8.1 Gb	27.1 M	40.33 %	181 x
	HiC - MluCI	15.3 Gb	50.9 M	10.31 %	340 x
Bolinopsis infundibulum	Hi-C DpnII	90.5 Gb	301 M	21.94%	350 x
	Hi-C MluCI	50.5 Gb	168 M	13.97%	196 x
	Hi-C FatI	57.1 Gb	190 M	1.71%	221 x
	Illumina WGS	15.6 Gb	52.0 M	NA	60 x
	PacBio CLR	52.3 Gb	7.05 M	NA	203 x

Table 3.1 - Sequencing libraries produced for this study. Each species had one DpnII and one MluCI library. The read depth was over 500x genome coverage for each species.

Macrosynteny comparisons

We wrote a software package called `odp` (Oxford Dot Plots) to perform chromosome-scale synteny analyses (github.com/conchoecia/odp). The software uses blastp or diamond blastp (Buchfink *et al.* 2015) to find reciprocal-best-hit proteins between species, generates chromosome-scale synteny plots using the reciprocal-best-hit proteins, and calculates the statistical significance of those syntenic relationships using Fisher's exact test (Simakov *et al.* 2020). The software also calculates a rolling-window measure of breaks in synteny, called D (Simakov *et al.* 2020), for each scaffold.

Scaffolding Existing Genomes

We scaffolded the existing genome assemblies for *S. rosetta* (GCF_000188695.1), *C. owczarzaki* (GCF_000151315.2), and *C. fragrantissima* (GCA_002024145.1) using the Hi-C sequencing reads with Dovetail Genomics HiRise (v Aug 2019). Hi-C heatmaps were prepared using the protocol described in (Schultz *et al.* 2021). Inversions from genome misassembly, and assembly misjoins, were identified using the Hi-C heatmaps. Erroneous sequence inversions were corrected by replacing the region with the reverse complement of that region. Assembly misjoins were split at the nearest gap of Ns. In the *C. owczarzaki* assembly we found and removed large regions and scaffolds that had no Hi-C reads that mapped, including one megabase-scale region of a scaffold and one

chromosome-scale scaffold. These regions were likely due to contamination in the original sequencing data, given that these regions had no Hi-C reads that mapped. We generated coordinates of gene positions along the chromosome-scale scaffolds by mapping the transcripts from the published genome annotation to the new chromosome-scale scaffolds with minimap2 v2.17 (Li 2017).

These libraries allowed us to scaffold the already-published genome assemblies to chromosome-scale. The Hi-C data also allowed us to identify several megabases of the original *Creolimax fragrantissima* assembly as contamination from another source, possibly a co-cultured organism. We identified the general location of the centromeres in *Creolimax fragrantissima* and *Capsaspora owczarzaki* using the Hi-C data. We found that *Salpingoeca rosetta* has 36 chromosomes, *Capsaspora owczarzaki* has 16 chromosomes, and *Creolimax fragrantissima* has 26 chromosomes (Figure 3.1). Comparing the published and new chromosome-scale assemblies revealed that the published assemblies of these three species were nearly chromosome-scale. Each genome required fewer than 15 joins to form the complete chromosomes.

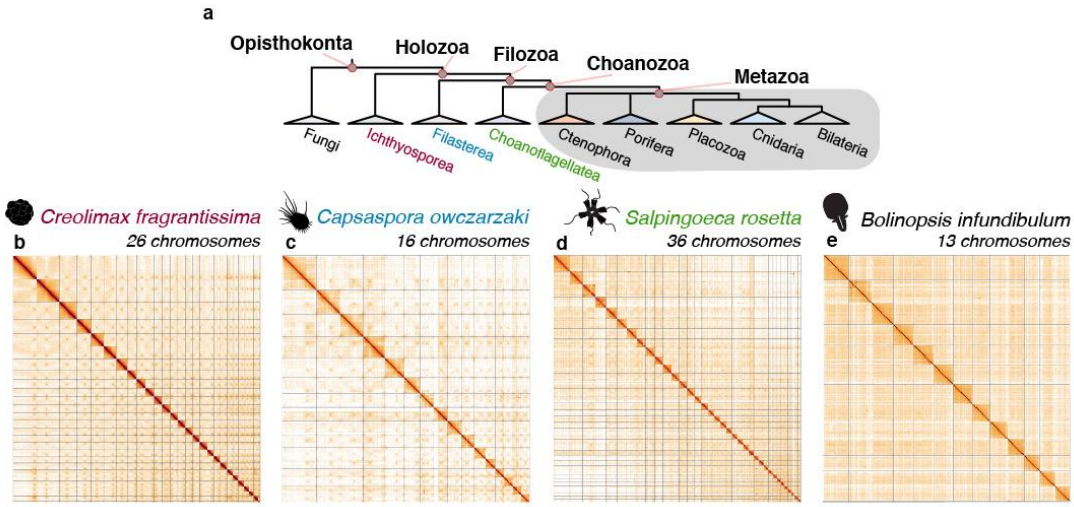

Figure 3.1 - New Chromosome-Scale Genomes. (a) The species sequenced in this study are unicellular, and are outgroups to the Metazoa. *Creolimax fragrantissima* is an ichthyosporean, *Capsaspora owczarzaki* is a filasterian, and *Salpingoeca* is a choanoflagellate. Together, these organisms, along with the animals, form a clade called the Holozoa. (b-e) The Hi-C heatmaps of the chromosome-scale presented in this manuscript are consistent with the pattern seen from Hi-C maps of chromosome-scale assemblies of other species, in which the inter-scaffold Hi-C signal is much stronger than any inter-scaffold Hi-C signal.

The conserved ctenophore karyotype and animal macrosynteny

We used odp to identify if the karyotype of 1n=13 in the *Hormiphora californensis* genome was consistent among ctenophores. We found that the *Bolinopsis infundibulum* genome assembly contained 13 chromosome-scale scaffolds that had 1-to-1 homology with the 13 *Hormiphora californensis* chromosomes (Figure 3.2, p<<1e-9). *Hormiphora* and *Bolinopsis* are estimated to have a most recent common ancestor dating approximately between 160-260 Mya (Whelan *et al.* 2017), and this clade contains much of known ctenophore diversity. Given the long divergence time, the karyotype conservation in ctenophores is remarkably high.

The conservation of karyotype is also high between bilaterians, sponges, and cnidarians (Figure 3.2). The genomes of the cnidarian *R. esculentum*, the sponge *E. muelleri*, and the bilaterian *B. floridae* can easily be described with in terms of another species' genome using fewer than 15 chromosomal fusions or fissions.

Using odp we found that the genomes of the ctenophores are highly rearranged relative to cnidarians, sponges, and bilaterians (Figure 3.2). Some ctenophore chromosomes have clear homology to one, two, or three chromosomes in a cnidarian, bilaterian, or sponge. However, there are some chromosomes in the sponge, the cnidarian, or the bilaterian genomes that have no clear homologs in ctenophores. For example, the sponge *Ephydatia muelleri* chromosomes 2, 7, 24, 17, 6, and 5 have no clear homologous chromosomes in ctenophores, despite have easily-identifiable homologous chromosomes in cnidarians or bilaterians.

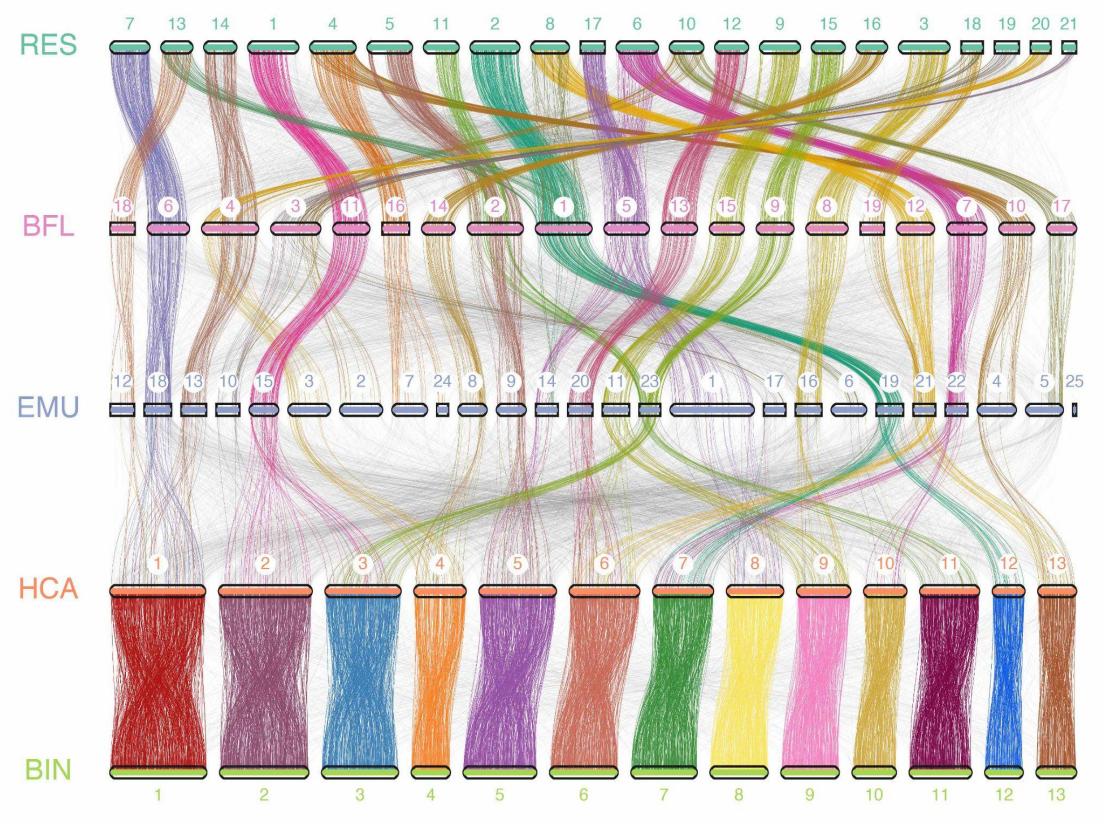

Figure 3.2 - Macrosynteny in the animals. Each row is the genome of one organism. Labelled bars are chromosomes with ranked gene coordinates. Lines connecting bars represent a single reciprocal-best blastp hit, likely orthologous proteins. Lines colored in (RES, BFL), (BFL, EMU), and (EMU, HCA) are colored based on significant gene groupings found in the sponge, cnidarian, and bilaterian genomes. In the ctenophores, only genes on orthologous chromosomes are colored. The ctenophores HCA and BIN share 13 homologous chromosomes. The cnidarian RES, the lancelet BFL, and the sponge EMU share many homologous chromosomes in very few rearrangements. The ctenophores are highly rearranged relative to the sponges, cnidarians, and bilaterians.

Identification of Filazoan Linkage Groups

We sought to identify groups of genes that are both highly conserved among diverse species in the Filozoa, and that persist on the same set of chromosomes in those species. Finding such groupings of genes provides evidence for ancestrally-linked groups of genes that were present in the common ancestor to the species in question (Lv *et al.* 2011; Simakov *et al.* 2020). In addition, finding sets of genes with common fusions in two species to the exclusion of others is phylogenetically informative, if the fissions or fusions are shown to be irreversible.

We first performed 4-way reciprocal-best-hit blastp searches between all combinations of *Hormiphora californensis* (HCA), *Capsaspora owczarzaki* (COW), *Ephydatia muelleri* (EMU), and *Rhopilema esculentum* (RES). We also performed the same analysis, but substituted *Rhopilema* with the cnidarian *Nematostella*, the placozoan *Trichoplax*, or the bilaterian *Branchiostoma*. For a set of genes to be retained as a 4-way reciprocal-best blastp hit, we required that each gene had the same best blastp hit to the same gene in all other species. Each set of reciprocal-best blast hits is a complete directed graph, in which the genes are nodes, and the edges are the top blast hits for each gene. There were between 1804 and 1862 gene sets that were identifiable four-way reciprocal-best hits from each analysis using this method. For the group HCA-COW-EMU-RES, there were 1856 four-way reciprocal-best gene groups. We then grouped reciprocal-best gene sets together based on their occupancy on the same combination of chromosomes in the four species, and used a randomized

permutation test to consider only groups of genes with a low false discovery rate ($\boldsymbol{\alpha} \leq 0.001$, see Supplementary results).

We found that this technique performed on COW-HCA-EMU-RES recovered significant groupings of genes that appeared on the same chromosomes in the four species ($\boldsymbol{\alpha} \leq 0.0003$, Table 2). Based on the pairings of EMU and RES chromosomes, we found groups of genes in COW-HCA-EMU-RES that corresponded to the ancestral linkage groups (ALGs) of genes A1a, B1, C1, C2, D, Ea, F, G, H, I, J1, J2, K, L, M, and N, first identified in (Simakov *et al.*). We note that the false discovery rate, $\boldsymbol{\alpha}$, of finding 6 or more genes on the same group of four chromosomes in four species was incalculably low after even 100 million permutation tests. Therefore, some of the reported $\boldsymbol{\alpha}$ values are maximum bounds (Table 3.2). Given that the clade including *Capsaspora* and animals is the filozoa, we call these significantly large groupings of genes filozoan linkage groups (FLGs). These represent groups of genes that appear to have been linked on the same chromosomes since the ancestor of the Filozoa.

Information Content	ALG	COW	HCA	EMU	RES	FLG/ALG	Number of Genes	alpha
exclusive sponge-cnidarian grouping	A1a	COW3	HCA7	EMU19	RES2	A1a_x	14	$\alpha < 2.89e\text{-}04$
		COW5	HCA12	EMU19	RES2	A1a_y	5	$\alpha = 2.89e\text{-}04$
	C1	COW1	HCA2	EMU9	RES5	C1_x	5	$\alpha = 2.89e\text{-}04$
		COW3	HCA2	EMU9	RES5	C1_x	6	$\alpha < 2.89e\text{-}04$
		COW14	HCA5	EMU9	RES5	C1_y	5	$\alpha = 2.89e\text{-}04$
		COW3	HCA5	EMU9	RES5	C1_y	8	$\alpha < 2.89e\text{-}04$
	Ea	COW3	HCA2	EMU1	RES17	Ea_x	7	$\alpha < 2.89e\text{-}04$
		COW4	HCA8	EMU1	RES17	Ea_z	5	$\alpha = 2.89e\text{-}04$
	F	COW1	HCA7	EMU22	RES6	F_x	5	$\alpha = 2.89e\text{-}04$
		COW8	HCA10	EMU22	RES6	F_y	5	$\alpha = 2.89e\text{-}04$
	G	COW10	HCA2	EMU15	RES1	G_x	8	$\alpha < 2.89e\text{-}04$
		COW12	HCA3	EMU15	RES1	G_y	6	$\alpha < 2.89e\text{-}04$
	N	COW2	HCA6	EMU21	RES8	N_x	7	$\alpha < 2.89e\text{-}04$
		COW2	HCA13	EMU21	RES8	N_y	8	$\alpha < 2.89e\text{-}04$
exclusive animal grouping	B1	COW1	HCA13	EMU4	RES4	B1	6	$\alpha < 2.89e\text{-}04$
		COW3	HCA13	EMU4	RES4	B1	7	$\alpha < 2.89e\text{-}04$
	H	COW4	HCA6	EMU20	RES12	H	10	$\alpha < 2.89e\text{-}04$
		COW6	HCA6	EMU20	RES12	H	5	$\alpha = 2.89e\text{-}04$
	I	COW13	HCA1	EMU13	RES14	I	5	$\alpha = 2.89e\text{-}04$
		COW4	HCA1	EMU13	RES14	I	6	$\alpha < 2.89e\text{-}04$
		COW9	HCA1	EMU13	RES14	I	7	$\alpha < 2.89e\text{-}04$
	K	COW4	HCA3	EMU23	RES15	K	7	$\alpha < 2.89e\text{-}04$
		COW6	HCA3	EMU23	RES15	K	8	$\alpha < 2.89e\text{-}04$
uninformative	C2	COW5	HCA2	EMU10	RES21	C2	5	$\alpha = 2.89e\text{-}04$
	D	COW12	HCA1	EMU18	RES7	D	11	$\alpha < 2.89e\text{-}04$
	J1	COW1	HCA11	EMU5	RES10	J1	5	$\alpha = 2.89e\text{-}04$
	J2	COW11	HCA11	EMU23	RES11	J2	6	$\alpha < 2.89e\text{-}04$
	L	COW1	HCA9	EMU11	RES9	L	7	$\alpha < 2.89e\text{-}04$
	M	COW4	HCA9	EMU16	RES3	M	6	$\alpha < 2.89e\text{-}04$

Table 3.2 - Four-way reciprocal best blast hit results between COW-HCA-EMU-RES. Each row is a group of genes that was present on the same set of chromosomes in COW-HCA-EMU-RES in a four-way reciprocal best blast search. Only rows with a significant false discovery rate ($\alpha < 0.001$, or at least 5 genes) are shown. The rows labelled "exclusive sponge-cnidarian grouping" includes

rows that groups sponges and cnidarians together to the exclusion of ctenophores and *Capsaspora*. There were no rows that grouped ctenophores and cnidarians together to the exclusion of sponges and *Capsaspora*. Rows labelled "exclusive animal grouping" are sets of genes that are grouped on the same chromosomes in the metazoans, but occurred on separate chromosomes in *Capsaspora*. Rows labelled "uninformative" had significant numbers of genes, but were unique chromosome configurations that were not shared with other rows. Scaffold columns are colored pastels or grays if they share common chromosomes. Rows are colored red or yellow if they are phylogenetically informative fusions-with-mixing, and correspond with Figure 3.3. The red and yellow components on COW3 for ALG C1 are located on non-overlapping intervals of COW3. Note that our permutation test for false discovery rates (α) found a maximum of four genes grouped together, so $\alpha < 1e^{-4}$ is a conservative upper limit to false discovery rate for chromosome groups with more than 5 genes. (End of caption for Table 3.1)

Several of the FLGs found in this analysis were not phylogenetically informative in that they had no pairs of shared chromosomes with other gene groups. This includes gene groupings corresponding to the Simakov *et al.* ALGs C2, D, J1, J2, L, and M were singletons.

Other FLGs shared the same chromosomes in HCA-EMU-RES to the exclusion of *Capsaspora*. The ALGs B1, H, I, and K fall in this category. These gene groupings indicate possible chromosome fusion events in the ancestor of metazoans.

Because this analysis does not include an outgroup to *Capsaspora*, it is also possible that these events represent lineage-specific chromosome fissions in *Capsaspora*. Because the chromosome configurations are shared between ctenophores, sponges, and cnidarians, these gene groupings are uninformative for resolving whether sponges or ctenophores are the sister to other animals.

The remaining FLGs, corresponding to the Simakov *et al.* ALGs A1a, C1, Ea, F, G, and N, have genes that are shared on single cnidarian or sponge chromosomes, but exist on separate ctenophore chromosomes and on separate *Capsaspora* chromosomes. We designate these groups as _x or _y subgroupings of single FLGs. For example, the genes on FLG A1a can be split into two gene sets, A1a_x and A1a_y. Both A1a_x and A1a_y reside on single chromosomes in EMU and RES, but A1a_x and A1a_y are on separate chromosomes in HCA and COW. Given that *Capsaspora* is the known outgroup species to all animals, these gene groupings can help polarize the evolutionary relationships between sponges, ctenophores, and the rest of animals. None of the FLGs existed on the same pairs of chromosomes in HCA-RES, but on the separate pairs in EMU-COW (Table 4.1).

The extent of mixing of the FLGs

To visualize the extent of mixing of the FLGs in the sponge, cnidarian, and bilaterian genomes, we plotted the coordinates of the FLG genes from the four-way COW-HCA-EMU-RES reciprocal-best blastp analysis in all four species (Figure 3.2). We also quantified the extent of mixing by calculating the percent of each chromosome that the _x and _y FLGs occupied, and the percent of each chromosome covered by a region which genes from _x and _y were interlaced.

All six FLGs with possibly phylogenetically-informative pairings, A1a, C1, Ea, F, G, and N, had the _x and _y components of the ALGs on single separate chromosomes in the genome of the cnidarian *R. esculentum* and of the sponge *E. muelleri*. The analysis of the extent of mixing of FLGs revealed that the total percent of genes from _x and _y that are mixed into one another are 76% of the *R. esculentum* genes, and 47%% of the *E. muelleri* genes. In other words, the _x and _y components of the FLGs are better mixed on the cnidarian chromosomes than on the sponge chromosomes. In addition, the average percent of the chromosomes that are covered by the interlacing portions of _x and _y are 22 % in *Ephydatia*, but 61 % in *Rhopilema*. This means that the _x and _y genes for FLGs in the sponge are not dispersed along the chromosomes as much as in the cnidarian. In one instance, despite all of the genes from the Ea_x and Ea_y components being on the same chromosome in *Ephydatia*, the genes are on opposite sides of the chromosome.

FLG edge cases

We found that C1_x and C1_y genes on *Capsaspora* chromosome 3 did not have overlapping coordinates and were discrete regions. Otherwise, C1_x genes exist on COW1, and C1_y genes exist on COW14. Given that *Capsaspora* is the known outgroup to animals, it is unclear whether there was a derived translocation in *Capsaspora* between ancestrally separate _x and _y groups of genes, or if there was a fission of COW3, followed by fusions with COW1 and COW14 to form the C1_x and C1_y groups before the common ancestor of Metazoa or of the Choanozoa. Given that we observe statistically significant groupings of C1_x and C1_y genes on completely separate chromosomes in COW1 and COW14, and that these groupings appear in *Hormiphora*, these findings do not refute the hypothesis of ctenophore-sister.

The N_x and N_y genes on COW2 overlapped on 3.8% of the length of COW2. Three out of seven genes from N_y were interlaced with two out of seven genes from N_x. Similarly to C1_x and C1_y on COW3, given that *Capsaspora* is an outgroup to animals it is unclear whether the presence of N_x and N_y on COW2 is a derived fusion or the ancestral state. In *Ephydatia*, the cnidarians, and in *Branchiostoma*, N_x and N_y are located on the same chromosome, and are well-mixed. Given the lack of conserved synteny in *Salpingoeca*, it is possible that the fission of N_x and N_y in *Hormiphora* is derived. These possibilities are not informative for determining whether sponges or ctenophores are the sister phylum to animals, but do not rule out either hypothesis.

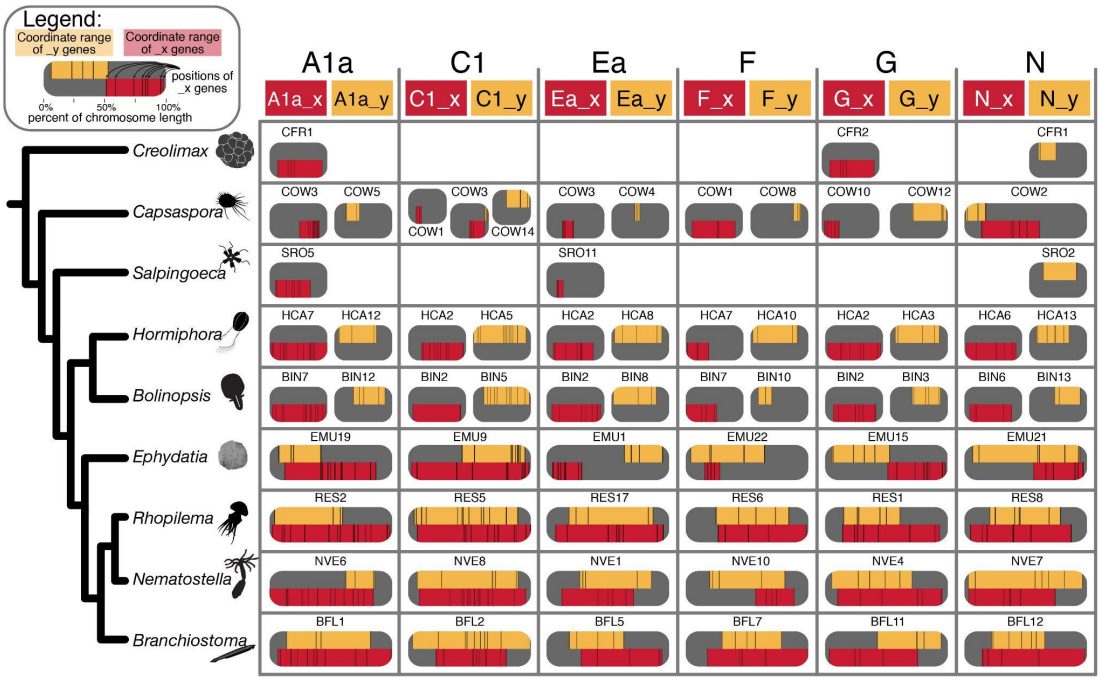

Figure 3.3 - The fate of Filasterean linkage groups. Columns separated by lines are linkage groups of genes identified by four-way reciprocal best blastp search with COW-HCA-EMU-RES, defined by the ALG nomenclature of Simakov et al. (2021 - in review). The genes in *Creolimax*, *Salpingoeca*, *Bolinopsis*, *Nematostella*, and *Branchiostoma* were identified using a HMM of the proteins identified from the four-way COW-HCA-EMU-RES reciprocal best blastp search. The _x linkage groups are colored red, and the _y linkage groups are colored yellow. The grey bars represent the chromosomes on which genes from the _x and _y components exist. Within each chromosome, the minimum and maximum position of the _x component genes are depicted as red bars, and the minimum and maximum position of the _y component genes are depicted as yellow bars. Vertical dark lines within each red and yellow bar show the coordinates of the each gene in that group. This plot shows that the _x and

_y components of genes from A1a, C1, Ea, F, G, and N exist on the same chromosomes in the sponges, cnidarians, and bilaterians, but exist on separate chromosomes in ctenophores and unicellular outgroup species to animals. On COW3, some genes from both C1_x and C1_y are present, but on non-overlapping coordinates. On COW2, genes from N_y and N_x exist on the same chromosome, with mostly non-overlapping intervals. Missing groups in *Creolimax* and *Salpingoeca* indicate that we did not find significantly large groupings of the _x or _y components in those genomes. The false discovery rate of findings groupings of genes this large is less than $\alpha < 1e^{-4}$ for each of the depicted chromosomes. (End of Figure 3.3 Caption)

Sample	Dominant Chrom.	_x	_y	% Chrom Covered by _x	Mean % Chrom Covered by _x	% Chrom Covered by _y	Mean % of Chrom Covered by _y	% of Chrom Covered by Interlacing _x and _y	Mean % of Chrom Covered by _x and _y Interlacing Region	Genes in _x	Genes in _y	Num _x Genes Mixed	Num _y Genes Mixed	% _x Genes Mixed	Mean % of _x Genes Mixed	% _y Genes Mixed	Mean % _y Genes Mixed	Mean % of All Genes Mixed
COW	COW3	C1_x	C1_y	38.5	42.7	8.6	12.6	0.0	1.9	6	8	0	0	0.0	14.3	0.0	28.6	28.6
	COW2	N_x	N_y	47.0		16.7		3.8		7	7	2	4	28.6		57.1		
EMU	EMU19	A1a_x	A1a_y	74.2	48.8	34.3	51.8	29.2	21.9	15	5	1	4	6.7	43.2	80.0	51.3	47.3
	EMU9	C1_x	C1_y	95.7		51.1		51.1		11	13	6	13	54.5		100.0		
	EMU1	Ea_x	Ea_y	23.9		31.0		0.0		8	5	0	0	0.0		0.0		
	EMU22	F_x	F_y	12.0		60.0		12.0		5	5	5	2	100.0		40.0		
	EMU15	G_x	G_y	46.9		47.2		1.6		8	6	1	1	12.5		16.7		
	EMU21	N_x	N_y	40.3		87.4		37.5		7	7	6	5	85.7		71.4		
RESLi	RES2	A1a_x	A1a_y	96.1	84.8	55.6	62.0	55.6	61.3	15	5	7	5	46.7	58.6	100.0	94.1	76.3
	RES5	C1_x	C1_y	92.4		84.7		82.0		11	13	9	11	81.8		84.6		
	RES17	Ea_x	Ea_y	87.7		68.5		68.5		8	5	4	5	50.0		100.0		
	RES6	F_x	F_y	72.0		59.2		57.8		5	5	4	4	80.0		80.0		
	RES1	G_x	G_y	78.9		46.1		46.1		8	6	4	6	50.0		100.0		
	RES8	N_x	N_y	82.0		57.7		57.7		7	7	3	7	42.9		100.0		

Table 3.3 - Mixing of the _x and _y gene groups. Each row is a chromosome on which there was mixing of _x and _y genes. COW3 has no overlap of the range of genes from C1_x and C1_y, despite genes from both sets being present on the same chromosome. COW2 has two of seven genes in N_x interlaced with four of seven genes from N_y. On average, the _x components and _y components were better dispersed along the chromosomes in *Rhopilema* than *Ephydatia*. In addition, the mixed portions of the _x and _y groups cover a larger percentage of the chromosomes in *Rhopilema* than in *Ephydatia*. Lastly, on average, a higher percentage of genes in the _x and _y groups were interlaced in *Rhopilema* than in *Ephydatia*.

Evidence of Filozoan Linkage Groups using other taxa

From the four-way reciprocal best blast analysis we found that five of the ALGs can be described as significantly numerous ($\alpha < 1e^{-4}$) groups of genes that are mixed and present on the same chromosomes in *Rhopilema* and *Ephydatia*, but are on separate chromosomes, or non-overlapping regions of single chromosomes, in *Hormiphora* and *Capsaspora*. These ALGs are A1a, C1, Ea, F, and G. We next

sought to determine if the patterns of _x/_y gene colocalization or separation were consistent in other species. Because reciprocal-best blast searches are increasingly stringent for each additional species added, we used hidden Markov models (HMMs) constructed from each of the four-way reciprocal best blast hits to find the best matches in the genomes of other species. Using these HMMs we searched in the genomes of the uniceullular *Creolimax*, the unicellular *Salpingoeca*, the ctenophore *Bolinopsis*, the anthozoan cnidarian *Nematostella*, and the chordate bilaterian *Branchiostoma*.

We found that there were very few significant gene groupings on single chromosomes of *Creolimax* and *Salpingoeca* relative to the _x and _y components of ALGs A1a, C1, Ea, F, and G. Oxford dot plots of *Capsaspora*-*Creolimax* and *Salpingoeca*-*Creolimax* show that there are many rearrangements between these species, and no regions strongly evident of macrosynteny. Of the significant groupings of genes in *Creolimax* and *Salpingoeca* chromosomes, we did not find any ALGs for which we recovered genes from both the _x and _y components. Therefore, these species are not phylogenetically informative when considering only macrosynteny. We do note that the most significant group of genes that we recovered in both species was A1a_x. This indicates that this grouping of genes may have been present on a single chromosome in the ancestor of all Holozoans.

Using the HMMs to search for genes in the *Bolinopsis* genome recovered significant groupings of genes of all of the _x and _y components of the ALGs A1a, C1, Ea, F, G, and N. Each of these groupings of genes, _x and _y, occurred on

separate chromosomes. Each grouping of genes present on a *Bolinopsis* chromosome was also present on the orthologous chromosome in *Hormiphora*, and corresponded to one of the FLGs identified in the COW-HCA-EMU-RES reciprocal best blastp analysis.

Both *Hormiphora* and *Bolinopsis* have a karyotype of 13 chromosomes, and each chromosome has an orthologous chromosome in the other species. Given that the _x and _y gene components were on separate, orthologous, chromosomes in both ctenophores, the separation of these linkage groups onto separate chromosomes in the Ctenophora can be interpreted as as a generality for the phylum, rather than a derived state in the *Hormiphora*.

Significant _x and _y gene groupings on *Nematostella* and *Branchiostoma* chromosomes revealed that the _x and _y gene groups were colocalized on the same chromosomes in both species for ALGs A1a, C1, Ea, F, G, and N. In addition, the genes from the _x and _y groups were interlaced, and distributed over more than 75% of each chromosome on which those gene groups were found. This is the same pattern that was found in the cnidarian *Rhopilema* and the sponge *Ephydatia*.

Corroboration of FLGs with macrosynteny blocks in Oxford dot plots

We next sought to determine whether any of the FLGs that we identified with four-way reciprocal-best blastp searches corresponded to regions of macrosynteny visible from dot plots. To test this, we plotted a monochromatic *Capsaspora-Rhopilema* Oxford dot plot to visualize regions of macrosynteny

(Figure 3.4a). We then produced the same Oxford dot plot, but colored the dots by the FLG identified in the four-way reciprocal best blast analysis with COW-HCA-EMU-RES (Figure 3.4b).

The monochromatic plot (Figure 3.4a) clearly shows regions of macrosynteny between COW1-RES19, COW2-RES8, COW3-RES2, COW3-RES5, COW3-RES17, COW4-RES12, COW5-RES2, COW8-RES2, and possibly more. These appear as dense rectangles of many orthologous protein hits in the same region.

Surprisingly, each FLG, with the except of some components of C1, clearly corresponds to a macrosynteny block identifiable in Figure 3.4a. Notably, A1a_x corresponds to the large macrosynteny block on COW3-RES2, while A1a_y corresponds to the macrosynteny block on COW5-RES. *Capsaspora* chromosome 3 contains many regions with genes conserved in the ctenophores, sponges, and bilaterians, as it contains genes from FLGs Ea_x, C1_x, A1a_x, C1_x, and C1_y in sequential order (Figure 3.4). We found that in *Bolinopsis*, the FLGs occurred on the homologous chromosomes as in *Hormiphora* (Figure 3.5).

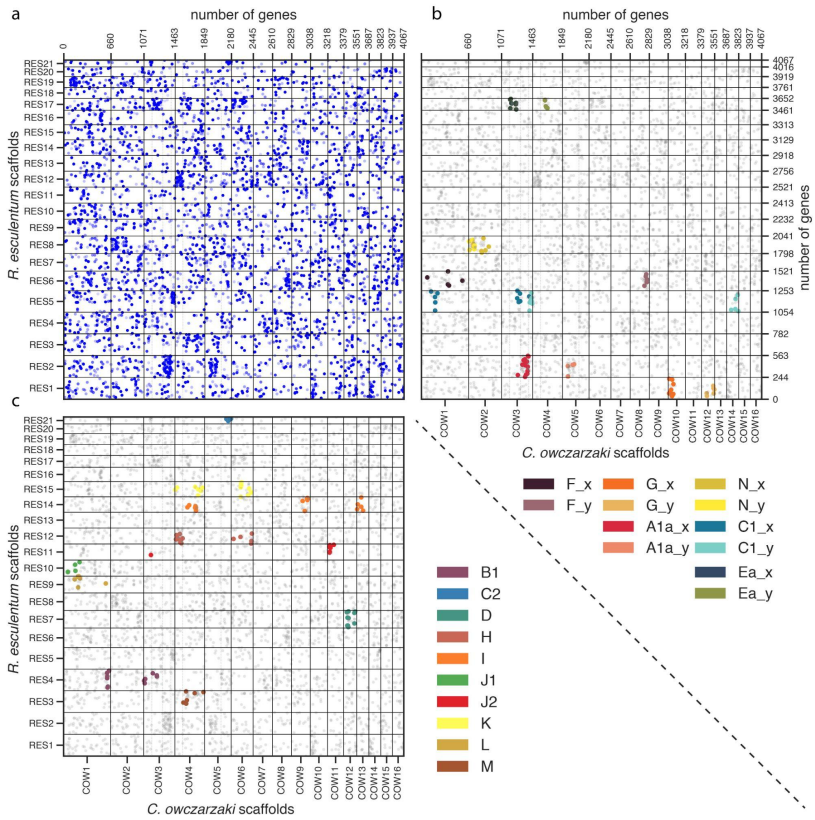

Figure 3.4 - Oxford dot plots of *Rhopilema* and *Capsaspora*. (a) Each blue dot is a reciprocal best blastp hit between a protein in the *Rhopilema* genome and a protein in the *Capsaspora* genome. There are rectangular regions of high density on COW1-RES19, COW2-RES8, COW3-RES2, COW3-RES5, COW3-RES17, COW4-RES12, COW5-RES2, COW8-RES2, and possibly more. (b) The same dotplot as panel **a**, but only genes from the four-way reciprocal best blastp search between COW-HCA-EMU-RES are shown. (c) The genes from the four-way COW-HCA-EMU-RES reciprocal best blast search that represent gene groupings present since the most recent animal ancestor, or since the ancestor of the filozoa. These gene linkages do not contain information in answering whether ctenophores or sponges are the sister phylum to other animals. (End of Figure 3.4 caption)

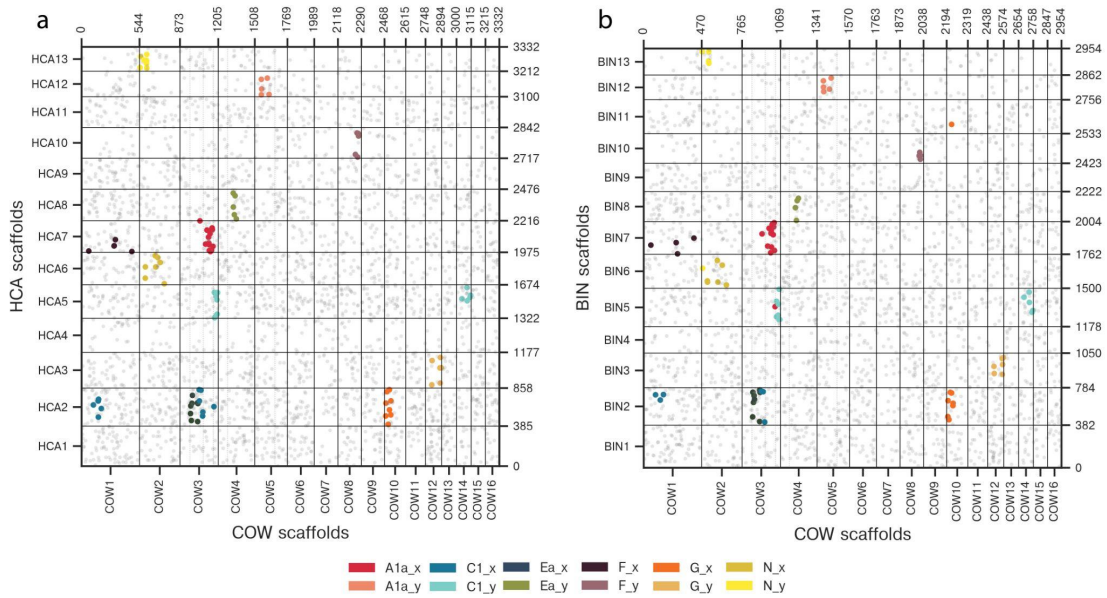

Figure 3.5 - Oxford dot plots of COW-HCA and COW-BIN. In both plots, genes from the COW-HCA-EMU-RES FLGs are colored, and other genes are gray. (a) The Oxford dot plot of COW-HCA. (b) The Oxford dot plot of COW-BIN. The FLG gene groupings occur on the same homologous chromosomes in BIN as they do in HCA.

Filozoan linkage groups in Trichoplax and Amphimedon

The genome assemblies of the placozoan *Trichoplax* and the sponge *Amphimedon* are not chromosome-scale. Despite this, we produced Oxford dot plots of *Capsaspora-Amphimedon* and *Trichoplax-Amphimedon* to see if genes from _x and _y ALG components were present on the same scaffolds. Finding _x and _y gene components on separate scaffolds is not a true positive that those components are on separate scaffolds, due to the genomes being fragmented. However, finding _x and _y

genes present on the same scaffold is evidence for an affinity to _x and _y fusions found on ALGs A1a, C1, Ea, F, G, and N in sponges, cnidarians, and bilaterians.

We found that the *Amphimedon* (Porifera) genome contained two scaffolds with _x and _y fusions (Figure 3.6). AQU1 contained genes from both C1_x and C1_y. AQU7 contained genes from both Ea_x and Ea_y. Similarly, in the *Trichoplax* (Placozoa) genome we found many genes from C1_x and C1_y present on TAD3, genes from Ea_x and Ea_y on TAD3, Ea_x and Ea_y on TAD5, A1a_x and A1a_y on TAD 6, and lastly we found G_x and G_y on TAD9 (Figure 3.7). This direct evidence that there are genes from the _x and _y components on the same scaffolds in these *Amphimedon* and *Trichoplax* suggests that their karyotypes more resemble those of sponges, cnidarians, and bilaterians rather than the karyotypes of ctenophores or unicellular outgroup species.

Our analyses contain evidence that placozoan chromosomes have derived fusions and fissions after the 5 FLG fissions in the ancestor to sponges, cnidarians, and ancestors. While it is impossible to absolutely determine the absolute phylogenetic position of placozoans without a complete chromosome-scale genome of at least one species, it is clear that they occur in a monophyletic clade containing sponges, cnidarians, and bilaterians, to the exclusion of ctenophores.

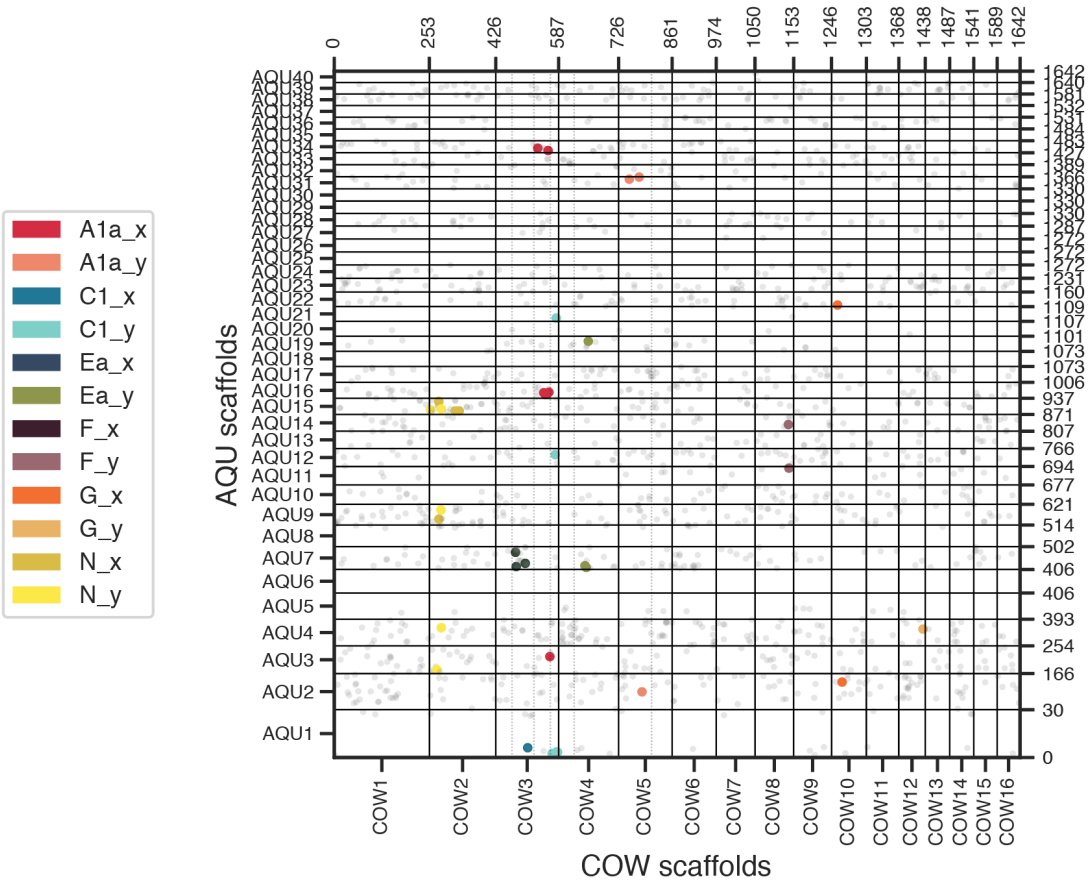

Figure 3.6 - Oxford dot plots of *Capsaspora* and *Amphimedon queenslandica*.

Best protein hits between *Capsaspora* and *Amphimedon* identified by an HMM. Genes corresponding to FLGs are colored. Only the largest 40 *Amphimedon* scaffolds are plotted. Scaffold AQU1 contained genes from both C1_x and C1_y. Scaffold AQU7 contained genes from both Ea_x and Ea_y.

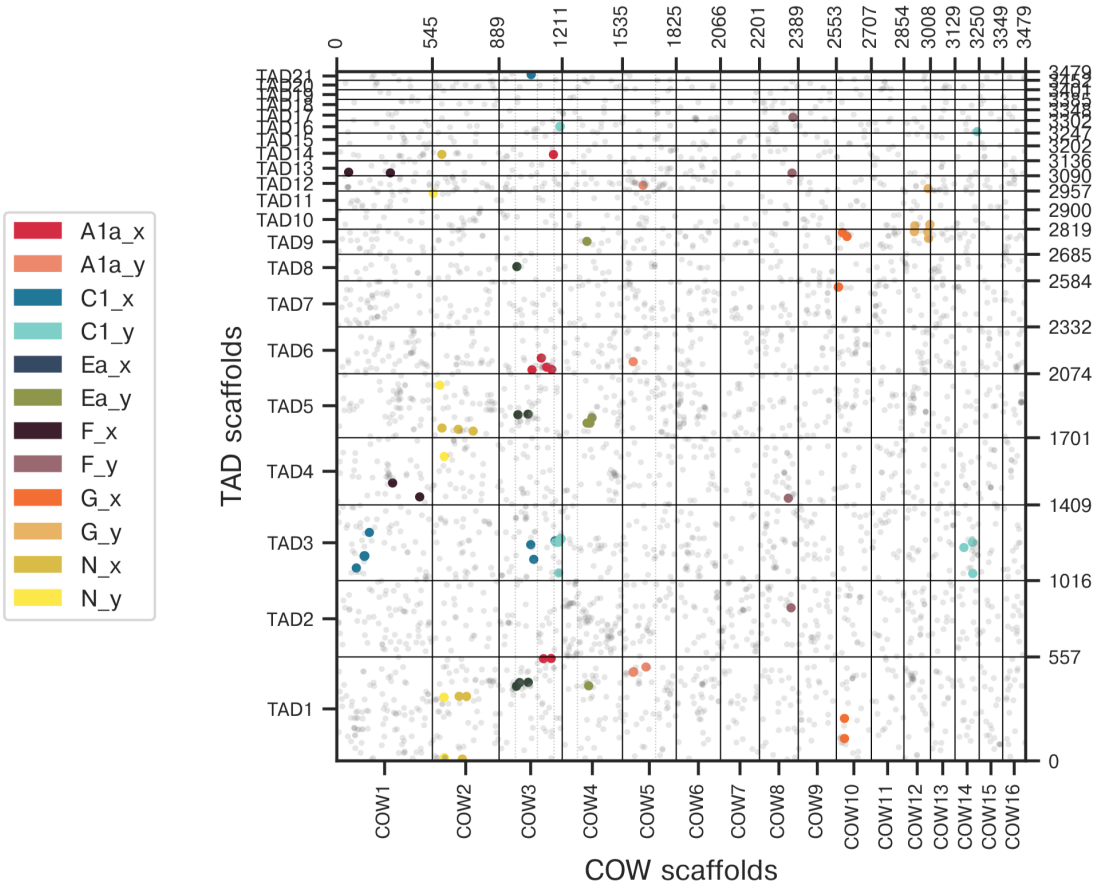

Figure 3.7 - Oxford dot plots of *Capsaspora* and *Trichoplax adhaerens*. Best protein hits between *Capsaspora* and *Trichoplax* identified by an HMM. Genes corresponding to FLGs are colored. Only the largest 21 *Trichoplax* scaffolds are plotted. Scaffold TAD3 contains genes from C1_x and C1_y, mixed, similar to the sponge, cnidarians, and bilaterians. Also, genes from Ea_x and Ea_y are both on TAD3, genes from Ea_x and Ea_y are both on TAD5, genes from A1a_x and A1a_y are both on TAD 6, and genes from G_x and G_y are both on TAD9.

Ctenophores are the sister phylum to other animals

Our analyses have shown eight pieces of evidence to help determine whether sponges or ctenophores are the sister clade to other animals. First, we found groupings of genes that occur on the same sets of chromosomes in *Capsaspora*, *Rhopilema*, *Hormiphora*, and *Ephydatia* using four-way reciprocal best blastp searches. The false discovery rate of finding that quantity of genes on that pairing of chromosomes, calculated with a permutation test, was significantly low for every grouping ($\alpha << 1e^{-4}$). Second, of these significantly large groups of genes, there were 10 groups of genes that were present on separate chromosomes, or non-overlapping chromosome intervals, in *Hormiphora* and *Capsaspora* (Figure 3.3). However, all of those ten groups of genes occurred on a single chromosome in *Rhopilema* and *Ephydatia* (Figure 3.3). Third, of the significantly large groups of genes from the four-way reciprocal best blastp search, none of the gene sets occurred on the same pairs of chromosomes in *Hormiphora* and *Rhopilema*, while on different sets of chromosomes in *Ephydatia* and *Capsaspora* (Figure 3.3). Fourth, the ten groups of genes that we call FLGs, or Filozoan linkage groups, are mixed on individual chromosomes in the sponge *Ephydatia*, and even more mixed and dispersed on homologous chromosomes in *Rhopilema* (Table 3.2). Fifth, the linkage groups that we identified are also present in the genomes, to some extent, of the unicellular species *Salpingoeca* and *Creolimax* (Figure 3.3). The groups that were present in these species were all on separate chromosomes. Sixth, The ten linkage groups were also all identified in the genome of another ctenophore species, *Bolinopsis infundibulum*,

with the same FLG configuration present in *Hormiphora* (Figure 3.5). As in the *Hormiphora* genome, there was no mixing between genes on the _x and _y pairs of FLGs in the *Bolinopsis* genome, with the exception of a single gene (Figure 3.5). Seventh, similar to the FLG configuration in the sponge *Ephydatia* and the cnidarian *Rhopilema*, we identified all ten FLGs as five mixed ALGs on single chromosomes in the genomes of the anemone *Nematostella* and in the chordate *Branchiostoma* (Figure 3.3). Eighth, as in the sponge, cnidarian, and bilaterian genomes, the fragmented genomes of the placozoan *Trichoplax* and the sponge *Amphimedon* contained scaffolds with genes from both _x and _y partner FLGs (Figure 3.6, Figure 3.7). For example, genes from both C1_x and C1_y were found on the same scaffold.

We must note that the mixing of two groups of genes on a chromosome is a probabilistically irreversible event. Imagine placing one deck of 500 blue cards on top of a deck of 500 red cards. Once the deck has been shuffled several times (chromosomal inversions), the probability that another shuffle will revert the deck back to a state with all red cards and blue cards separate is very unlikely. In this way, gene order mixing is an irreversible character that is phylogenetically informative, and gives context about where, in a phylogenetic tree, different fusions occurred.

The findings presented above leave two possible explanations for resolving the phylogenetic position of sponges and ctenophores. The first possibility is that sponges are the sister phylum to the rest of animals. In this scenario, the FLGs that we identified fused into common shared chromosomes in the ancestor to all animals. We see these fused-and-mixed FLGs in chromosomes of extant cnidarians, bilaterians, in

the sponge *Ephydatia*, and in the placozoan *Trichoplax*. In this scenario, the ancestral ctenophore would have had to have undergo a series of inversions to un-mix the FLGs on individual chromosomes, then undergo seven chromosomal fissions to split the genes into the same subgroups as found in the outgroup to all animals, *Capsaspora*. The probability of an entire phylum having a complete reversal to an ancestral karyotype state through seven fusions, then seven identical fissions, is a probability too small beyond reasonable consideration.

The alternative explanation for the shared sub-ALG linkages on cnidarian, bilaterian, placozoan, and sponge chromosomes is that the phylum Ctenophora is the sister phylum to all other animals. In this scenario, the 16 FLGs that we identified were present on separate chromosomes in the genomes of the ancestor to the Filozoa, the ancestor to the Choanozoa, and the ancestor to the Metazoa. These FLGs persist on separate chromosomes today in filasterians, choanoflagellates, and ctenophores. However, FLGs underwent chromosome fusion-then-mixing events in the ancestor to sponges, cnidarians, and bilaterians.

Given that the sponge-sister hypothesis can be rejected due to the improbability of the reversal of ctenophore genomes to the ancestral karyotype state, we conclude that the most parsimonious explanation is that ctenophores are the phylum sister to the rest of the Metazoa.

Evolutionary significance of ctenophore-sister

While these results may seem to simply resolve an outstanding open phylogenetic question, the implications of this result impact many fields of biology. Understanding the evolution of the neuron in animals is important to understanding the evolution of all animals, and may help us better understand neuron physiology. The fact that ctenophores have neurons, but sponges and placozoans do not, has been one of the phenetic characters used to argue that that sponges are clade sister to all other animals. However, the biology of ctenophore neurons is different from other animals, at least at the protein level. For instance, ctenophores use many glutamate receptors (Moroz *et al.* 2014; Moroz 2015) where bilaterians use fewer; ctenophore neurons appear to lack polarity in the same sense that bilaterians neurons have (Hernandez-Nicaise 1973). However, many of the genes essential for synaptic function in bilaterians are also present in ctenophores (Ryan *et al.* 2013). Resolving that ctenophores are the sister phylum to other animals leaves open another critical question: did ctenophores evolve neurons independently, or were neurons lost in the sponges and in the placozoans (Ryan and Chiodin 2015)?

Biological significance of genes in FLGs

For every _x and _y component of the FLGs A1a, C1, Ea, F, G, and N, we found the closest human gene using a hidden Markov model. Using these genes, we identified human orthologs, and performed a GO enrichment analysis to look for a biological signal grouping any of the FLGs together. We did not find significant

enrichment of any GO term in any of the gene groupings of an FLG. However, we note that ctenophores, sponges, and cnidarians are not model organisms, and the GO terms of the human orthologous proteins may not reflect the biology of the proteins' functions in non-bilaterian animals.

One alternative explanation for why these genes have persisted on the same chromosomes is dosage-dependency. If genes require expression at similar doses to maintain fitness, then there is negative selection when a chromosomal fusion splits up groups of genes (Lv *et al.* 2011).

Conclusion

Comparing the chromosome-scale genome assemblies of non-bilaterian animals to unicellular outgroups has provided several lines of evidence that ctenophores are the sister phylum to other animals. Phylogenomic methods for resolving whether sponges or ctenophores are the sister clade to other organisms have not been able to resolve the question without uncertainty, and have inherent issues with model choice and long-branch artifacts. In this study, the discovery of fourteen groups of genes that are present on different chromosomes in ctenophores an unicellular outgroups, but coexist and are interlaced on the same chromosomes in all other animals, unambiguously polarizes the early events in animal evolution. Other plausible alternatives to this finding, such as the possibility that sponges or placozoans are the sister clade to other animals, had no supporting evidence in the genomes of any species we studied. These results were consistent when looking at multiple genomes of unicellular outgroups, multiple ctenophore genomes, multiple

sponge genomes, multiple cnidarian genomes, and the only contiguous placozoan genome assembly available.

Further genome sequencing of species that are unicellular outgroups to animals may reveal other ancestral linkages of genes. While it is not yet clear if these ancestrally linked genes share a common biological function, these relationships represent gene linkages that date to over one billion years ago (Dohrmann and Wörheide 2017). We note that it is possible that the phylogenetic relationship of ctenophores and sponges may have never been answered with only a few more chromosomal rearrangements in the genome of *Capsaspora*.

Chapter 4

Bioluminescence biochemistry: a novel luciferase from the syllid polychaete, *Odontosyllis undecimdonta*, and determining the luminescence system of an undescribed cladorhizid sponge.

This text is adapted from two published articles:

Séverine Martini[†], Darrin T. Schultz[†], Lonny Lundsten, Steven H.D. Haddock. ***Bioluminescence in an Undescribed Species of Carnivorous Sponge (Cladorhizidae) From the Deep Sea.*** Frontiers in Marine Science 7 (2020): 1041.

Darrin T. Schultz[†], Alexey A. Kotlobay[†], Rustam Ziganshin, Artyom Bannikov, Nadezhda M. Markina, Tatiana V. Chepurnyh et al. ***Luciferase of the Japanese syllid polychaete Odontosyllis undecimdonta.*** Biochemical and Biophysical Research Communications 502, no. 3 (2018): 318-323.

[†] - Indicates co-first authorship

Introduction - *Odontosyllis* bioluminescence

One class of marine organisms that has many bioluminescent species is the Polychaeta. Polychaete worms occupy many marine habitats including the intertidal, reefs and inshore ecosystems, the midwater, and the seafloor. Six families within the polychaetes contain bioluminescent species: the Chaetopteridae, the Polynoidae, the Syllidae, the Tomopteridae, the Cirratuliformia, and the Terebelliformia (Verdes and Gruber 2017). However, the protein and small molecule components causing bioluminescence are not known in any of the species in these families.

Of the luminous polychaetes, species in the genus *Odontosyllis* are perhaps the most widely distributed around the world. There are luminous *Odontosyllis* species reported in Toyama Bay in Japan (掘井直二郎 1982), Bermuda (Huntsman 1948), San Diego (Tsuji and Hill 1983), and Belize (Gaston and Hall 2000). These animals dwell in tubes in and on substrates on the seafloor, and bioluminesce en masse only several times per year during courtship displays, which have been characterized in several studies (Wilkens and Wolken 1981; Tsuji and Hill 1983; Fischer and Fischer 1995).

Odontosyllis courtship displays are both mesmerizing and have provided excellent opportunities to collect biomaterial for studying the bioluminescence biochemistry of these species (Shimomura *et al.* 1963, 1964; Trainor 1979; Deheyn and Latz 2009). However, *Odontosyllis* secretes its luminescence in a mucus, from which it is difficult to purify proteins. Also, the luciferin molecule was difficult to collect in sufficient quantity, and proved to be unstable in the presence of oxygen.

These studies did not resolve the protein sequence or luciferin structure of the *Odontosyllis* luminescence system.

Odontosyllis - Materials and Methods

Specimen collection

Odontosyllis undecimdonta worms used in this study for protein purification, MS transcript identification, and nucleic acid purification were collected on October 06, 2016 in Toyama Prefecture Japan, Namerikawa City, at the coordinates 36° 46′ 40.3032” N 137° 20′ 42.378” E. At dusk, *Odontosyllis* worms were attracted to a handheld light at the surface and collected with a hand dip net. Worms were individually preserved in Invitrogen RNAlater or lyophilized for later analysis.

RNA Isolation and sequencing

Total RNA intended for Illumina RNA-seq and Oxford Nanopore cDNA libraries was isolated using the Trizol protocol (Rio *et al.* 2010). A template-switching Illumina RNA-seq library from OdonB total RNA was prepared at Evrogen (Moscow, Russia) with a TruSeq Stranded mRNA Library Prep Kit v2 with the i7 index ACAGTG(A). The library was sequenced at the UC Davis DNA Technologies Core on an Illumina HiSeq 4000 2x150 PE run to a depth of 32,457,166 read pairs.

For cDNA sequencing on the Oxford Nanopore Technologies Minion, we first synthesized cDNA using an established synthesis protocol (Picelli *et al.* 2014). We

captured mRNAs from 50µL of *O. undecimdonta* RNA using the oligo (5'-/5Me-isodC/AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTT TTTTTTTT TTTTTVN-3'). performed strand-switching with the oligo (5'-AAGCAGTGGTATCAACGCAGAGTACATrGrGrG-3'), then amplified the cDNA using the ISPCR primer (5′-AAGCAGTGGTATCAACGCAGAGT-3′) for fifteen cycles.

From the amplified cDNA 1 µg was used as input for the SQK-LSK208 Oxford Nanopore Technologies 2D Strand switching cDNA sequencing protocol. The final library concentration after 2D adapter-ligated capture and prior to sequencing was 8.18 ng/µl. The final library mass loaded to the flow cell was 98.16 ng in 12 µl of library. The flow cell used was a model FLO-MIN106. We used MinKNOW v1.3.30 to control the sequencing run. The sequencing run produced 428,172 fast5 read files. We used Albacore v1.1.1 to perform 2D basecalling on the reads and poretools v0.6.0 to extract reads from the basecalled fast5 files (Loman and Quinlan 2014).

Transcriptome Assembly

Adapters were trimmed from the Illumina RNA-seq reads using SeqPrep2 . We then assembled the transcriptome using Trinity v2.1.1 with the option --SS_lib_type FR for read directionality and the --long_reads option using all 2D reads extracted from the Albacore-basecalled Oxford Nanopore reads (Grabherr *et al.* 2011).

Luciferase identification and expression

The *O. undecimdonta* luciferase was purified using chromatographic techniques, and the luminous fractions from were analyzed on a mass spectrometer. We used the mass spectrometry data to identify putative luciferase sequences from the transcriptome based on their high abundance in the protein purification fraction that had the highest bioluminescence activity.

We mapped the long Oxford Nanopore cDNA reads, and Illumina RNA-seq reads, to the putative transcripts to determine whether the sequence we identified from the transcriptome was supported by the presence of RNA-seq read data mapping to the area.

Protein expression and characterization of the luciferase

After assaying whether the putative luciferase transcripts appeared to be real isoforms based on support with long-read cDNA data, the putative luciferase was expressed in a mammalian cell culture. The cells expressing the putative luciferase were lysed, and were found to be luminous upon mixing with an aqueous luciferin solution. The emission spectrum of the in vitro luminescence reaction matched the emission spectrum of the live worms.

Homology search in other polychaetes

We sought to determine whether the luciferase from *O. undecimdonta* had orthologs in other polychaete species, luminous or not. Answering this question can

help determine if bioluminescence evolved independently in this lineage, or if the evolution was shared with another clade.

We found RNA-seq data of many polychaete species on the NCBI Sequence Read Archive (SRA), but did not find transcriptomes assembled from these reads on any publicly available repositories. So, we used Trinity (Grabherr *et al.* 2011) to assemble transcriptomes from publicly available RNA-seq data of the following species: Amphinomidae (*Pareurythoe californica* SRR1926090 (Andrade *et al.* 2015)), Chaetopteridae (*Chaetopterus* sp. SRR1646443 (Lemer *et al.* 2015), *Chaetopterus variopedatus* SRR5590967, *Mesochaetopterus minutus* SRR1925760 (Andrade *et al.* 2015), *Phyllochaetopterus* sp. SRR1257898 (Weigert *et al.* 2014), *Spiochaetopterus* sp. SRR1224605 (Weigert *et al.* 2014)), Eunicida (*Eunice pennata* SRR2040479, *Eunice torquata* SRR2005375 (Andrade *et al.* 2015)), Cirratulidae (*Cirratulus cirratus* SRR5590966, *Cirratulus spectabilis* SRR3574861 (Li *et al.* 2017)), Flabelligeridae (*Flabelligera mundata* SRR3574613 (Li *et al.* 2017)), Acrocirridae (*Macrochaeta clavicornis* SRR1221445 (Weigert *et al.* 2014)), Phyllodocidae (*Phyllodoce medipapillata* SRR2016923 (Andrade *et al.* 2015)), Polynoidae (*Harmothoe extenuata* SRR1237766 (Weigert *et al.* 2014), *Harmothoe imbricata* SRR2005364 (Andrade *et al.* 2015) and SRR4841788 (Francis *et al.* 2013)), Syllidae (*Syllis* sp. SRR1224604 (Weigert *et al.* 2014)), and Tomopteridae (*Tomopteris helgolandica* SRR1237767 (Weigert *et al.* 2014)). The confirmed luminous species included in this analysis were *Chaetopterus variopedatus* (Shimomura 2006), *Harmothoe extenuata* (Bassot and Nicolas 1978), *Harmothoe*

imbricata (Miron *et al.* 1987), and *Tomopteris helgolandica* (Gouveneaux and Mallefet 2013). All other species mentioned above may be luminous, with the exception of: *Eunice* spp., *Pareurythoe californica*, and *Phyllodoce medipapillata* (Herring 1987).

To search for homologs of the putative *O. undecimdonta* luciferase, we queried the *O. undecimdonta* luciferase against individual translated polychaete transcriptomes using blastp v2.7.1. For each blastp search, we limited the search to the top hit using -max_target_seqs 1. The top hit for the luciferase in each species' genome was queried against the nr database to determine the possible protein identity against annotated proteomes.

Odontosyllis - Results

The isolation and purification of *O. undecimdonta* luciferase required ion exchange chromatography, size exclusion chromatography, and ultrafiltration. The presence of luciferase in samples was controlled by a bioluminescence assay for all stages of purification. Several bands that corresponded to bioluminescence activity in the size exclusion chromatography fractions were identified by polyacrylamide gel electrophoresis. These bands were excised from gel and were identified by liquid chromatography mass spectrometry.

The transcriptome assembled with Illumina paired-end reads and ONT 2D reads extracted with poretools "fwd" parameter yielded 256,027 transcripts and a

median transcript length of 737 base pairs. Four transcripts were identified as potential luciferases based on coverage and quantity of MS matches. Three long transcripts c9g1i2 (990 bp), c9g1i3 (993 bp), c9g1i6 (990 bp) had c-terminal amino acid variation. Transcript c9g1i5 (711 bp) was homologous to the aforementioned three transcripts but lacked 118 n-terminal amino acids. These four transcripts were verified by presence of two ONT whole-cDNA reads that spanned from the 5′ UTR to the 3′UTR (Figure 4.1). Non-spliced mapping of an Illumina paired-end polyA RNA-seq library also confirmed that the longest of the four transcripts were present in the cDNA library of the worm. The protein products of these four transcripts were identical at 92% of sites (Figure 4.2).

Figure 4.1 - Supporting evidence for putative luciferase sequences. The Peptide Matches track shows unique peptide hits to any of the four transcript models that match by DNA and amino acid sequence similarity. All transcript models except c9g1i5 share the same structure, whereas c9g1i5 lacks 93 N-terminal amino acids. The ONT cDNA Reads track shows individual Oxford Nanopore 2D cDNA reads that align to the c9g1i2 transcript. Three reads span the complete 5′ UTR, transcript, and 3′ UTR of the long isoforms (c9g1i2, c9g1i3, c9g1i6), and four additional reads support the 5′ UTR of the long isoforms. The RNA-seq coverage track supports the 5′ and 3′ UTR of the long isoforms, despite the predictable 3′ bias inherent to polyA-selecting library preparation techniques.

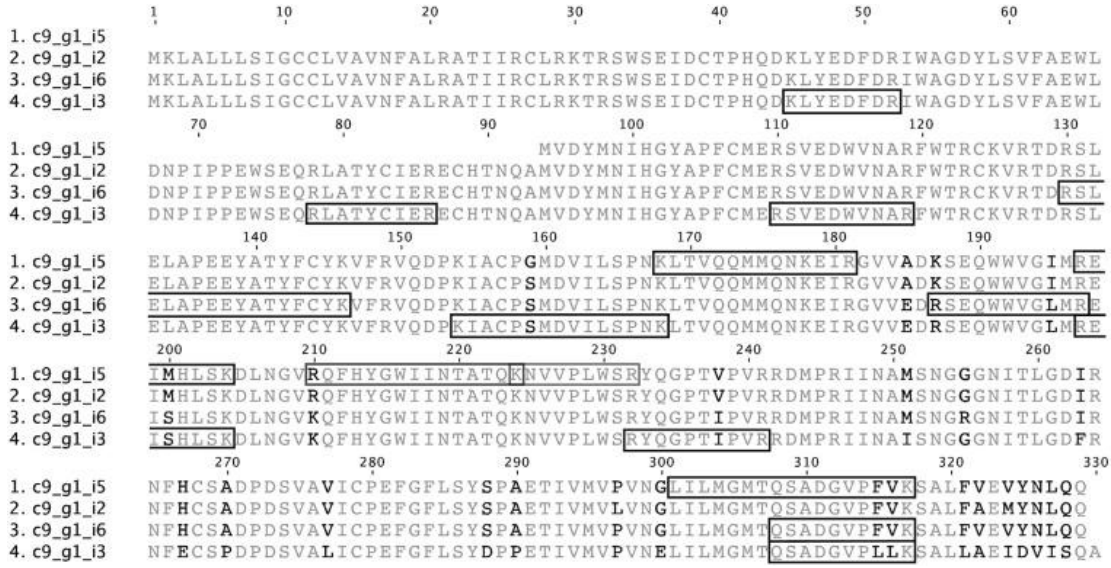

Figure 4.2 - The amino acid alignment of the four putative luciferase transcripts. Black boxes surrounding the alignment indicate regions to which there were exact MS peptide matches. The four transcripts are, on average, 92% identical to one another. Transcript c9g1i5 lacks 93 N-terminal amino acids. All transcripts have a highly variable C-terminus. Alignment columns highlighted black are variable sites.

All four candidate DNA sequences were synthesized as linear dsDNA fragments and cloned using MoClo technology. Then, mammalian cells were transfected by resulting constructs. Mammalian cell culture lysate from two of the above four candidates produced bioluminescence when assayed with purified luciferin (c9g1i2 and c9g1i6). The bioluminescence spectra of positive clones were similar to that of native *O. undecimdonta* worms. However, cell culture lysate from expressed transcripts c9g1i3 and c9g1i5 were not luminous. None of the non-lysed cell cultures produced luminescence when purified luciferin was applied.

The protein product of c9g1i2 is 329 amino acids. A tblastn search with the c9g1i2 protein product only found an insignificant match (E-value = 3.1) to a predicted transcription factor (sequence XM_021634012.1) from gerbil (*Meriones unguiculatus*). A blastn search returned no significant matches. Blast searches against the assembled transcriptomes of publicly available polychaete RNA-seq read data also yielded no significant matches in any of the species.

Odontosyllis - Discussion

Given our lack of fresh specimens we opted to extract and purify the *Odontosyllis* luciferase directly from the lyophilized worms and successfully identified the luciferase gene using classic protein purification, luciferin purification, and recent whole-cDNA sequencing techniques. We then reconstructed native *Odontosyllis* bioluminescence *in vitro* using purified protein and highly purified luciferin with no additional cofactors. Lastly, we verified the identity of the *Odontosyllis* luciferase gene by showing that recombinant protein and purified luciferin in cell-lysate is luminous, in which the luminescence spectra (λ_{max}, near 510 nm) matches that of the *Odontosyllis* in vivo luminescence.

While the bioluminescence emitted during mating is well-characterized in *Odontosyllis* spp., the luciferase structure and the mechanism of the luciferin-luciferase reaction remains unclear. Despite this uncertainty, protein ortology searches using BLAST show that syllid luciferase is unique both among sequenced polychaetes and other sequenced organisms in public databases. The lack

of evidence for a conserved protein in the transcriptomes of other luminous polychaetes leaves open the theory that bioluminescence evolved more than three times in the annelids. In this conservative estimate, we only include the evolution of two unique bioluminescent systems for which either the structure of the luciferin, luciferase, or both have been determined (earthworms (Petushkov *et al.* 2014) and *Odontosyllis*) plus at least one event for other annelids with uncharacterized bioluminescent systems. Given that the structure of other polychaete luciferins is still unknown, this leaves the question of polychaete bioluminescence unanswered. Identification of the *O. umdecimdonta* luciferase sequence is the most important step to further characterization of this worm's bioluminescent system and the screening of other purified polychaete luciferins for cross-reactivity.

Introduction - Bioluminescence in sponges

Another group of marine organisms that has poorly understood or dubious accounts of bioluminescence is the sponges (phylum Porifera). The first record of autogenic sponge luminescence is a 19th-century mention of light-emitting sponge embryos (Pagenstecher 1881). Other observations of luminous sponges over the next several decades were ultimately attributed to worms or other invertebrates living in the pores of the sponge tissue (Dahlgren 1916; Okada 1925). One record from the 20th century describes light being emitted from a sponge itself (Harvey 1921), but this has also been considered doubtful by later authors (Herring 1987). A recent publication claims the discovery of a luminous sponge (Demospongiae, *Suberites domuncula*) and a luciferase, however the authors argue that this marine sponge uses a firefly luciferase homolog and firefly luciferin as its luminescence system (Wiens *et al.* 2010; Wang *et al.* 2012). Because sponges filter large volumes of water, it is challenging to distinguish between luminescence of the animal itself or light produced by other organisms concentrated within its tissues. Bioluminescence observations in sponges may have been induced by numerous bioluminescent symbiotic, entrained, or captured bacteria living in sponges (Hentschel *et al.* 2006), or other associated eukaryotes. Given the unclear records of luminescence in sponges, whether or not autogenically luminous sponges exist remains a mystery.

In these studies, we determined the protein sequence of the luciferase of *Odontosyllis undecimdonta*, discovered luminescence in a sponge, and determined some components of the luminescence system in the sponge. My contributions to the

polychaete luciferase project were sequencing the transcriptome of *Odontosyllis undecimdonta*, helping to identify the putative luciferase sequences using mass spectrometry data, corroborating the transcripts with long-read sequencing data, and performing homology searches for the *Odontosyllis* luciferase in other polychaetes. The co-first author of this study, Dr. Alexey Kotlobay, purified the luciferase, performed mass spectrometry, characterized the biochemical properties of the luminescence system, and expressed the luciferase.

My contributions to the sponge luminescence project were performing the molecular identification of the sponge's taxonomic placement, photographing and recording video of sponge bioluminescence, performing biochemical experiments on the bioluminescence system of the sponge, and performing a metagenomics analysis on the sponge tissue to consider alternative hypotheses of the provenance of the sponge's luminescence. The first co-first author on the manuscript, Dr. Séverine Martini, first discovered that the sponge was luminous, and also photographed and recorded videos of the animals.

Sponge - Materials and Methods

Sample Collection

In three consecutive years, between June, 2017 and July, 2019, MBARI's ROV Doc Ricketts, was used to collect six specimens of an undescribed poriferan (hereafter named individuals Clado1-Clado6), in the deep Northeast Pacific Ocean (Figure 4.3) approximately two hundred kilometers offshore from Big Sur, California (Table 4.1). All specimens were sampled on the seafloor composed of silt and clay.

Other macro-organisms were observed in the surrounding area, including the Holothuroidea (*Scotoplanes* sp.) and benthic ctenophores. These sponges were anchored to the substrate with rhizoids. The sponges were collected using the ROV's robotic arm to clasp the stem and were dropped into a sampling container dedicated to biological collections. Several sponge specimens were collected with one or more benthic ctenophores attached to the rigid stalk (Figure 4.3A). Once the specimens were retrieved from the ROV, they were kept in a 4°C, dark cold room in seawater.

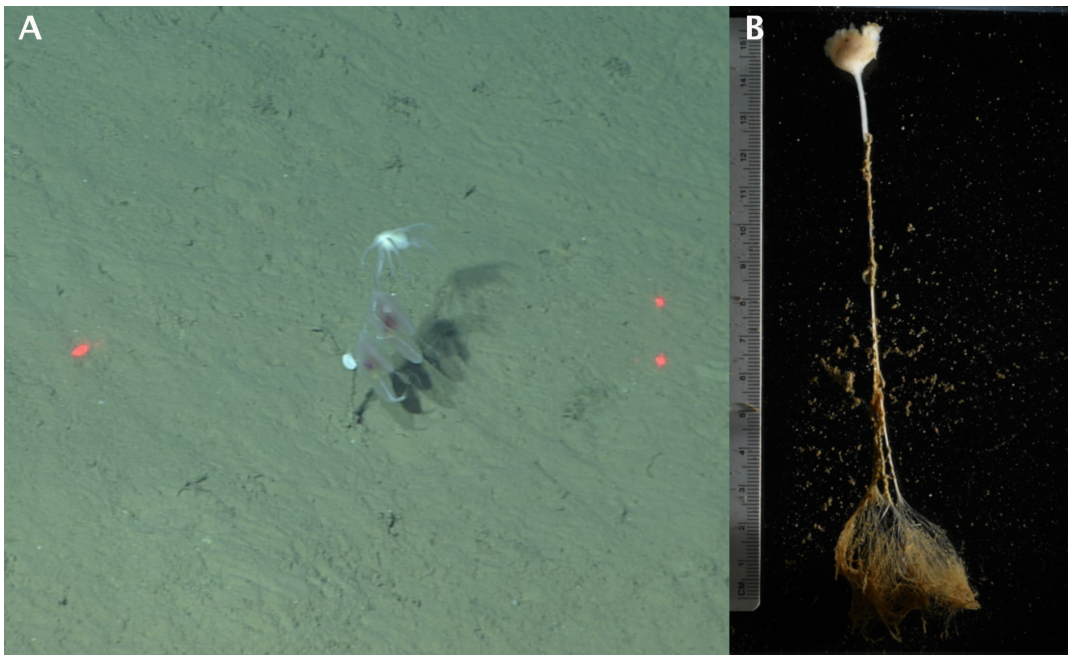

Figure 4.3 - Sampling of the Cladorhizidae sponge and morphological observation. (A) *In situ* observation of Clado6 by ROV, with two benthic ctenophores attached on the stalk. (B) In the lab observation of Clado6. Filaments are retracted.

Bioluminescence Observations

Remaining sediments were cleaned from the samples (Figure 4.3B), then the samples were transferred to fresh 4°C seawater and were left to rest for 10 minutes

before being gently stimulated to determine whether they emitted light. To document light emission, images were taken using a Sony α7S camera.

In Vitro Bioluminescence Assays

After we mechanically stimulated bioluminescence in the sponges, we cut approximately 200 mg portions of sponge individuals Clado3 and Clado5 to use for biochemical tests. Each sample was ground in a loose-fitting 1 mL Dounce homogenizer on ice in approximately 750µL of 750 mM NaCl 20 mM Tris-HCL buffer equilibrated to pH 8.0 at 3°C. The homogenate was centrifuged at 14,000 rcf at 4°C for ten minutes on the ship. The supernatant was removed and placed in a new tube. The pellet was resuspended in 500µL of the Tris-HCl, NaCl buffer described above. Bioluminescence measurements were taken in a custom-built integrating sphere that enables using a micropipettor to inject samples into the measurement chamber while recording. Measurements were made at a sampling frequency of 40ms (Haddock *et al.* 2001). The measurement protocol was as follows: (1) place 1-10 µL of buffer, coelenterazine, or *Renilla* luciferase solution in a clear 1.5 mL microcentrifuge tube inside the integrating sphere, (2) place a micropipettor with 90-190 µL of homogenate or other sample inside the tube in the integrating sphere, but not touching the 1-10 µL of liquid, (3) close the door to the integrating sphere and record the baseline luminescence of the unmixed analytes for four seconds, (4) inject the contents of the micropipettor into the tube to mix, without removing the micropipettor, (5) allow the assay to continue to completion at 20s total.

Using Clado3, we conducted the following assays. (A) To test if coelenterazine was the luciferin causing light emission in the sponge sample we added 90 µL of homogenate to 10 µL of 1 µg/mL coelenterazine in Tris/NaCl buffer. We also added 90 µL of homogenate to 10µL of 0.1 µg/mL coelenterazine. As a negative control we used 10 µL of Tris/NaCl buffer and 90 µL of homogenate. (B) To determine if the sponge homogenate contained coelenterazine we added 98 µL of homogenate to 2µL of 1.4 µg/mL of *Renilla* luciferase (RLuc) in the Tris/NaCl buffer described above. We conducted this test twice. As a negative control we used 2 µL of Tris/NaCl buffer and 90 µL of homogenate. (C) We assayed the heat stability of the luciferase by measuring the baseline activity of 60 µL of homogenate and 10 µL of 1 µg/mL coelenterazine in Tris/NaCl buffer. Unused homogenate was incubated in a heat block at 96°C for three minutes, then placed on ice. 60 µL of heat-treated homogenate was mixed with 10 µL of 1 µg/mL coelenterazine in Tris/NaCl buffer. (D) Lastly, we mixed 90 µL of Tris/NaCl buffer with 10 µL of Tris/NaCl buffer as an absolute control, and conducted the same assay with a homogenate preparation from a nonluminous sponge from the genus *Caulophacus.*

Using Clado5, we prepared homogenate as described above, then performed additional sample clarification to ensure that the observed luminescence was from soluble proteins and not bacteria. The homogenate described above was passed through a 0.45µm filter spin column (Millipore Sigma Ultrafree-MC) at 12,000 rcf for ten minutes at 4°C, then that filtrate was passed through a 0.1µm filter spin column under the same conditions. This material was used to conduct the following

tests. 90 µL of homogenate was mixed with 10 µL of the following compounds to assay for cofactors triggering luminescence: (A) 3% H_2O_2, (B) 200 mM calcium acetate, (C) 1M $CaCl_2$, (D) 2M KCl, and (E) 3M NaCl. We also used the clarified sample to conduct the following assays: (F) 90 µL homogenate and 10 µL of 1 µg/mL coelenterazine in Tris/NaCl buffer, (G) 90 µL homogenate and 10 µL of Tris/NaCl buffer, (H) 90 µL homogenate and 10 µL of 1.4 µg/mL of *Renilla* luciferase in Tris/NaCl buffer.

To determine the size of the protein or protein complex responsible for light emission, we concentrated 150 µL of the 0.45 µm- and 0.1 µm-filtered Clado5 homogenate on a 50 kDa spin column (Millipore Sigma Ultrafree-MC) at 12,000 rcf until the sample had 75 µL in the retentate and 75 µL in the filtrate. We assayed for luminescence by mixing (I) 60 µL of the retentate with 10 µL of 1 µg/mL coelenterazine solution, and (J) 60 µL of the filtrate with 10 µL of 1 µg/mL coelenterazine solution. The Clado5 0.45 µm and 0.1 µm filter-clarified homogenate still emitted light when it was mixed with coelenterazine. The Clado5 clarified homogenate's light-emitting activity with coelenterazine was concentrated on a 50 kDa spin column, and the 50 kDa spin column filtrate had little light-emitting activity when mixed with coelenterazine.

Sequencing and COI Assembly

The Clado1 sponge was rinsed in filtered seawater, then a subsample was frozen in liquid nitrogen. The sample was later pulverized with a blue plastic pestle in tissue lysis buffer and genomic DNA was isolated with an E.Z.N.A. Mollusc DNA kit (Omega Bio-tek). DNA was sheared with a Bioruptor sonicator and a whole-genome shotgun (WGS) sequencing library was prepared (Meyer and Kircher 2010). This library, DS137, was sequenced on a 2x75PE MiSeq run in the UC Santa Cruz Paleogenomics laboratory to a depth of approximately two million read pairs.

Reads were mapped to the mitochondrial genome of *Negombata magnifica (Belinky et al. 2008)* using bwa mem (Li 2013), both SNPs and indels were corrected using pilon v1.22 (Walker *et al.* 2014). The corrected assembly was then used again for another round of mapping and correction. This process was iteratively repeated ten times to generate mitochondrial regions containing the sequence from DS137. The corrected *N. magnifica* COX1 sequence had two regions two which DS137 reads mapped. These two regions were used as seeds for two independent runs of MITObim v1.9.1 (Hahn *et al.* 2013). The sequences from the two MITObim assemblies were aligned into a single contig of approximately 3kbp. DS137 reads were aligned to the contig using bwa mem, then SNPs and indels were corrected with pilon. The assembly was verified by visualizing reads mapped to the contig using IGV (Robinson *et al.* 2011) and by visually inspecting a COX1 protein alignment between DS137 and closely related species. Geneious v11.1.5 was used to identify ORFs in

the contig, and blastx was used to determine the identity of the ORFs (Altschul *et al.* 1997).

Molecular Identification and Phylogeny

To identify the most similar species to the sample, the complete COX1 ORF in the mitochondrial contig was used as a blastn query against all sponge nucleotide sequences in NCBI.

To generate a COX1 phylogeny we downloaded select all *Poecilosclerida* (NCBI:txid27925) sponge sequences from NCBI, largely based on the literature (Hestetun et al., 2016) removed sequences that had not been identified to the species level, and removed identical sequences. The COX1 sequences were aligned using MUSCLE v3.8.425 (Edgar 2004) and the alignment was trimmed to a 575bp region contained in all sequences. A maximum likelihood phylogeny was generated with raxmlHPC-PTHREADS using model GTRGAMMA with the extended majority-rule consensus tree criterion (-# autoMRE) using *Guitarra antarctica* LN870510MK833943 as an outgroup (Stamatakis 2014). The analysis completed after 45000 bootstrap replicates. A Bayesian phylogeny was generated using MrBayes v3.2.6 using a chain length of 630000, four heated chains, a heated chain temperature of 0.2, subsampling frequency of 200, a burn-in of 2500, a random see of 2020145, and a HKY85 model with gamma rate variation (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003).

Metagenomic Analysis

We used bbmerge.sh to merge the DS137 read pairs (Bushnell *et al.* 2017). In order to have one query per pair of unmerged reads, the unmerged reads were concatenated together with 40 Ns, and the R2 read for each pair was in the reverse complement orientation. The 2,020,492 merged reads were queried against the NCBI nt database using blastn (Altschul *et al.* 1997). MEGAN v6.5.4 was used to perform metagenomic binning of the sequences (Huson *et al.* 2007). DS137 reads were mapped against the lux operons of *Vibrio*, *Photobacterium*, *Aliivibrio*, *Photorhabdus*, and *Shewanella* using bwa mem.

Sponge - Results

Bioluminescence observation

Bioluminescence was tested and observed in all of the six specimens sampled (Table 1). After gently touching each individual sponge with a gloved hand or round-pointed forceps, we observed blue-green bioluminescence localized around the point of mechanical depression (Figures 4.4). Composite images of bioluminescence video frame grabs show that the luminescence in the animals occurs in various parts of the animal rather than in a single location (Figures 4.5, 4.6, 4.7). Local light responses were emitted from the globular mass, filamentous processes along the siliceous stalk, but not from the rhizoid structure. The light kinetics were bright and visible to the naked eye for 5-10 seconds. Repetitive stimulations were reproducible, and the light did not appreciably dim over time. We attempted to stimulate bioluminescence in Clado1 by flashing a white light at the organism, but it did not

produce a visible bioluminescent response. Potassium chloride, calcium chloride, and freshwater did not cause the animal to emit light. Only mechanical stimulation caused the sponge to bioluminesce.

Furthermore, the luminescence does not correspond with patches of residual marine snow or detritus that remain on the animal even after cleaning. We found no visible animals living in or on the sponges that could be responsible for the luminescence.

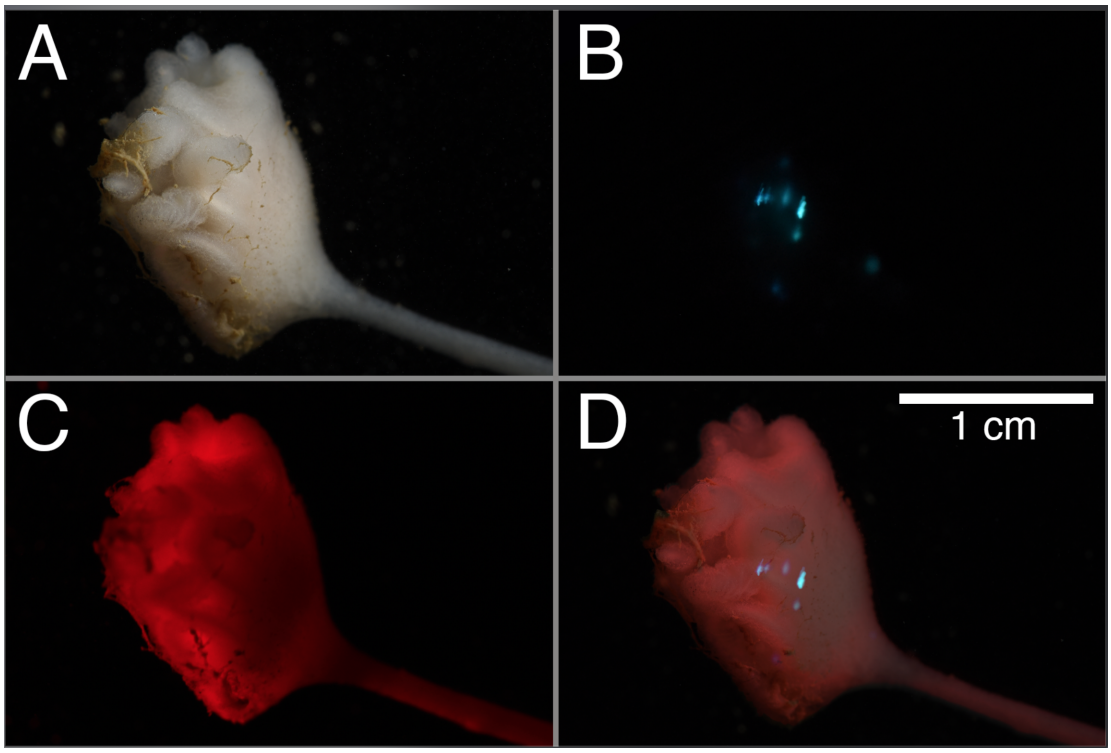

Figure 4.4 - Composite observation of bioluminescence from Clado6. (A) observation of the sponge under white-light, (B) 2.5s exposure of the bioluminescence after mechanical stimulation, (C) red light observation of the individual, and (D) composite of (A-C).

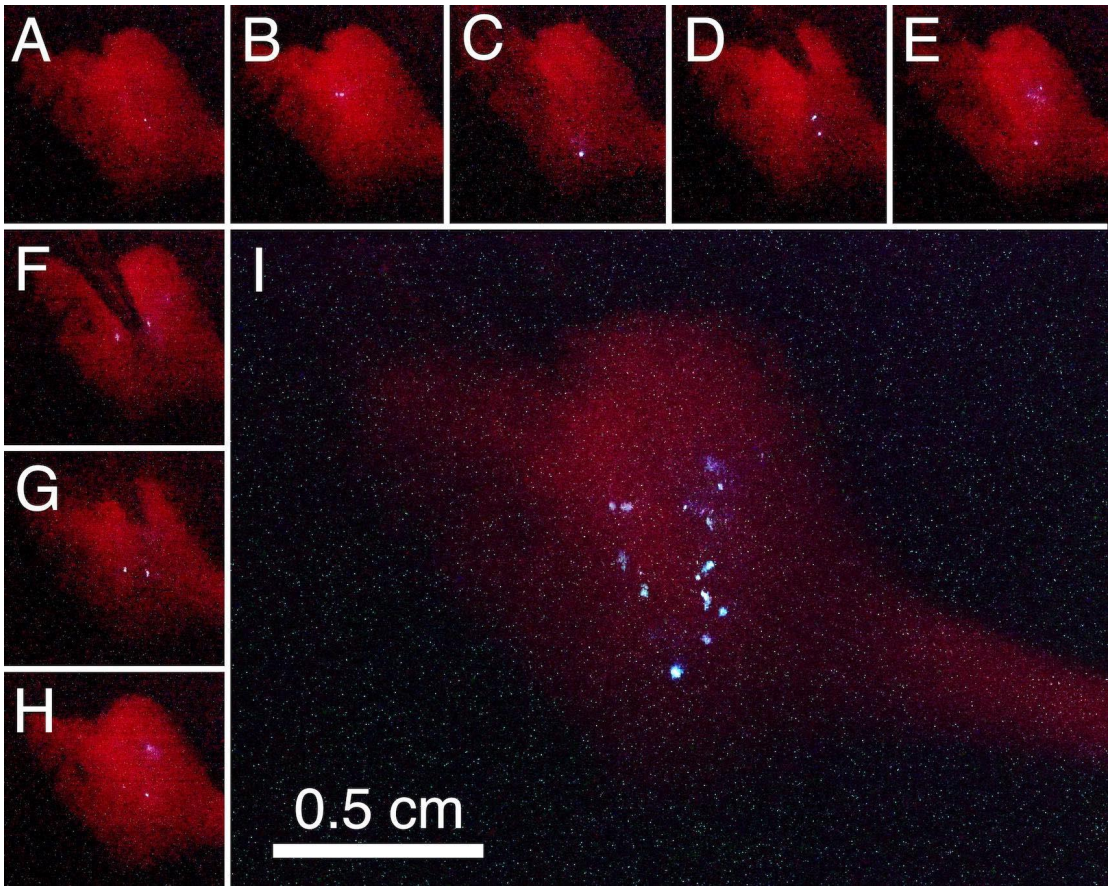

Figure 4.5 - Composite bioluminescence from Clado3. Individual video frames (A-H) of discrete luminescence events from mechanically stimulating Clado3. (I) is a composite of the individual frames.

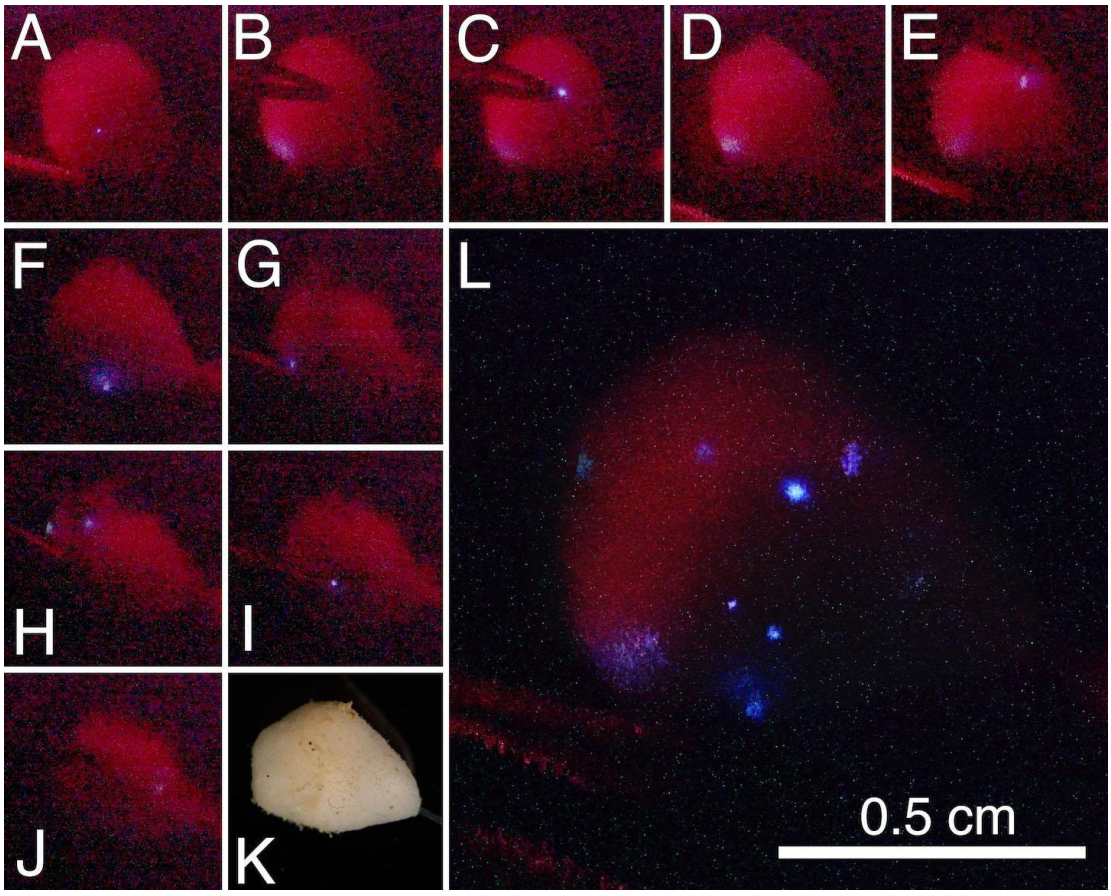

Figure 4.6 - Composite bioluminescence from Clado4. Individual video frames (A-J) of discrete luminescence events from mechanically stimulating Clado4. (K) is a white-light image. (L) is a composite of the individual frames.

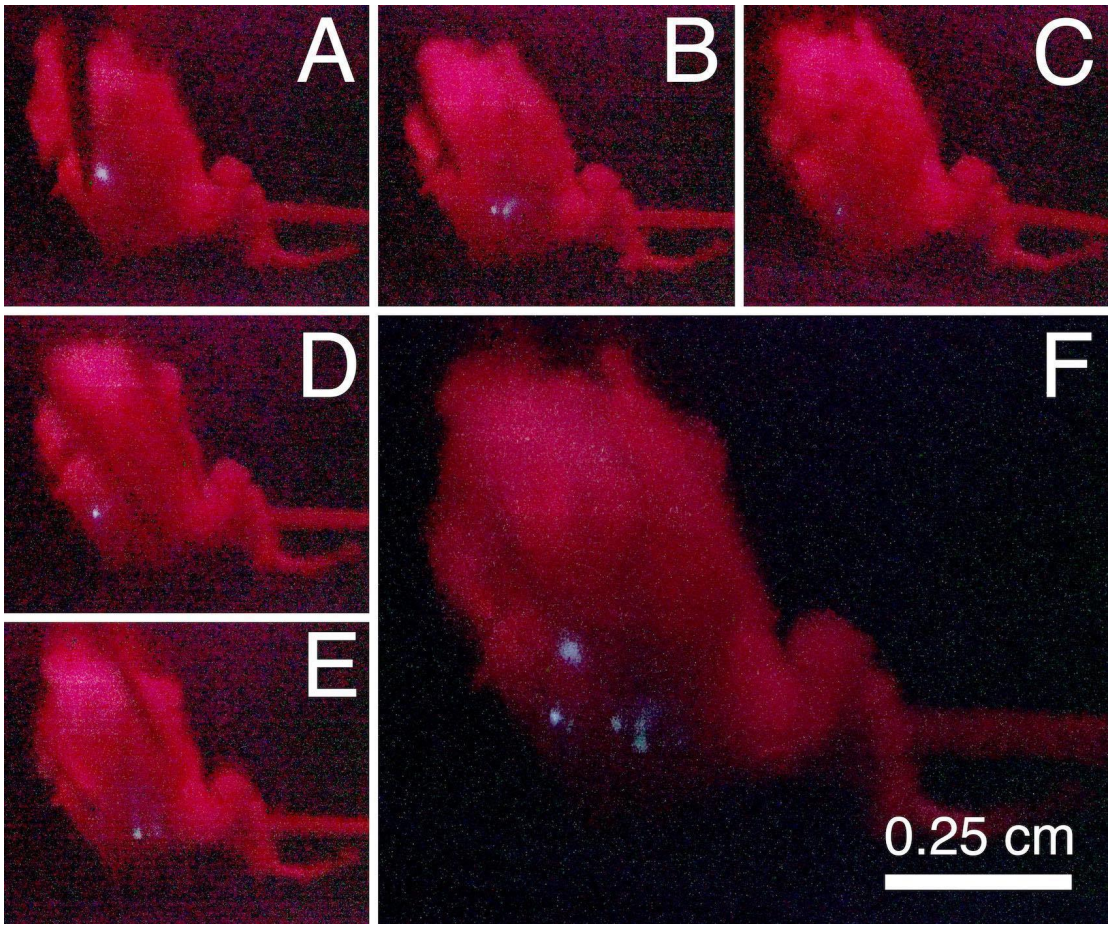

Figure 4.7 - Composite bioluminescence from Clado5. Individual video frames (A-E) of discrete luminescence events from mechanically stimulating Clado5. (F) is a composite of the individual frames.

Sample Name	Collection Latitude	Collection Longitude	Collection Depth	Collection Date	Bioluminescence	Experiments performed
Clado1	35.498°N	123.99°W	3,958 m	June 10, 2017	observed, photographed (Figure S3-A)	morphological, spicules observation (Figure 2), sequencing (Figure 6)
Clado2	35.50°N	124.01°W	3,978 m	May 4, 2018	observed, photographed (Figure S3-B)	morphological observation
Clado3	35.50°N	123.99°W	3,976 m	July 12, 2019	observed, video (Figure S4)	morphological observation, bioluminescence essays (Figure 4)
Clado4	35.50°N	123.99°W	3,977 m	July 12, 2019	observed, video (Figure S5)	morphological observation
Clado5	35.50°N	123.99°W	3,977 m	July 12, 2019	observed, video (Figure S6)	morphological observation, bioluminescence assays (Figure 4 and 5)
Clado6	35.49°N	124.00°W	3,979 m	July 14, 2019	observed, photograph (Figure 3)	morphological observation (Figure 1)

Table 4.1 - Sample collection information for all six sponge specimens. Sequence Read Archive accession for Clado 1 is “PRJNA556048”.

In Vitro Bioluminescence Assays

Biochemical tests were performed to verify the presence of a luciferin and luciferase light-emitting reaction. The homogenate's supernatant from both individuals Clado 3 and 5 luminesced above background without the addition of any cofactors (Figure 4.8). Adding homogenate to a coelenterazine-containing solution caused a light-emitting reaction consistent with luciferase-luciferin type reactions from other species. The intensity of the light-emitting reaction was proportional to the quantity of coelenterazine mixed with the Clado3 homogenate. Adding Clado3 and Clado5 homogenate to *Renilla* luciferase resulted in a light-emitting reaction (Figure 4.8). The light-emitting reaction was not triggered in Clado5 homogenate by any of the following compounds: calcium chloride, calcium acetate, hydrogen peroxide, potassium chloride, sodium chloride (Figure 4.8). In comparison, homogenized tissue from a nonluminous *Caulophacus* species sponge did not produce a light-emitting reaction when added to coelenterazine. Mixing nonluminous sponge homogenate and *Renilla* luciferase also did not produce light (Figure 4.8).

Heating the Clado3 homogenate at 96°C for three minutes reduced the light-emitting activity of the solution by 93% (Figure 4.9). In comparison, homogenized tissue from a nonluminous sponge species did not produce a light-emitting reaction when added to coelenterazine. Mixing nonluminous sponge homogenate and *Renilla* luciferase also did not produce light.

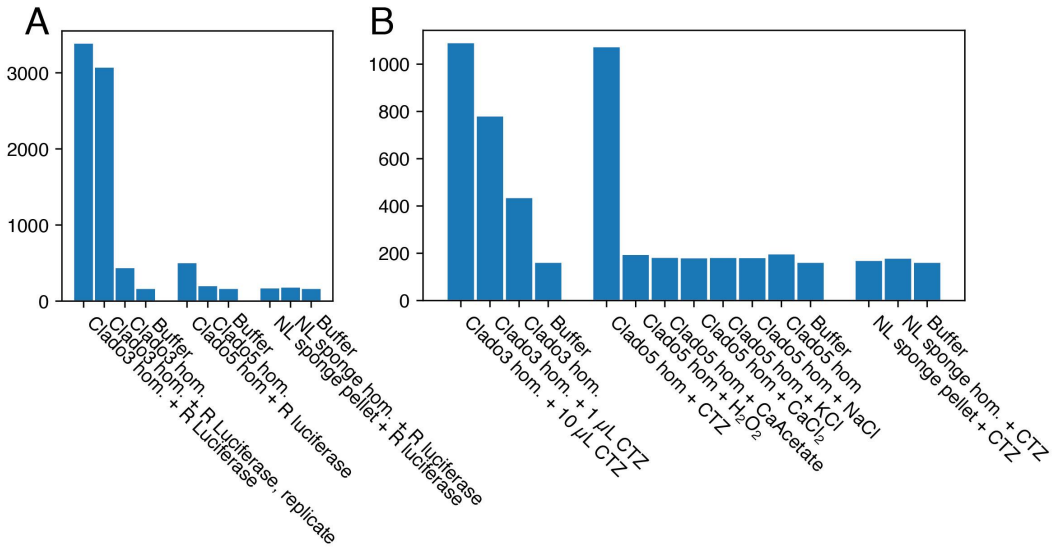

Figure 4.8 - Bioluminescence Assays for the Cladorhizid sponges. Relative Light Units (RLU) is integrated from 6 to 20 seconds**.** (A) Shows that *Renilla* luciferase (R luciferase) added to Clado3 and Clado5 homogenate (hom.) results in a bioluminescent reaction. Renilla luciferase added to a nonluminous (NL) sponge did not result in a bioluminescent reaction. This suggests that Clado3 and Clado5 contained coelenterazine while the non-luminous sponge did not. (B) Shows that coelenterazine (CTZ) added to Clado3 and Clado5 homogenate resulted in a bioluminescent reaction in a dose-dependent fashion (1µL and 10µL). Compounds that are possible triggers for bioluminescent reactions in photoproteins do not cause a bioluminescent reaction. Coelenterazine added to the non-luminous sponge homogenate does not produce a bioluminescent reaction.

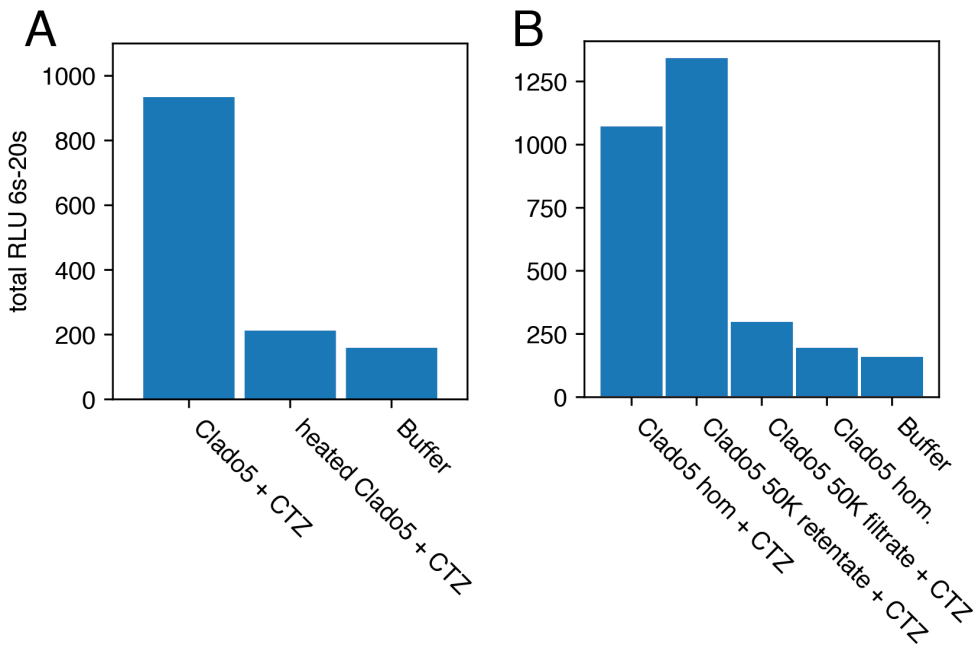

Figure 4.9 - Heat deactivation and activity concentration on a 50K filter. Relative Light Units (RLU) are integrated from 6 to 20 seconds. (A) Shows that heat treating the homogenate (hom.) caused a reduction in bioluminescence activity when coelenterazine (CTZ) was added. (B) Clado5 homogenate bioluminescence activity passed through a 0.2µm filter, but was concentrated on a 50KDa filter. These results together (Clado5 50K filtrate and retentate) imply that the bioluminescent system is protein-based and consists of a protein or protein complex larger than 50KDa. The Clado5 0.45 µm and 0.1 µm filter-clarified homogenate still emitted light when it was mixed with coelenterazine.

Molecular identification and Phylogeny

A contig containing COX1 was assembled using the whole-genome shotgun (WGS) library and validated using read coverage and AA similarity to closely related species.

The top two blastn hits for the complete COX1 nucleotide sequence were close to the COX1 sequences from *Cladorhiza* sp. (KX266208.1) and from *Chondrocladia* sp. (LN870486.1). Both blastn hits were 97.7% identical to the Clado1 query (Genbank accession number is MN418897). The Clado1 COX1 sequence was most closely related to species in the genus *Cladorhiza* (Figure 4.10).

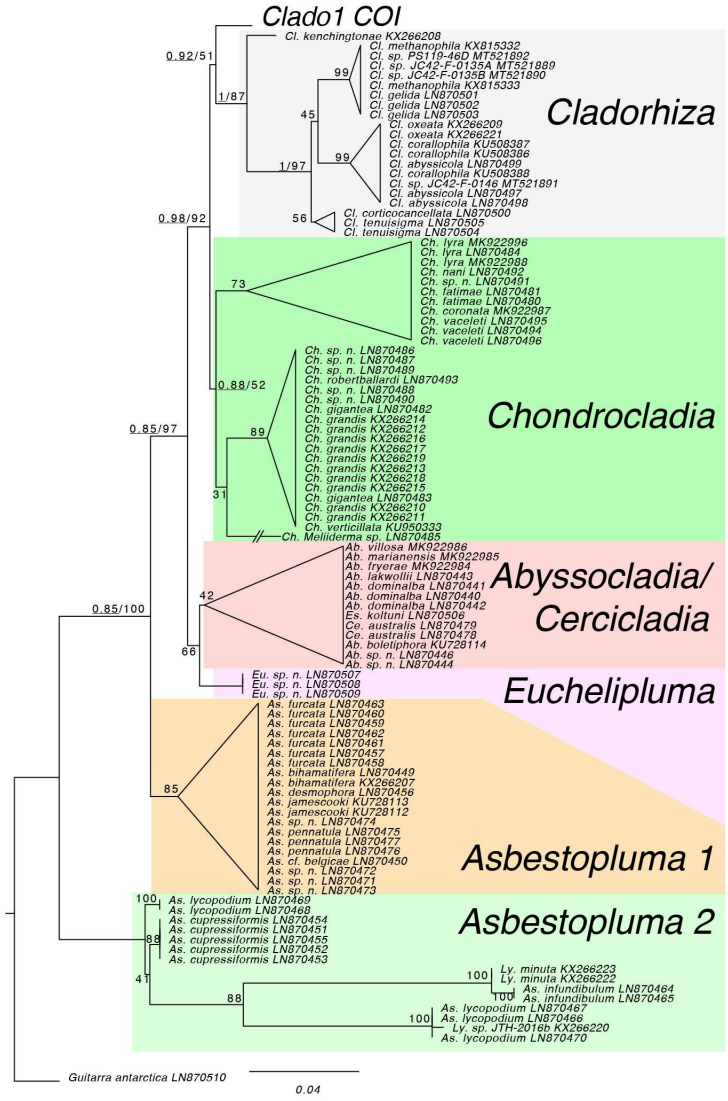

Figure 4.10 - Maximum likelihood and Bayesian phylogenetic analysis of COI locus. The branch lengths here are from the Bayesian analysis. The Bayesian and maximum likelihood analyses had the same tree topology, so support values show both the bootstrap values from RAxML and the posterior probability from MrBayes. The topology of the RAxML tree matched the genus-level topology of the 28S rDNA, COI and ALG11 tree from Hestetun et al 2016. The COI sequence from Clado1 falls as an outgroup of the *Chondrocladia* sp. and *Cladorhiza* genus. Bayesian support for major nodes are prepended, underlined, to bootstrap values.

Metagenomic Analysis

The analysis of 2,020,492 Illumina read pairs found 64,221 reads, or 3.17% of all read pairs, reads that were identifiable to a taxonomic unit (Figure 4.11). This low percent of identifiable reads is likely due to the fact that this sponge's genome, the primary component of the sequencing library, is not present in any online database. 31.6% of the identifiable read pairs (20,322) were of bacterial origin, although only 133 read pairs were identifiable as belonging to the bioluminescent bacterial genera *Vibrio* (133 read pairs - 0.2% of 64,221), *Photorhabdus* (42 read pairs - 0.007% of 64,221), or *Photobacterium* (11 reads pairs - 0.0002% of 64,221). The remaining largest contributors of identifiable reads were derived from Euteleost fish (8784 reads, 13% of 64,221). Most reads that map to fish were identifiable to the *Cyprinus carpio* genome, and were attributable to Illumina adapters erroneously included in the *Cyprinus carpio* genome and occurring at a low rate in the Illumina library. Other reads with a large number of identifiable reads were identifiable as coming from Poecilosclerida sponges (2786 reads, 4.3%), and protostomes (mostly nematodes, 1087 reads, 1.7%). We did not find any reads that were identifiable as ctenophore-origin, suggesting that the sponges do not consume the ctenophores that live on their silicious stalks. We also attempted to map the Clado1 reads to the complete bacterial luciferase operons from multiple bacterial species, but no reads mapped.

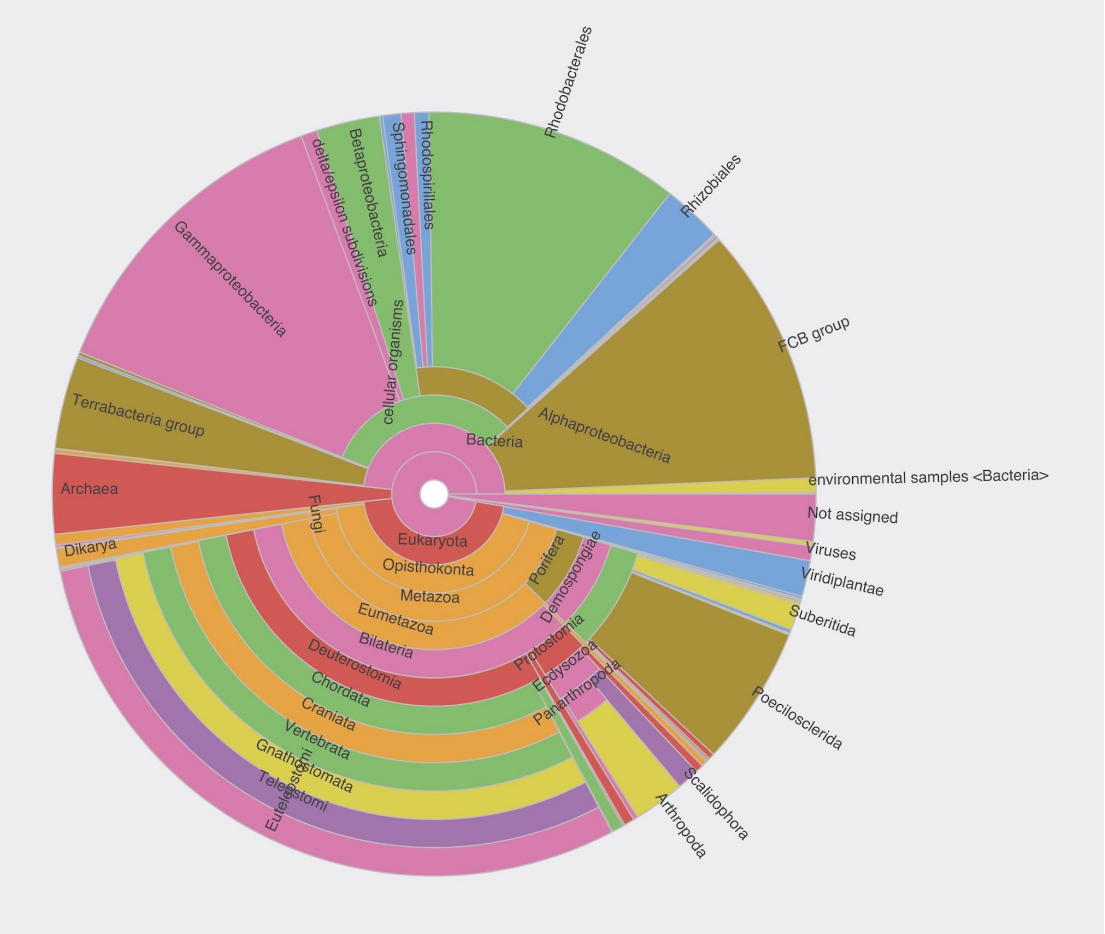

Figure 4.11 - Composition of taxa from metagenomic analysis. The proportion of 64,221 identifiable reads are shown here, split into taxa. Reads mapping to fishes were mostly false positive hits to the *Cyprinus carpio* genome, which erroneously has many Illumina sequencing adapters. (End of caption)

Sponge - Discussion

MBARI's Video Annotation Reference System (VARS (Schlining and Stout 2006)), a 30+ year record of deep-sea observations, contains many occurrences of Cladorhizidae species. They were observed from different locations where MBARI

has conducted ROV research in the Canadian Arctic Ocean, the northeast Pacific, the Gulf of California, and Hawaii. Numerous morphologically similar individuals of our new deep-sea sponge were observed at these sampling locations. Further phylogenetic analyses using additional loci, as well as an extensive morphological characterization, will be necessary to determine this bioluminescent sponge to species level. However, it is clear that these bioluminescent sponges are within a clade containing *Chondrocladia* and *Cladorhiza (Hestetun et al. 2016)*.

Doubtful observations of bioluminescence in sponges have been reported since the beginning of the 19th century (e.g, Pagenstecher 1881; Okada 1925). False observations may be induced by numerous bioluminescent symbiotic (Hentschel *et al.* 2006), entrained, or captured bacteria in sponges (Dahlgren 1916; Okada 1925). In our observations, light emitted upon mechanical stimulation was localized to the general area of stimulation, and only lasted for several seconds. The brief glow of several seconds observed in these sponges does not correspond to the continuous glow characteristic of bacterial bioluminescence. We showed that the sponge homogenate cross reacts with the luciferin coelenterazine, a molecule not used in any known bioluminescence biochemical system in bacteria or annelids. Another potential source of misinterpretation is the presence of partially digested luminous planktonic organisms, since these are carnivorous sponges. However, the sponge was cleaned before stimulation and light emission was observed in various parts of the sponge mechanically stimulated and for each specimen.

The sponge homogenate released light when mixed with coelenterazine even when it had passed through a 0.1 μm filter, a pore size through which only the smallest bacteria are known to pass (Wang *et al.* 2008). Moreover, in metagenomic analyses only a very small proportion of the reads were related to bioluminescent bacteria. We also checked carefully for contamination from other small organisms such as copepods or annelids. While we cannot verify with certainty that our observations are autogenic luminescence, the observation of similar bioluminescence on various body parts of six different specimens, from different collection times and locations, with corresponding light kinetics, and an absence of data suggesting that the luminescence is caused by bacteria or other contaminating animals certainly warrants further investigation. A future study confirming that the coelenterazine-luciferase is encoded in the sponge genome will be necessary to validate that these cladorhizid sponges are autogenically luminous.

Lastly, these findings raise questions around the evolution of marine bioluminescence in animals. In our hypothesis, these sponges use coelenterazine as their light-emitting molecule. Coelenterazine is the only luciferin used in the luminescent systems of a few protists and the other non-bilaterian clades, the Cnidaria and the Ctenophora (Haddock *et al.* 2010). The ancestors of these three phyla likely diverged before the Cambrian Explosion over 600 million years ago (Dohrmann and Wörheide 2017), and bioluminescence is thought to play a role in speciation (Ellis and Oakley 2016). Our results spark interest in the numerous roles of

bioluminescence and its involvement in the speciation and diversification of life in the ocean.

Acknowledgments

We thank the pilots of the ROV Doc Ricketts and the crew of the R/V Western Flyer. Kyra Schlining provided expertise onboard during both of these expeditions. Funding sources: This work was supported by the David and Lucile Packard Foundation support of MBARI, and S. Martini was funded, in part, by a grant from the Bettencourt-Schueller Foundation. Darrin Schultz was funded in part by GRFP DGE 1339067.

Data and materials availability

The Illumina sequencing reads for this project are available on NCBI at PRJNA556048.

Appendix A

Supplementary Materials for Chapter 2

A chromosome-scale genome assembly and karyotype of the ctenophore *Hormiphora californensis*

This text is adapted from a published article:

Darrin T. Schultz[†],Warren R. Francis[†], Jakob D. McBroome, Lynne M. Christianson, Steven H.D. Haddock, Richard E. Green. ***A chromosome-scale genome assembly and karyotype of the ctenophore Hormiphora californensis*.** (2021) G3: Genes, Genomes, Genetics

[†] - Indicates co-first authorship

Supplementary Materials and Methods

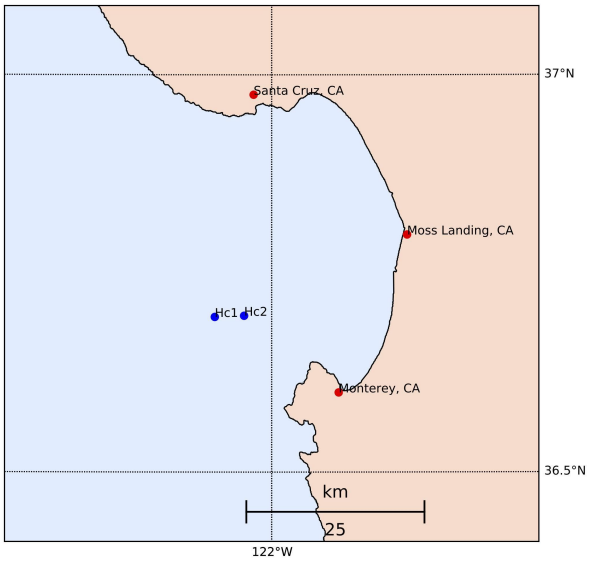

Figure A1 - *Hormiphora californensis* and *B. forskalii* sample collection map. *H. californensis* individuals Hc1 and Hc2 were collected within two kilometers of one another three years apart. See collection conditions and parameters in Table A1.

Sequencing data preparation

H. californensis HMW DNA was isolated from individual Hc1 by lysing tissue in CTAB buffer (Dawson *et al.* 1998), then purifying the DNA with a chloroform, phenol:chloroform, chloroform, ethanol precipitation protocol (Sambrook and Russell 2006). Two PacBio SMRT CLR sequencing libraries were constructed and sequenced on three SMRT cells on a PacBio Sequel or Sequel II machine at UCD, yielding 27.4 Gbp of CLR subreads (Figure S2). A Hc1 HMW DNA extract was also used to create three Dovetail Chicago libraries at University of California Santa Cruz (UCSC) (Putnam *et al.* 2016), using either the DpnII or MluCI

enzyme. The Chicago libraries were sequenced to a depth of 105 million read pairs. One Hc1 HMW DNA extract was used to construct a 10X chromium library at UCSC, and was sequenced to a depth of 74 million read pairs (Weisenfeld *et al.* 2017). Eight Hi-C libraries for individual Hc1 were constructed using less than 50mg of flash-frozen tissue per prep (Adams *et al.* 2020). Six libraries were made with DpnII, and two were made with MluCI. Four of these libraries were sequenced to a depth of 616.4 million read pairs, with each replicate having at least 95.9 million read pairs. In addition, we prepared one DpnII Hi-C library with tentacle tissue from individual Hc3, sequenced to a depth of 233.9 million read pairs.

Total RNA was isolated from *H. californensis* individual Hc1 by pulverizing 100 mg of frozen tissue under liquid nitrogen, then proceeding with a Trizol RNA isolation protocol (Rio *et al.* 2010). The RNA was assayed at the UC Davis (UCD) DNA Technologies Core. One Illumina TruSeq RNA Library Prep Kit v2 library was constructed from this RNA at UCD. This library was sequenced to a depth of 95 million read pairs. The UCD DNA Technologies Core also prepared an Iso-Seq library and sequenced this library on a single Sequel II SMRT cell.

Lastly, *H. californensis* shotgun libraries were prepared from Hc1 and Hc2 by isolating DNA using the Omega Biotek EZNA Mollusc DNA kit, shearing the DNA using a Bioruptor, and preparing libraries with insert sizes of 400-500bp using the NEBNext Ultra II WGS, NEBNext Ultra II FS, or Illumina TruSeq Nano DNA library prep kits. Hc1 libraries were sequenced to a depth of 120 million read pairs.

The Hc2 library was sequenced to a depth of 64 million 100PE reads on a HiSeq 2500 at the University of Utah DNA Sequencing Core Facility.

Trimming raw sequencing data

All Illumina libraries were trimmed with Trimmomatic v0.35 (Bolger *et al.* 2014) using the options `ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10:1:TRUE LEADING:3 TRAILING: 3 SLIDINGWINDOW:4:15 MINLEN:36`. All Hi-C and Chicago libraries were additionally trimmed by removing the 3′ end of reads after the restriction enzyme's junction sequence.

The PacBio Iso-Seq and WGS data were converted to circular consensus sequences using ccs v4.0 (github.com/PacificBiosciences/ccs). The Iso-Seq data then had the 5′ and 3′ cDNA primers removed using lima v1.10.0 (github.com/PacificBiosciences/barcoding), then polyA tails and chimeric sequences were removed with isoseq3 v3.2 (github.com/PacificBiosciences/IsoSeq). We used pauvre marginplot commit 13uhtt7 (github.com/conchoecia/pauvre) to check the overall consensus quality and length of transcripts (Figure S3) (De Coster *et al.* 2018).

Mitochondrial genome assembly

To assemble the *H. californensis* mitochondrial genome, we first mapped the PacBio Sequel CCS reads to the corrected *P. bachei* mitochondrial genome (Kohn *et*

al. 2012; Arafat *et al.* 2018) using minimap2 v2.17 (Li 2017) with parameters -ax asm20. Reads that mapped to the *P. bachei* mitochondrial genome were assembled using canu v2.1.1 (Koren *et al.* 2017) with the options `genomeSize=15kb -pacbio-corrected`. We used Geneious v11 to identify the largest ORF, and used blastn v2.6.0 (Altschul *et al.* 1997) to identify that the ORF encoded COX1. We selected the start codon of the COX1 gene to be the 5′-most position of the mitochondrial genome, as is conventional with previous ctenophore mitochondrial genome annotations (Kohn *et al.* 2012; Pett and Lavrov 2015; Arafat *et al.* 2018). The sequence was trimmed up to the start codon of the COX1 gene on the canu contig. To confirm that the sequence was circular, we mapped the CCS reads to two concatenated copies of the mitochondrial genome using minimap2.

The mitochondrial genome assembly for individual Hc2 was generated by mapping the trimmed Hc2 Illumina WGS reads to the Hc1 mitochondrial assembly using BWA-MEM (Li 2013), then correcting the reference using pilon (Walker *et al.* 2014). We mapped the reads back to the pilon-corrected reference to verify that it was correct.

The final 12564 bp Hc1 mitochondrial genome assembly was annotated by mapping the rRNA and CDS sequences from the corrected *P. bachei* mitochondrial genome (Arafat *et al.* 2018) to the assembly using Geneious v11. Geneious was then used to predict ORFs using the Mold Protozoan Mitochondrial translation table. ORF start sites that were conserved between Hc1 and Hc2 were used to delimit the beginning of the transcripts.

To annotate the ribosomal RNA boundaries we mapped the untrimmed RNA-seq reads to the final assembly with BWA-MEM (Li 2013). The start and stop sites for each ribosomal RNA were selected by finding positions that had several reads with the same start/stop site followed by a fast attenuation in coverage, also guided by the length of the *P. bachei* ribosomal RNA sequences. I-TASSER was to predict the protein structure and to find the best structural analogs for the conserved URFs present in the genomes (Yang *et al.* 2015). We used the TMHMM tool to predict transmembrane domains for the URFs (Krogh *et al.* 2001). We used tRNAscanSE and ARWEN to search for mitochondrially-encoded tRNAs (Lowe and Eddy 1997; Laslett and Canback 2008).

Phylogeny construction

Full-length ctenophore 18S sequences were downloaded from NCBI, aligned using MUSCLE, then trimmed such that each sequence had greater than 90% occupancy. This alignment was used in a rapid bootstrapping maximum likelihood (ML) search of 250 trees with the GTR GAMMA model using RAxML v7.2.8 (Stamatakis 2014). A tree for COX1 nucleotide sequences was constructed in the same fashion. The mitochondrial nucleotide alignment was constructed by individually performing translation alignments on the COX1, COX2, COX3, CYTB, ND1, ND2, ND4, and ND5 loci from multiple species using MAFFT v7.388 (Katoh *et al.* 2002). The alignments were concatenated, and a RAxML ML tree was constructed using the parameters described above. A Bayesian tree was constructed with the same concatenated protein alignment using MrBayes v3.2.6 (Ronquist and Huelsenbeck 2003), with *Tethya actinia* as an outgroup, the HKY85 substitution model, gamma rate variation, chain length of 30000, 4 heated chains, 0.2 heated chain temp, subsampling frequency every 200 trees, a 2500-tree burn-in, and a random seed of 1910.

Genome assembly

The wtdbg2 assembler v2.4 (Ruan and Li 2019) with parameters `-g 85m -p 0 -k 15 -e 3 -A -S 2 -s 0.05 -L 5000 -R --aln-dovetail 10240` was used to *de novo* assemble the PacBio CLR subreads. The assembly was polished with arrow v2.2 (github.com/PacificBiosciences/gcpp), then with pilon v1.22 (Walker *et al.* 2014) using the Illumina WGS libraries. Haplotigs were removed

with Purge Haplotigs v1.0.4 (Roach *et al.* 2018) using parameters `purge_haplotigs cov -l 50 -m 175 -h 600 -j 70 -s 80` and `purge_haplotigs purge -a 30`. We then ran purge_haplotigs clip to remove overlapping contig ends.

Dovetail Genomics HiRise (v Aug 2019) was used to scaffold the genome first using the Chicago libraries, then using the Hi-C libraries (Putnam *et al.* 2016). We mapped shotgun reads to the contig assembly with BWA-MEM v0.7.17 (Li 2013) and calculated the mean coverage and GC content using BlobTools v1.1.1 (Laetsch and Blaxter 2017). Scaffolds with a mean coverage of less than 100, or having greater than 50% GC, were removed from the assembly. The resulting assembly was gapfilled using LR Gapcloser with the PacBio subreads (commit 156381a) (Xu *et al.* 2019). The assembly was then polished with pilon using the Illumina WGS libraries.

Hi-C heatmap generation

We generated a Hi-C heatmap to check for genome misassemblies. The Hi-C reads were mapped to the genome assembly using BWA-MEM with options `-5SPM` (Li 2013), the BAM was converted to a sorted and deduplicated pair file with pairtools v0.3.0 (github.com/mirnylab/pairtools), the pairs file was indexed with pairix v0.3.7 (github.com/4dn-dcic/pairix), then the pairs file was converted to a normalized mcool file using Cooler v0.8.10 (Abdennur and Mirny 2020). Additionally, we generated a PretextMap Hi-C matrix (github.com/wtsi-hpag/PretextMap commit ee1bf66). To visualize the matrices we

used HiGlass v1.10.0 (Kerpedjiev *et al.* 2018) or PretextView v0.1.0 (github.com/wtsi-hpag/PretextView).

Variant Calling

To call variants to be used in phasing and in other analyses, we first mapped the PacBio CLR WGS reads to the genome using minimap2 v2.17 (Li 2017), and mapped the Hc1 Illumina WGS reads to the genome using BWA-MEM and samtools (Li *et al.* 2009; Li 2013). We then called variants using these two BAM files as inputs to the software freebayes and gnu parallel (Tange and Others 2011; Garrison and Marth 2012). We filtered the VCF file to only include diploid calls.

To phase the variants we then marked duplicates in the Hi-C BAM file using Picard v2.25.1 ("Picard Toolkit" 2016), then used HapCUT2 v1.3.1 (Edge *et al.* 2017) extractHairs on the Hi-C, Chicago, and PacBio CCS BAMs. For the PacBio subreads we used the extractHairs parameters `--pacbio 1 --new_format 1 --indels 1`. For the Hi-C reads we used the HapCUT2 extractHairs parameters `--hic 1 --new_format 1 --indels 1`. For the Chicago reads we used the HapCUT2 extractHairs parameters `--maxIS 10000000 --new_format 1 --indels 1`. We then concatenated these fragment files and used them as input to phase the genome using HapCUT2 with the parameters `--hic 1 --outvcf 1` (Edge *et al.* 2017).

Genome annotation

The genome annotation is composed of manually-selected transcripts from several software packages, including BRAKER, GeneMark-ES/ET, AUGUSTUS, Stringtie, pinfish, and the cDNA cupcake pipeline. Blast results to the *Mnemiopsis leidyi* v2.2 proteins or the SwissProt database (Skinner *et al.* 2009; Robinson *et al.* 2011) were also used as additional sources of evidence. To generate the individual annotations, we performed the following:

BRAKER, GeneMark-ES/ET **and AUGUSTUS**: Illumina RNA-seq reads were aligned to the genome assembly using STAR v2.7.1a (Dobin *et al.* 2013), and the Trinity transcriptome and PacBio Iso-Seq reads were aligned to the assembly using minimap2 with option `-x splice:hq`. AUGUSTUS and GeneMark-ES/ET annotations were generated by running BRAKER v2.14 with the Illumina RNA-seq, PacBio Iso-Seq, and Trinity transcriptome BAM files as inputs (Stanke *et al.* 2004; Lomsadze *et al.* 2014; Hoff *et al.* 2019).

Cupcake: We mapped the full length, non-chimeric (FLNC) PacBio Iso-Seq reads mentioned above in "Sequencing read preparation" to the *H. californensis* genome using minimap2 with the parameters `-ax splice -uf --secondary=no -C5`. We then used the PacBio Cupcake tools to collapse the FLNC reads into transcript models (github.com/Magdoll/cDNA_Cupcake). We generated one set of

transcripts containing singletons, and one dataset without singletons, using the command `filter_away_subset.py --fuzzy_junction 5`.

Stringtie: Transcripts were predicted from the BAM file output of the minimap2 FLNC PacBio Iso-Seq-to-genome alignment using StringTie v2.0.4 (Pertea *et al.* 2015). Long parameters were used (`-L`) and the minimum isoform fraction was set to 0.1 (`-f 0.1`), with otherwise default parameters.

Pinfish: Transcripts were also predicted from the long reads using pinfish (github.com/nanoporetech/pinfish), with minimum isoform percentage set to 20, a minimum cluster size of 2 reads (`-p 20 -c 2`) and otherwise default parameters.

Manual inspection of each of the four annotations revealed many genes were erroneously fused or broken, compared to the true isoforms evident from the Iso-Seq data mapped to the reference. Because we found that each of the four annotations described above were imperfect, we chose to manually curate the annotation of the *H. californensis* genome. To ensure that the quality of the manual annotation was consistent across all 110 Mb, we developed a set of rules for difficult-to-annotate genes, like nested genes, gene clusters that appeared to have a trans-spliced leader

exon, and how to combine multiple annotations into a single gene. These guidelines are available for download (github.com/conchoecia/hormiphora and Zenodo DOI: 10.5281/zenodo.4074309).

Transcript phasing

We first generated a transcript sequence for each isoform in the genome annotation with gffread (github.com/gpertea/gffread), then non-splice aligned the Illumina RNA-seq and PacBio Iso-Seq reads to the transcripts with BWA-MEM and minimap2 (Li 2013, 2017). We then used freebayes to call variants for each isoform in parallel (github.com/ekg/freebayes) (Tange and Others 2011), then phased each isoform with WhatsHap (Patterson *et al.* 2015). A new reference sequence for each haplotype was generated using bcftools consensus (Li 2011), then haplotype-specific Iso-Seq reads were used to correct the new haplotype-specific isoform using pilon v1.22 (Walker *et al.* 2014). These isoforms were then mapped to the reference genome using minimap2 `-ax splice`, phased with WhatsHap, then matched with the whole-genome phase variant phase blocks.

The longest ORFs from the phased and polished transcript isoforms were predicted using prottrans.py using the parameters `-a 50 -r` (bitbucket.org/wrf/sequences/src/master/prottrans.py). For each gene isoform we randomly selected one of the amino acid sequences from one of the haplotypes. When the amino acid sequence from one haplotype was longer than the amino acid sequence on the other haplotype, we selected the longer one.

P. bachei genome reannotation

As no structural annotation, specifically no GFF file, was provided with the *P. bachei* genome, we created an exon-by-exon annotation file in GFF format from the reported scaffolds and transcripts for use in our whole-genome comparisons with *H. californensis*. The transcripts were mapped to the scaffolds with minimap2, using the options `-x splice --secondary=no` . Based on the mapping positions of each transcript in the BAM file, a GFF file was generated using pinfish (github.com/nanoporetech/pinfish) with the option `-g`. Of the 18950 transcripts, 18947 mapped back to the genome. For many protein comparisons, the proteins and transcripts provided with the P. bachei genome were insufficient due to the fragmented nature of the source scaffolds.

Next, we generated gene models using the AUGUSTUS web server (http://bioinf.uni-greifswald.de/webaugustus/index) (Hoff and Stanke, 2013) using the transcript models as the training set. This yielded two versions, the "hints" set and the "ab initio" set. As the "hints" version closely matched the transcript models, and likewise any gene fusions or breaks of that dataset, we instead used the "ab initio" set for downstream analyses.

Lastly, due to the relatedness between *H. californensis* and *P. bachei*, we examined whether we could simply map the *H. californensis* model transcripts to the *P. bachei* scaffolds using minimap2, with the options `-x splice --secondary=no`. With this strategy, 99% of *H. californensis* transcripts mapped

to *P. bachei*. 8000 of the transcripts had an additional mapping, likely due to fragmentation across different scaffolds or matching to both of a pair of uncollapsed haplotigs. We then used pinfish to generate a GTF file, as used above for the transcript model set.

Assessing fragmentation and fusion of genes

Using the *H. californensis* protein set, we used a custom Python script (compare_hcal_ref_proteins.py) to examine fragmentation of the *M. leidyi* protein set. The script uses the coordinates of local alignments generated by diamond (Buchfink *et al.* 2015) to check whether a protein in *H. californensis* contains multiple non-overlapping alignments to *M. leidyi* proteins on the same scaffold. Although this could mutually imply an erroneous fusion of two genes in *H. californensis*, the use of Isoseq reads for annotation makes this scenario unlikely. Nonetheless, out of around 1200 *M. leidyi* proteins that were identified as fragmented, we then manually checked a set of 384 genes (all those with 3 or more fragments, as well as others) and found that all of them were indeed fragmented. Most of these had correct isoforms from *de novo* transcriptome assemblies.

Supplementary Results

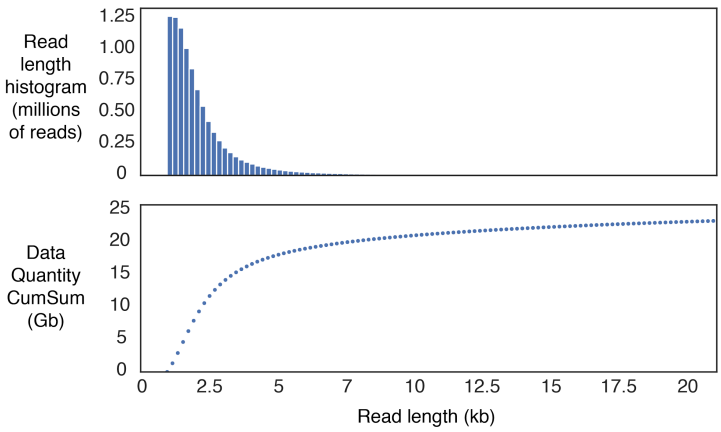

Figure A2 - PacBio subread size distribution. Read length distribution (top) and cumulative sum of total basepairs (bottom) of the PacBio Sequel and Sequel II subreads.

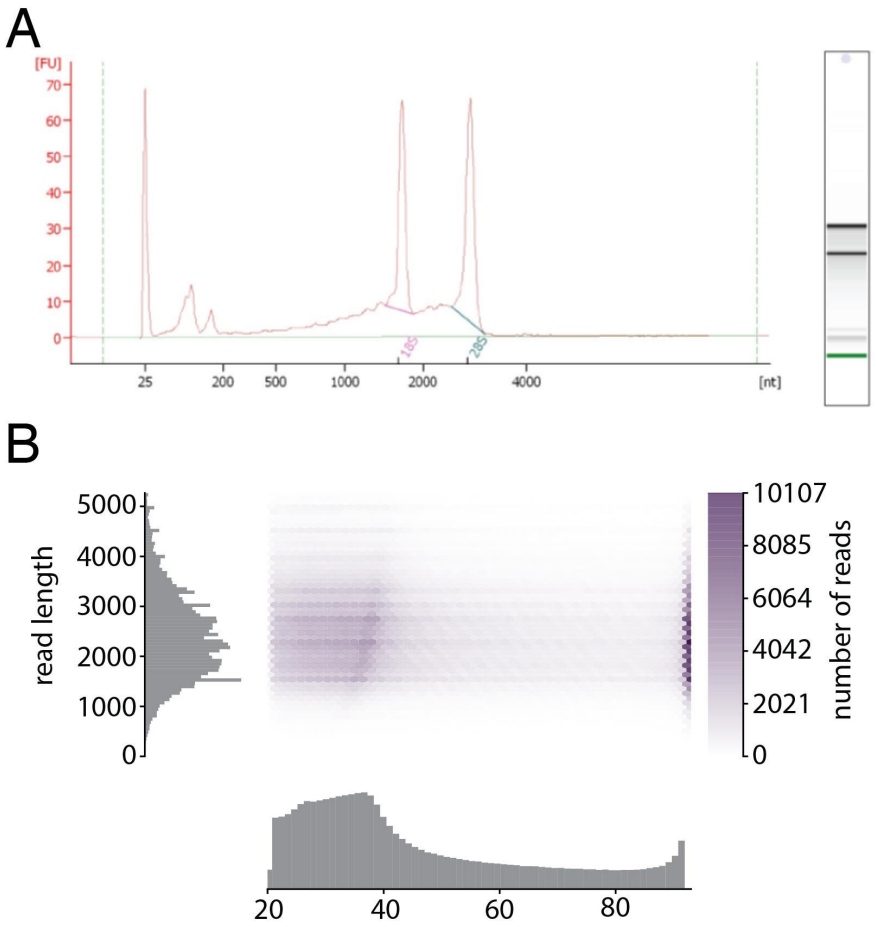

Figure A3 - RNA and IsoSeq size distribution. (A) The Agilent Bioanalyzer trace of the RNA used to create the PacBio Iso-Seq library Hc1_lib18_run1_PB_Iso-Seq (SRR10403581 and SRR10403849). The RNA used for the library was largely intact. (B) A heatmap of the Iso-Seq reads after consensus calling with the ccs software and filtering to retain full-length, non-chimeric sequences. The read length histogram roughly resembles the RNA size distribution in Panel A.

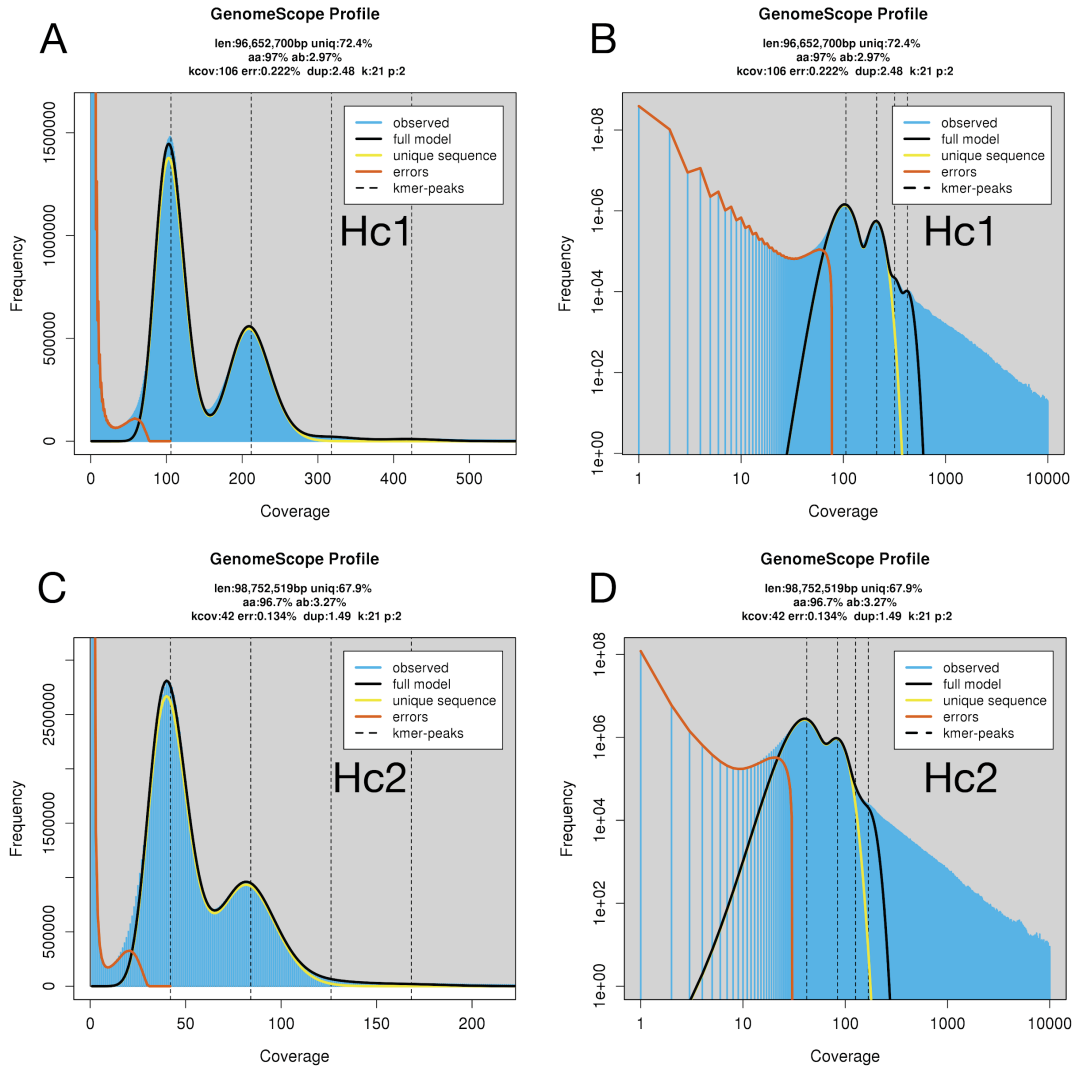

Figure A4 - *H. californensis* k-mer based genome size prediction. (A,C) The haploid genome size estimate from the GenomeScope2 for Hc1 (A) was 96.6 Mb, and for Hc2 (C) was 98.72 Mb. Altogether, the Illumina WGS libraries from Hc1 had 212x genome coverage, and the Hc2 Illumina WGS library had 82x genome coverage. The *H. californensis* genome appears to be diploid from the k-mer spectrum based on the presence of two peaks in both A and C. Panels B and D are log-log plots of panels A and C.

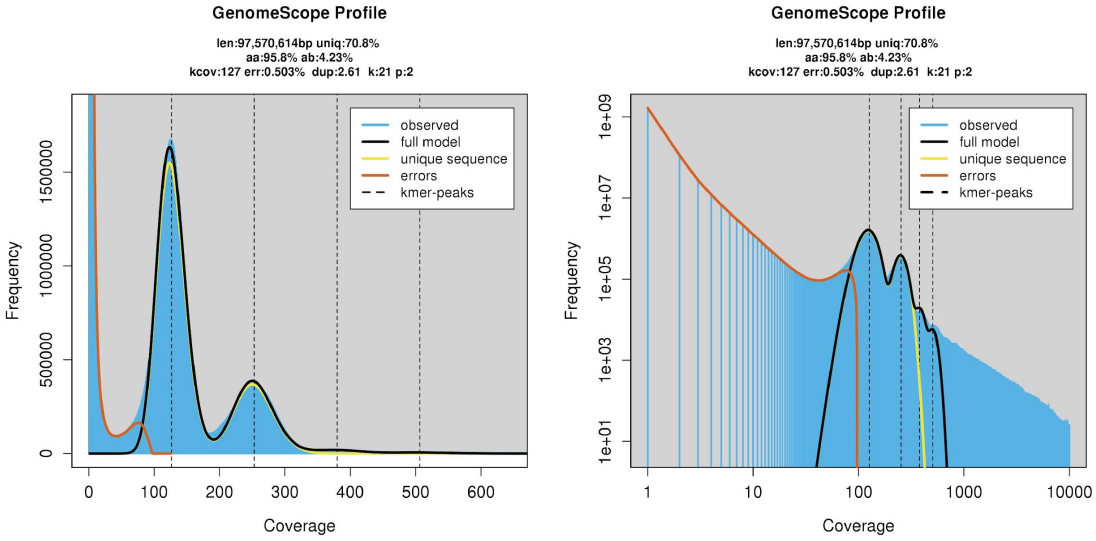

Figure A5 - *P. bachei* k-mer based genome size prediction. We predicted the *P. bachei* genome size using publicly-available single-individual WGS data from SRA, the jellyfish k-mer counter, and GenomeScope2. The predicted haploid size was 97.57 Mb. This predicted size is very close to the predicted size of the *H. californensis* genome (96-98 Mb) . Altogether, the shotgun libraries had approximately 250x coverage of the genome. Similar to the *H. californensis* k-mer spectrum, this plot suggests that the animal is diploid.

Assembly Step	% (Ill./ PB) WGS reads mapping	Number of Contigs	Number of Scaffolds	Assembly Size (Mb)	contig N50 (kb)	scaffold N50	BUSCO stats				
							(C)	(S)	(D)	(F)	(M)
wtdbg2	(97.85 / 94.08)	1769	1769	113.14	144	143 kb	58.8	58.1	0.7	18.8	22.4
arrow + pilon	(97.85 / 95.14)	1769	1769	113.15	144	143 kb	88.8	86.8	2	5.3	5.9
purge haplotigs	(98.03 / 94.57)	1309	1309	106.89	152	152 kb	87.5	85.8	1.7	5.6	6.9
blobtools	(/ 94.06)	1283	1283	106.44	153	152 kb	87.5	85.8	1.7	5.6	6.9
HiRise Chicago	(/)	1334	287	106.55	150	822 kb	88.4	87.1	1.3	4.6	7
HiRise Hi-C	(/)	1340	44	106.57	150	8.14 Mb	87.8	86.1	1.7	5.3	6.9
PBjelly	(/)	975	44	110.67	204	8.54 Mb	88.4	87.1	1.3	5.3	6.3
LRGC	(/ 95.16)	355	44	110.67	581	8.54 Mb	88.5	86.8	1.7	5.3	6.2
pilon	(98.24 / 95.32)	355	44	110.66	580	8.54 Mb	89.4	88.1	1.3	4.6	6

Table A1 - Statistics through the *H. californensis* genome assembly stages. Each row of this table shows various statistics after each step of the assembly. The percent of Illumina and PacBio WGS reads that map to the genome, the contig N50, the scaffold N50, and the BUSCO nucleotide mode completeness scores increase with subsequent assembly steps.

		Protein-coding genes				
Annotation Step	Number of non-Protein Coding Genes	Number of Protein-Coding Genes	Number of proteins with hits >1e-5 to nr	Number of proteins without hits >1e-5 to nr	Number of Proteins that do not appear in Mnemiopsis or Pleurobrachia genomes	Total Number of Genes
1. Genes added from Iso-Seq	248	12,987	8,420	4,567	619	13,235
2. Genes added from AUGUSTUS	38	1,170	585	585	95	1,208
3. Genes added from Pleurobrachia transcripts	23	108	20	88	10	131
Totals	309	14,265	8,945	5,320	714	14,574

Table A2 - Genome Annotation Steps. This table includes the total number of genes added at each annotation step. There were 13,236 genes that had Iso-Seq read support. There were 12,987 that were protein-coding and 249 that were not protein-coding. For each step we also included the number of protein coding genes that had significant hits to nr, and the number of protein-coding genes that did not appear in the *Mnemiopsis* or *Pleurobrachia* genomes' proteins, but appeared in the transcriptomes of other ctenophores. The total number of protein-coding genes that we identified was 14,265. The total number of genes that we identified, including non-protein-coding genes, was 14,574.

Dataset	Complete	Complete + Partial	Number of missing core genes	Average number of orthologs per core genes	% of detected core genes that have more than 1 ortholog	BUSCO string
Protein Models	281 (92.74%)	291 (96.04%)	12 (3.96%)	1.18	10.68%	C:92.7%[S:82.8%,D:9.9%] F:3.3%,M:4%
Genome	270 (89.11%)	286 (94.39%)	17 (5.61%)	1.01	0.74%	C:89.1%[S:88.4%,D:0.7%] F:5.3%,M:5.6%
IsoSeq FLNC	290 (95.71%)	296 (97.69%)	7 (2.31%)	101.57	96.55%	C:95.7%[S:3.3%,D:92.4%] F:2.0%,M:2.3%
Illumina *de novo* Transcrip -tome	299 (98.68%)	300 (99.01%)	3 (0.99%)	2.35	71.57%	C:98.7%[S:28.1%,D:70.6%] F:0.3%,M:1.0%

Table A3 - BUSCO scores. These BUSCO protein mode scores were calculated using gVolante (Nishimura *et al.* 2017).

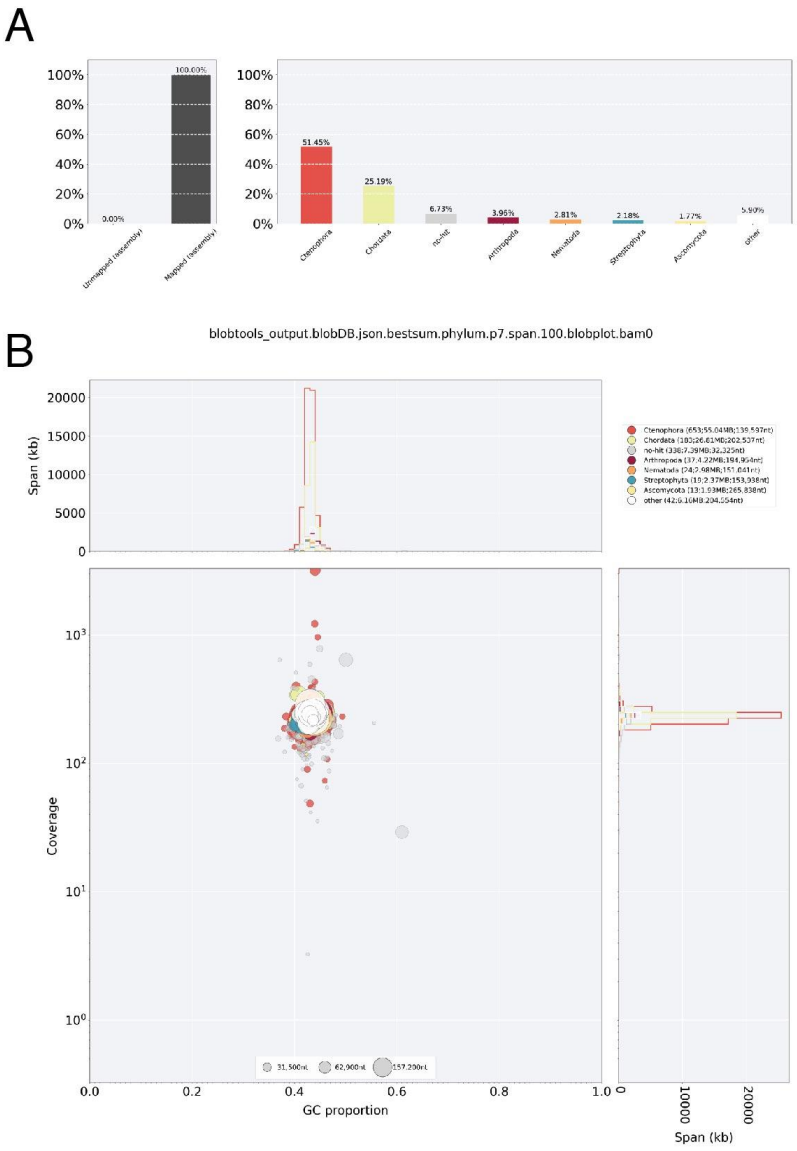

Figure A6 - *H. californensis* assembly intermediate blobtools plot. (A) While the BlobTools taxonomic classification suggests that there is a large amount of contamination in the DNA sequencing libraries, the GC and read depth coverage plot (B) suggests otherwise. Most contigs had close to 42% GC and had a mean read depth coverage around the haploid k-mer coverage of 212.

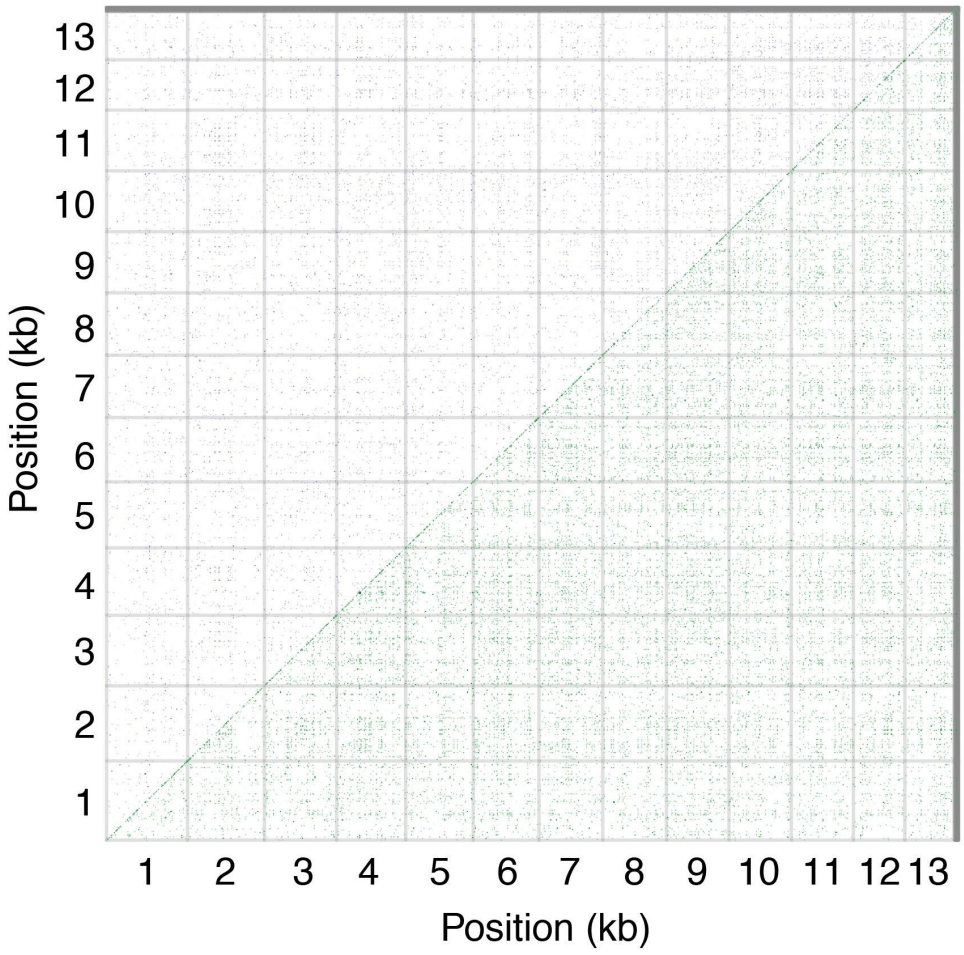

Figure A7 - D-genies genome dotplot. A dot-plot of the entire genome aligned against itself, showing that very few regions are duplicated, and there are no large segmental duplications. Light-green lines indicate matches (below diagonal), purple lines indicate reverse-complement matches (above diagonal). Short lines appear as dots.

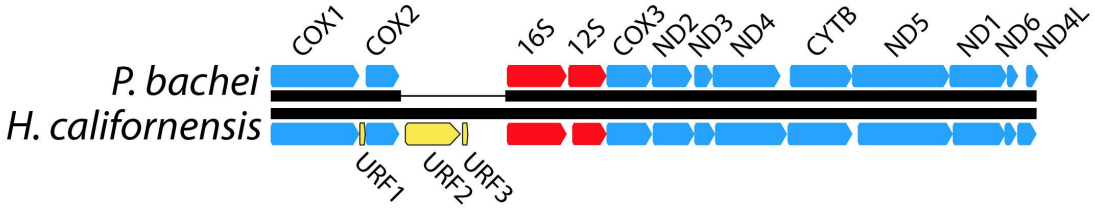

Figure A8 - A synteny plot of *P. bachei* and *H. californensis* mtDNA. These two species share the same gene order, except that *H. californensis* has a large insertion between the COX2 gene and the 16S gene. The insertion in the *H. californensis* mtDNA contains two URFs. URF1, URF2, and URF3 occur in both Hc1 and Hc2.

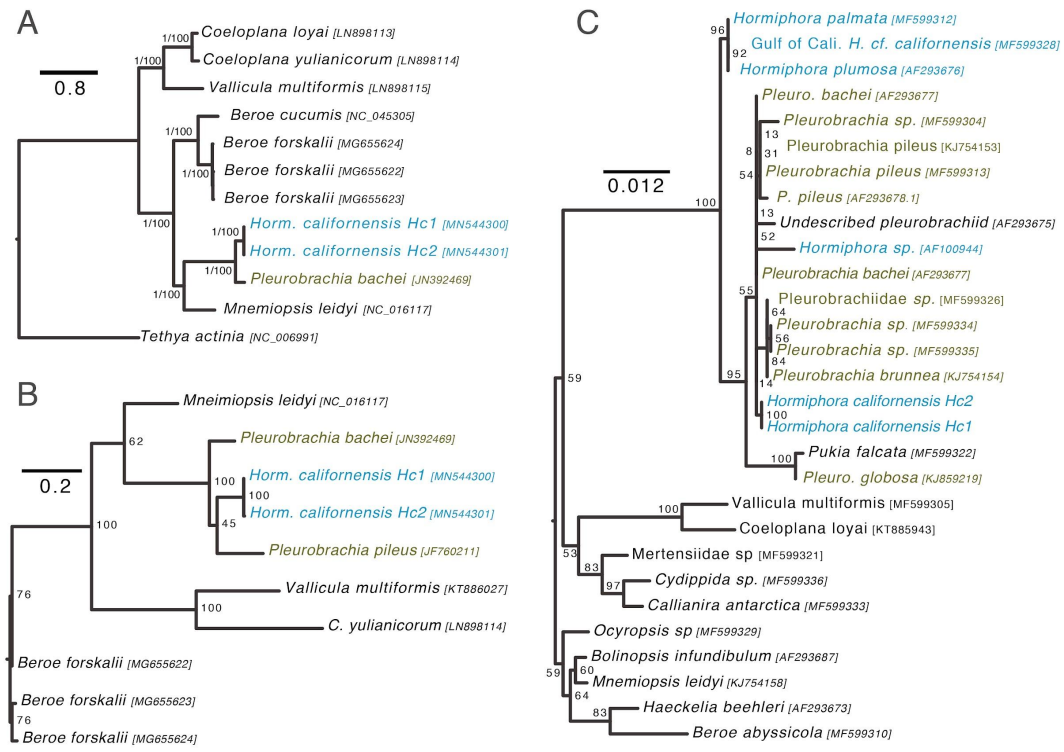

Figure A9 - Phylogenetic position of Hc1 and Hc2. (A) Ctenophore mitochondrial protein tree, including the COX1, COX2, COX3, CYTB, ND1, ND2, ND4, and ND5 loci. Node labels are posterior probability from the Bayesian tree, and the bootstrap value from the maximum likelihood tree. All nodes had a posterior probability of 1 and a bootstrap value of 100. (B) A COX1 nucleotide tree using additional COX1 sequences from NCBI. Node labels are bootstrap values. Samples Hc1 and Hc2 are in a clade within the genus *Pleurobrachia*. (C) A 18S ctenophore tree. Node labels are bootstrap values. Samples Hc1 and Hc2 lie within a polytomy of other *Pleurobrachia* species, but are distinct from *H. palmata*, *H. plumosa*, and a *H. californensis* sample from the Gulf of California.

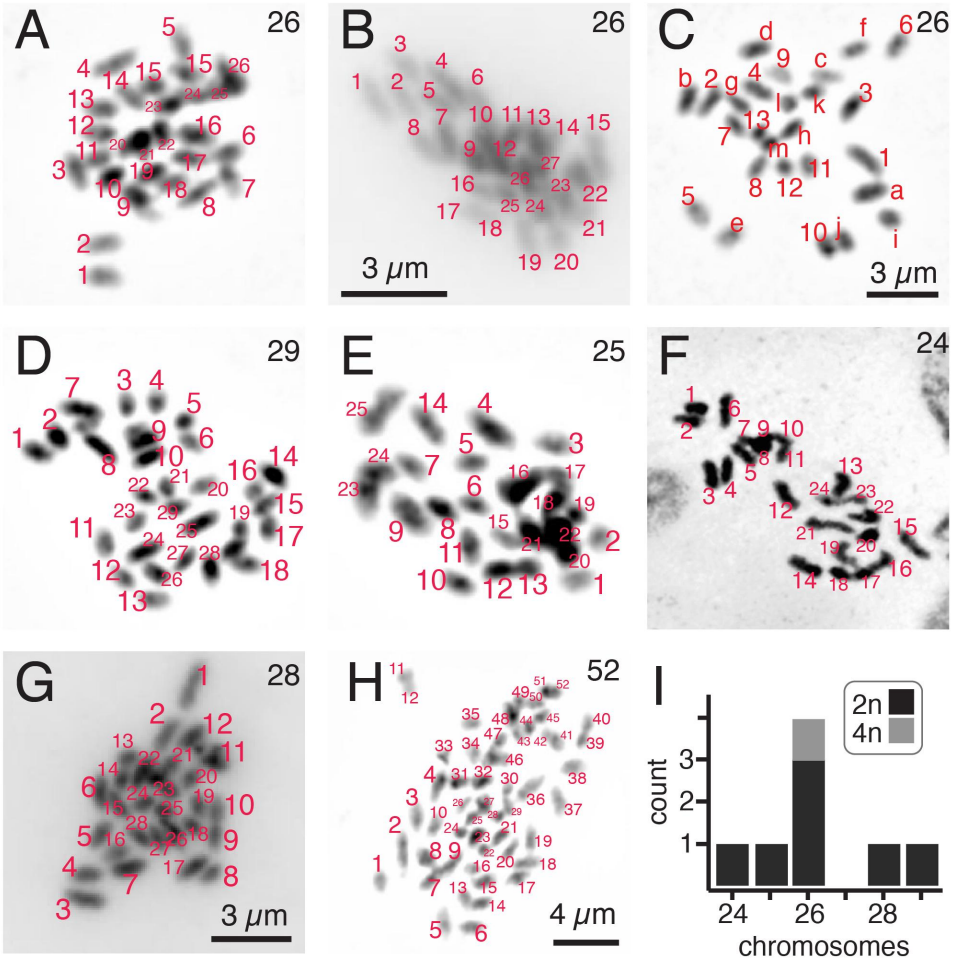

Figure A10 - *Hormiphora californensis* karyotyping results. Panels (A-H) are the karyotyping results from individual embryos stained with DAPI, image color inverted and grayscale. Each chromosome is numbered 1-N. Numbers in black is the total number of chromosomes estimated in that panel. Panel (C) is numbered in pairs using the same scheme as Figure 1. (I) shows a histogram of the number of times that each chromosome count was observed. There is one 4n count included in the 26 chromosomes bin, as the count was 52 and likely corresponded to a 2n of 26. A 2n of 26 corresponds to 13 pairs of chromosomes.

Gene number and synteny with other ctenophores

For *M. leidyi*, the ML2.2 annotation and protein set were compared to the *H. californensis* proteins with the script scaffold_synteny.py (Zenodo DOI: 10.5281/zenodo.4074309). We examined gene positions for 450 of the longest scaffolds in *M. leidyi*, accounting for 110Mb, whereby the smallest scaffold examined was 114kb in length. Of the 8685 query proteins with matches to any *M. leidyi* protein, 6422 gene matches were retained after filtering for quality and matches to the longest scaffolds.

We then examined collinearity of genes between the two genomes, again based on unidirectional BLAST hits, requiring at least 3 genes in a row, allowing up to 5 intervening genes. This identified 571 blocks containing 2258 total genes, though 439 of these blocks contained either 3 or 4 genes, suggesting that collinearity is limited between the two species. As the script that identified these blocks allows for two tandem genes to hit the same query, false positives from fragmented genes may account for some of these. For instance, we found 279 cases where the gene in *H. californensis* spans two or more genes in the ML2.2 annotation.

The *P. bachei* genome size prediction using GenomeScope2 was 97.57 Mb (Figure S5) - only 62.5% of the size of the published assembly, 156.1 Mb (Moroz *et al.* 2014). The predicted *P. bachei* genome size of 97.57 Mb is very close to the predicted genome size of *H. californensis*. Based on the mean read mapping depth per-scaffold, it appeared that haplotypes were collapsed for 5310 scaffolds, but that over half of the *P. bachei* scaffolds were unmerged haplotypes. If the remaining

16669 scaffolds were collapsed into a haploid representation this would yield a final estimated genome size of 107Mb, close to the size of the *H. californensis* genome. This suggests that only one third of the *P. bachei* assembly represents a haploid assembly. This may also account for the additional ~7000 proteins predicted in the *P. bachei* genome compared to *H. californensis*.

Therefore, we used two approaches to estimate coliniarity between *H. californensis* and *P. bachei*. First, we tried an analysis of only the 59Mb of scaffolds that had a mean coverage close to the haploid k-mer coverage of 250x. Of these scaffolds, the longest was only 221kb, therefore broad scale synteny could not be effectively analyzed. Of the original 18950 *P. bachei* transcripts, 7076 mapped to one of the 5310 haplotype-collapsed *P. bachei* scaffolds. We used this geneset for microsynteny analyses with *H. californensis*. In total, 299 putative collinear gene blocks of at least 3 genes were identified, accounting for 1280 genes. Overall, the high number of scaffolds in the *P. bachei* genome hampered our ability to detect microsynteny between *P. bachei* and *H. californensis*. Despite their relatedness, this was lower than the detectable synteny between the more phylogenetically distant *M. leidyi* and *H. californensis*.

Next, we tried reanalyzing the *P. bachei* scaffolds using an ab initio annotation from AUGUSTUS. This program had predicted 32683 total proteins across the *P. bachei* assembly, though the density is much higher than the v1 transcripts. For example, on the longest Pbac scaffold of 320kb, there are 12 mapped transcripts but 31 AUGUSTUS genes are predicted. *Ab initio* gene predictions have difficulty

resolving nested genes, which is disabled by default in AUGUSTUS, thus many of these predictions are likely to be fragments of larger genes that are split by nested genes. We analyzed microsynteny between *H. californensis* and the *P. bachei* AUGUSTUS annotation, using *H. californensis* as the query. This had identified 983 blocks for 5025 genes, more than twice the count from the original transcript annotation. If the *P. bachei* AUGUSTUS predictions were instead used as the query, this identified 1648 blocks with 7803 genes, in many cases spanning the entire scaffold. Because multiple query genes were allowed to map to a single target gene, this increase of almost 3000 genes is likely due to the fragmented AUGUSTUS predictions. Nonetheless, it is evident that there is substantial synteny between *H. californensis* and *P. bachei*.

Comparison to ML2 Assembly and Annotation

The *M. leidyi* ML2 annotation (Ryan *et al.* 2013) had 16545 proteins, almost 2000 more than the *H. californensis* v1 annotation from this study. We sought to explain the large difference in protein number using synteny and orthology information from blast searches.

We found 1200 neighboring ML2 proteins that were bridged by a single *H. californensis* protein, suggesting that either the *M. leidyi* proteins are falsely split, or the *H. californensis* protein is a false fusion. The majority of these cases only had two neighboring *M. leidyi* genes, though there were 8 cases of 4 or 5 neighboring *M.*

leidyi genes that were bridged by a single *H. californensis* protein. In all cases, these transcripts were supported with single Iso-Seq reads in *H. californensis*.

We manually corroborated these 8 cases by comparing the *M. leidyi* genes to the *H. californensis* ortholog, matches in publicly-available transcriptomes of other ctenophores (Francis *et al.* 2015; Whelan *et al.* 2015, 2017), and orthologs in other animals. This analysis revealed that all 8 proteins appear to be fragmented in *M. leidyi* and the *H. californensis* version appears to be complete. Generally these genes were large, and many included nested intronic genes. These included homologs of Midasin (4284AAs), Pecanex (2096AAs), Dynein heavy chain 14 (4735AAs), Piezo (2335AAs), a possible homolog of Centriolin (2141AAs), glycogen synthase (1214AAs), oxysterol binding protein (894AAs), and a putative homolog of SZT2 (3031AAs). Large genes such as dynein heavy chain required manual reannotation in *H. californensis* as well, as only 2/17 dynein genes were correctly annotated in the Iso-Seq-based Stringtie annotation.

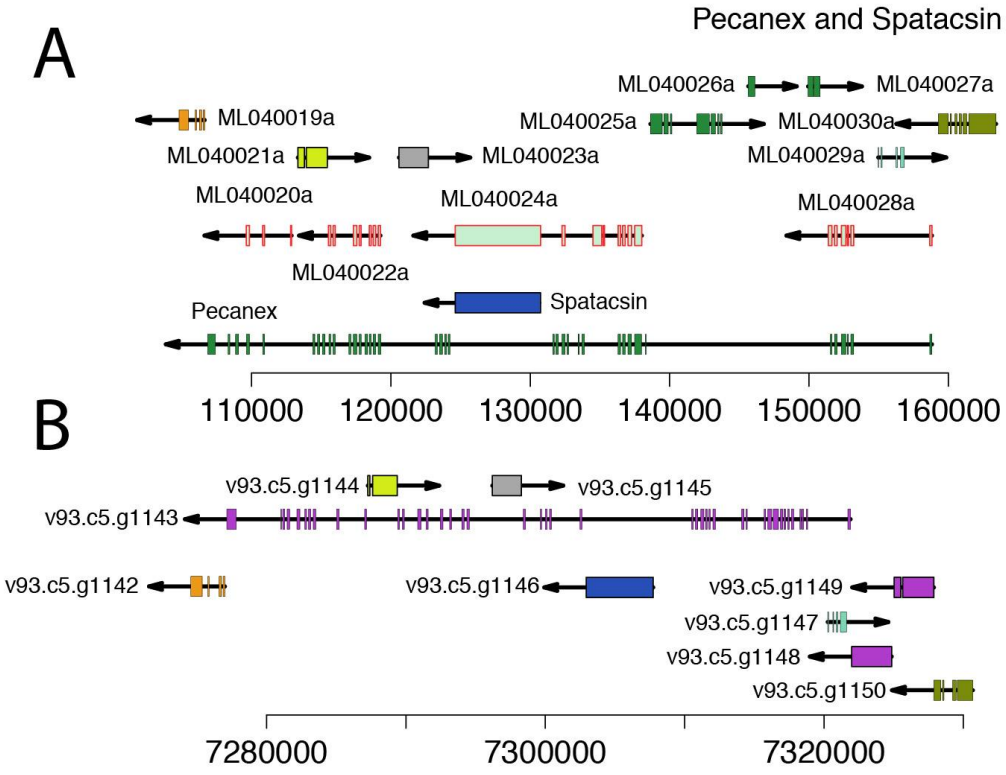

Figure A11 - Pecanex and Spatacsin loci. Loci of the homolog of pecanex in *M. leidyi* (A) and *H. californensis* (B). In *M. leidyi*, the full-length gene joins 4 genes from the ML2 annotation, and contains 6 nested intronic genes, one of which was falsely fused. Four of these genes have homologs in *H. californensis* in the orthologous introns. The gene ML040024a fuses the single-exon homolog of spatacsin, though this is not supported by the transcripts or de novo assembly. Many of the surrounding or nested genes are homologous between the two species, as ML040019a, ML040021a, ML040023a, ML040029a, and ML040030a, match with *H. californensis* c5.g977, c5.g979, c5.g980, c5.g982, and c5.g984, respectively, and are colored pairwise.

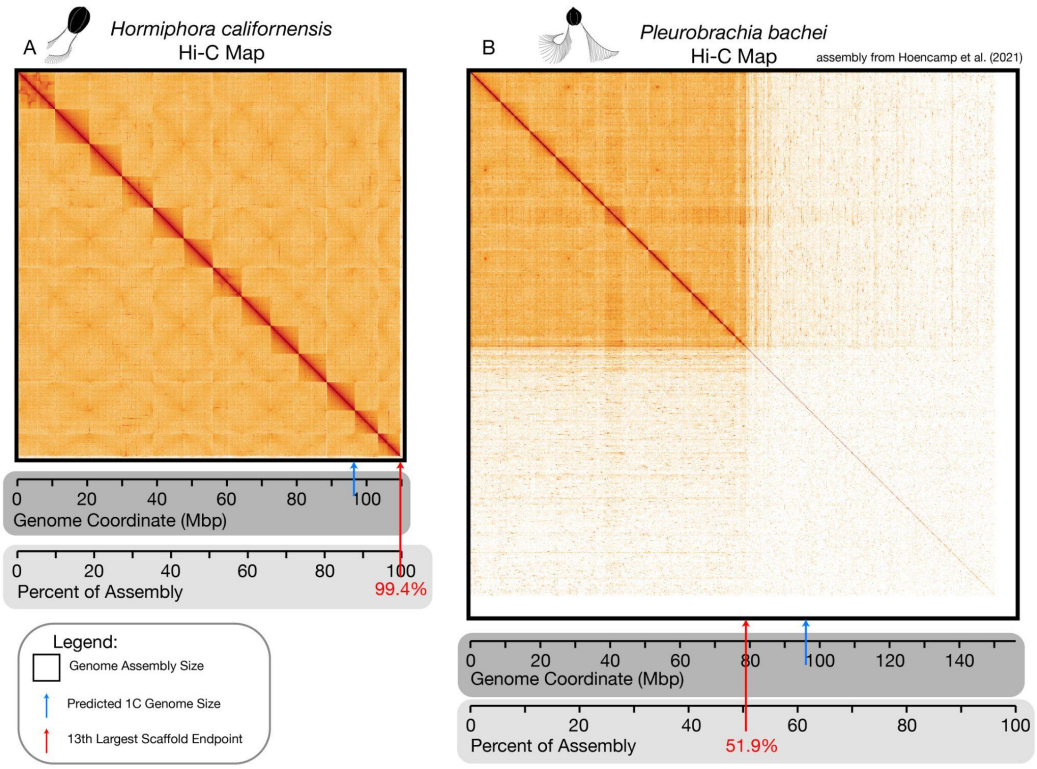

Figure A12 - Hi-C map of *H. californensis* and *P. bachei*. These are the Hi-C maps of *H. californensis* and *P. bachei*, shown without individual lines separating scaffolds. The x-y scale of megabase pairs (Mbp) in both plots is the same. The genome assembly sizes are shown with a black bounding border. The predicted genome size for both species based on k-mer spectra, 96.6 Mbp, is shown with a blue arrow. The amount of the genome in the 13 largest scaffolds is shown with a red arrow, and the percent of the assembly in those 13 scaffolds is shown in red text. (A) The Hi-C map for *H. californensis*. The largest 13 scaffolds contain 99.4% of the total bases in the assembly. (B) The Hi-C map for *P. bachei* from Hoencamp et al (2014). The largest 13 scaffolds contain 51.9% of the assembly. 48.1% of the genome is not in chromosome-scale scaffolds, yet has Hi-C connections to the chromosome-scale scaffolds.

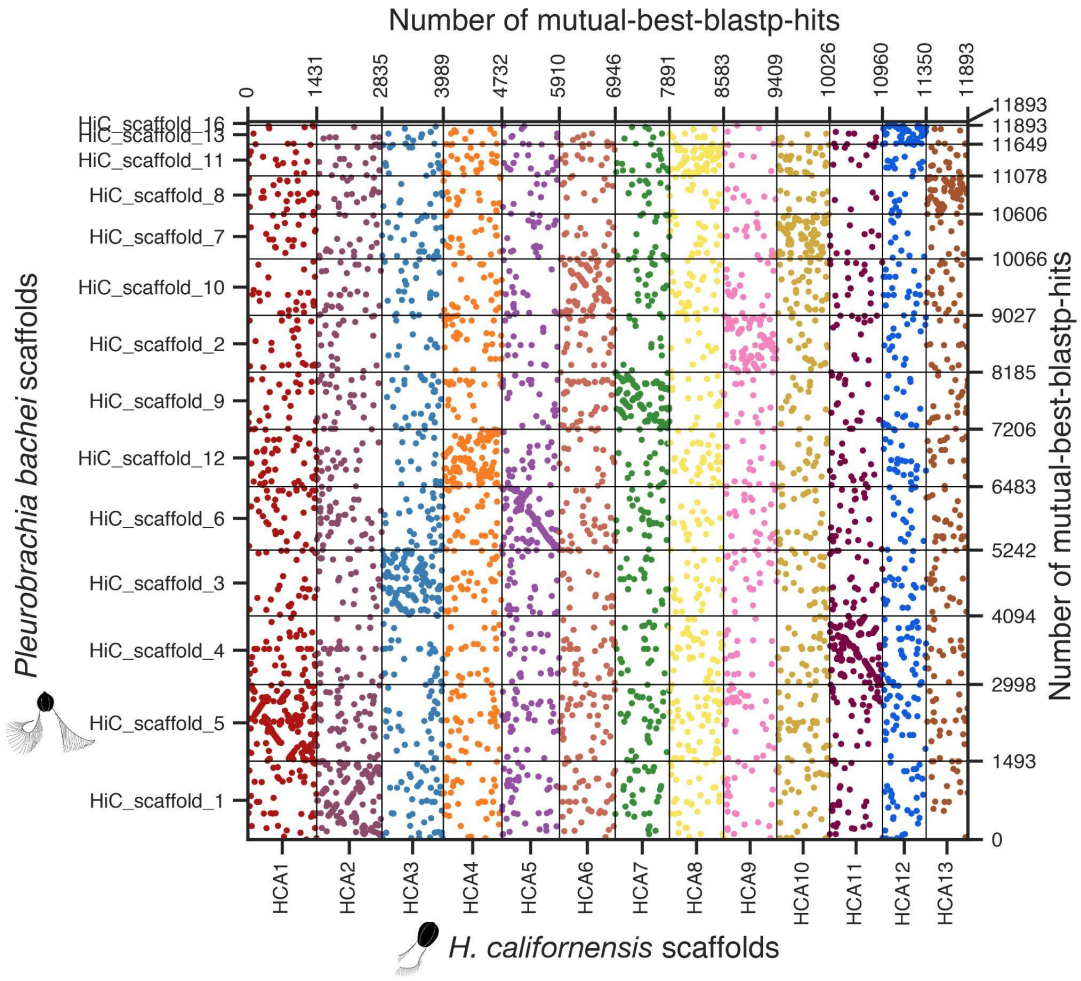

Figure A13 - *Pleurobrachia-Hormiphora* Oxford dot plot. This plot shows the coordinates of mutual best blastp hits when comparing the proteins in the genomes of *P. bachei* to *H. californensis*, and *H. californensis* to *P. bachei*. Only the first 13 *H. californensis* scaffolds, and the largest 14 *P. bachei* scaffolds, are plotted. One dot is one putatively orthologous protein shared by the two species. The dots are colored by *Hormiphora* chromosome. This plot shows that each *H. californensis* chromosomal scaffold has a homologous chromosomal scaffold in *P. bachei*. For example, *H. californensis* 1 predominantly shares genes with *P. bachei* HiC_scaffold_5. Moreover, this plot shows that while shared chromosomes 5 and 11 have large regions with gene

colinearity, most of the other homologous chromosomes are highly rearranged between *H. californensis* and *P. bachei*. Lastly, we see that only the 13 largest *P. bachei* scaffolds have enough information to assign them to homologous *H. californensis* scaffolds. The 14th-largest *P. bachei* scaffold has no proteins that had reciprocal best matches to the 13 chromosomal *H. californensis* scaffolds.

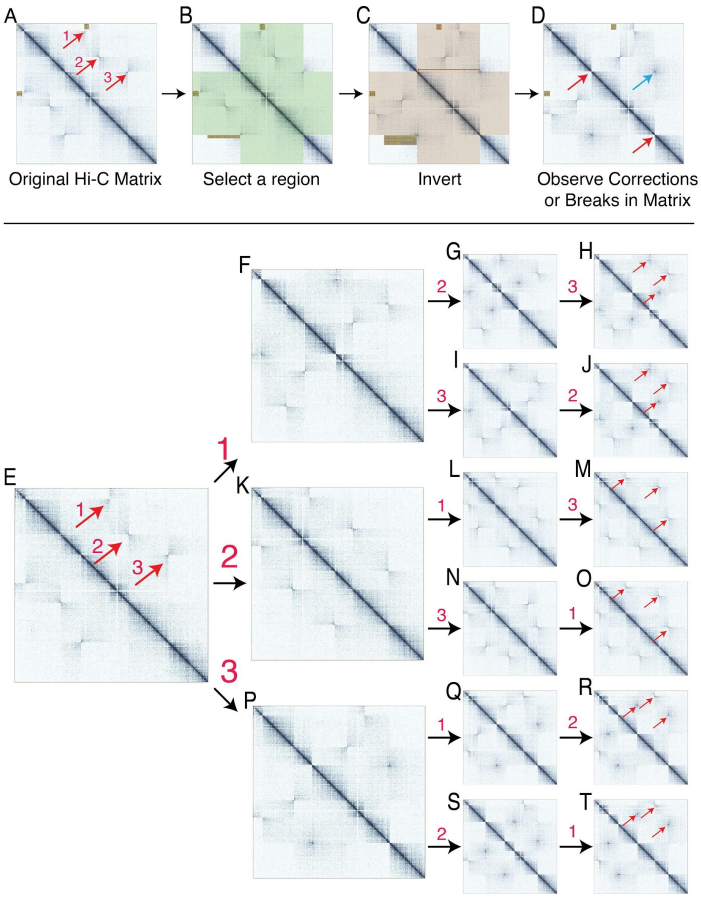

Figure A14. Scaffold 1 heterozygous inversion. Off-diagonal hotspots in Hi-C contact matrices (A, red arrows) indicate assembly errors, or heterozygous inversions. One method of determining if off-diagonal Hi-C hotspots are misassemblies was to manually invert the assembly at the suspect break points (B,C). The manipulation will result in removing the off-diagonal signal while preserving the diagonal signal, or preserving the off-diagonal signal while degrading the diagonal signal (D, red arrows). Panels (E-T) show all the possible combinations of rearrangements to attempt to correct the off-diagonal signal. Red numbers above black arrows indicate which off-diagonal signal was inverted. The right-most panels show that the off-diagonal signals remain after manipulating the heatmaps, and the continuity of the diagonal signal is interrupted. Therefore, this signal is likely from heterogyzous inversions.

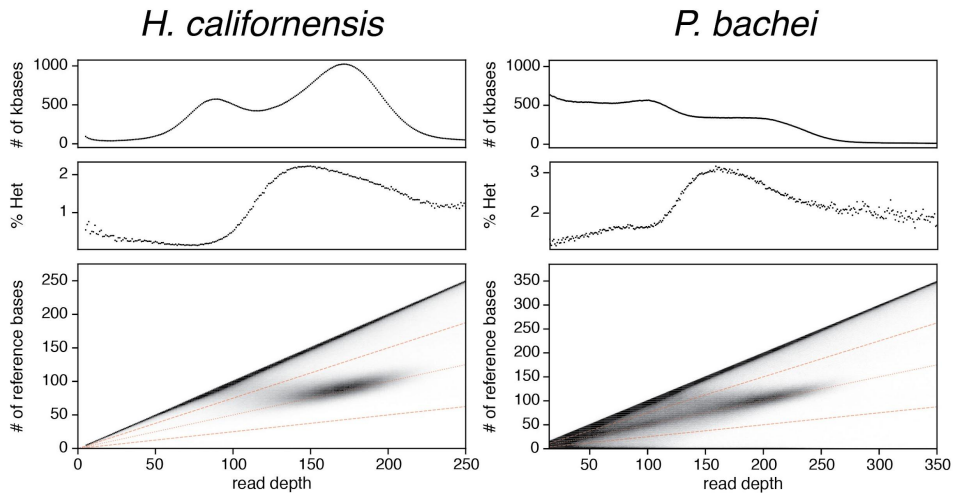

Figure A15 - Plots pertaining to the heterozygosity of *H. californensis* and *P. bachei*. The bottom-most panels show a heat map of the number of positions in the genome that have X-number of reads with the reference allele when the total read depth at that position is Y. A smear at 1x sequencing depth coverage (x-axis) with only 50% of bases matching the reference allele, shows that the animals are diploid. The top-most panel is a histogram of the total number of positions in the genome (Y-axis) that have X number of reads at that position. This plot is useful to visualize the proportion of bases that are either located on uncollapsed haplotigs, or are indels present in the assembly. The middle panel shows the heterozygosity at each read depth. The most reliable window for calculating heterozygosity is at the mode of the mapping depth where reads from both haplotypes map to the reference. This point is approximately 160x read depth for *H. californensis* and 205x read depth for *P. bachei*. The top panel of the *P. bachei* analysis shows that there are many positions in the genome that have reads mapped from only one haplotype, indicated by the peak around 102x read depth.

Individual	Acc. Number	Species	Method	k-mer size	% SNV Het (min)	% SNV Het (max)
SAMN00216730	SAMN00216730	*P. bachei*	mpileup	NA	2.63%	NA
			angsd	NA	2.40%	NA
			vcftools	NA	0	NA
			GenomeScope2	21	4.20%	4.25%
			GenomeScope2	41	3.03%	3.08%
Hc1	SAMN12924379	*H. californensis*	mpileup	NA	2.00%	NA
			angsd	NA	1.65%	NA
			vcftools	NA	1.51%	NA
			GenomeScope2	21	2.95%	2.98%
			GenomeScope2	41	2.36%	2.39%
Hc2	SAMN12924380	*H. californensis*	angsd	NA	1.85%	NA
			vcftools	NA	1.56%	NA
			GenomeScope2	21	3.25%	3.28%
			GenomeScope2	41	2.55%	2.58%

Table A4 - Estimated heterozygosity of *H. californensis* and *P. bachei*. We measured the heterozygosity of *P. bachei* SAMN00216730 and *H. californensis* Hc1 using the mpileup method (Saremi *et al.* 2019). In this table, the mpileup method only measures the single-nucleotide heterozygosity. In addition we measured the heterozygosity using angsd, vcftools, and GenomeScope2 (Danecek *et al.* 2011; Korneliussen *et al.* 2014; Ranallo-Benavidez *et al.* 2019). The k-mer size used and the window of heterozygosity values were reported for the GenomeScope method. Vcftools reported zero heterozygous sites for the *P. bachei* individual, which we attribute to a software error given the results of the mpileup and angsd analyses.

Species	Genome accession used	SRA accessions used
T. wilhelma	(Mills *et al.* 2018)	SRR2163223
T. adhaerens	GCF_000150275.1	SRX6204530 through SRX6204554
N. nomurai	GCA_003864495.1	SRR6298213
D. melanogaster	GCF_000001215.4	SRR10512945
S. purpuratus	GCF_000002235.5	SRR7211988
H. sapiens	GRCh38	(Zook *et al.* 2016)

Table A5 - Genome samples used in heterozygosity measurements. These genome assemblies and SRAs were used as a comparison for genome heterozygosity measurements compared to *H. californensis*.

Bibliography

Abdennur, N., and L. A. Mirny, 2020 Cooler: scalable storage for Hi-C data and other genomically labeled arrays. Bioinformatics 36: 311–316.

Adams, M., J. McBroome, N. Maurer, E. Pepper-Tunick, N. F. Saremi *et al.*, 2020 One fly-one genome: chromosome-scale genome assembly of a single outbred Drosophila melanogaster. Nucleic Acids Res. 48: e75.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Andrade, S. C. S., M. Novo, G. Y. Kawauchi, K. Worsaae, F. Pleijel *et al.*, 2015 Articulating "Archiannelids": Phylogenomics and Annelid Relationships, with Emphasis on Meiofaunal Taxa. Mol. Biol. Evol. 32: 2860–2875.

Antcliffe, J. B., R. H. T. Callow, and M. D. Brasier, 2014 Giving the early fossil record of sponges a squeeze. Biol. Rev. Camb. Philos. Soc. 89: 972–1004.

Arafat, H., A. Alamaru, C. Gissi, and D. Huchon, 2018 Extensive mitochondrial gene rearrangements in Ctenophora: insights from benthic Platyctenida. BMC Evol. Biol. 18: 65.

Aristotle, 350 BC *Aristotle: On the Parts of Animals*. Clarendon Press.

Bassot, J. M., and M. T. Nicolas, 1978 Similar paracrystals of endoplasmic reticulum in the photoemitters and the photoreceptors of scale-worms. Experientia 34: 726–728.

Belinky, F., C. Rot, M. Ilan, and D. Huchon, 2008 The complete mitochondrial genome of the demosponge Negombata magnifica (Poecilosclerida). Mol. Phylogenet. Evol. 47: 1238–1243.

Bolger, A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.

Boroni, M., M. Sammeth, S. G. Gava, N. A. N. Jorge, A. M. Macedo *et al.*, 2018 Landscape of the spliced leader trans-splicing mechanism in Schistosoma mansoni. Sci. Rep. 8: 3877.

Bråte, J., R. S. Neumann, B. Fromm, A. A. B. Haraldsen, J. E. Tarver *et al.*, 2018 Unicellular Origin of the Animal MicroRNA Machinery. Curr. Biol. 28: 3288–3295.e5.

Buchfink, B., C. Xie, and D. H. Huson, 2015 Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12: 59–60.

Burton, J. N., A. Adey, R. P. Patwardhan, R. Qiu, J. O. Kitzman *et al.*, 2013 Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat. Biotechnol. 31: 1119–1125.

Bushnell, B., J. Rood, and E. Singer, 2017 BBMerge - Accurate paired shotgun read merging via overlap. PLoS One 12: e0185056.

Cabanettes, F., and C. Klopp, 2018 D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ 6: e4958.

Chapman, J. A., I. Ho, S. Sunkara, S. Luo, G. P. Schroth *et al.*, 2011 Meraculous: De Novo Genome Assembly with Short Paired-End Reads (S. L. Salzberg, Ed.). PLoS One 6: e23501.

Chida, A. R., S. Ravi, S. Jayaprasad, K. Paul, J. Saha *et al.*, 2020 A Near-Chromosome Level Genome Assembly of Anopheles stephensi. Front. Genet. 11: 565626.

Clark, H. J. A. J. of S. A. A., 1866 ART. XLIII.--Conclusive proofs of the animality of the ciliate Sponges, and of their affinities with the Infusoria Flagellata; New Haven 42: 1820.

Corbett-Detig, R. B., I. Said, M. Calzetta, M. Genetti, J. McBroome *et al.*, 2019 Fine-Mapping Complex Inversion Breakpoints and Investigating Somatic Pairing in the Anopheles gambiae Species Complex Using Proximity-Ligation Sequencing. Genetics 213: 1495–1511.

Dahlgren, U., 1916 The production of light by animals. J. Franklin Inst. 181: 243–261.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.*, 2011 The variant call format and VCFtools. Bioinformatics 27: 2156–2158.

Dawson, M. N., K. A. Raskoff, and D. K. Jacobs, 1998 Field preservation of marine invertebrate tissue for DNA analyses. Mol. Mar. Biol. Biotechnol. 7: 145–152.

De Coster, W., S. D'Hert, D. T. Schultz, M. Cruts, and C. Van Broeckhoven, 2018 NanoPack: visualizing and processing long read sequencing data. Bioinformatics 34: 2666–2669.

Deheyn, D. D., and M. I. Latz, 2009 Internal and secreted bioluminescence of the marine polychaete Odontosyllis phosphorea (Syllidae). Invertebr. Biol. 128: 31–45.

Derelle, R., T. Momose, M. Manuel, C. Da Silva, P. Wincker *et al.*, 2010 Convergent origins and rapid evolution of spliced leader trans-splicing in metazoa: insights from the ctenophora and hydrozoa. RNA 16: 696–707.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski *et al.*, 2013 STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21.

Dohrmann, M., and G. Wörheide, 2017 Dating early animal evolution using phylogenomic data. Sci. Rep. 7: 3599.

Dunn, C. W., A. Hejnol, D. Q. Matus, K. Pang, W. E. Browne *et al.*, 2008 Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452: 745–749.

Edgar, R. C., 2004 MUSCLE: multiple sequence alignment with high accuracy and high

throughput. Nucleic Acids Res. 32: 1792–1797.

Edge, P., V. Bafna, and V. Bansal, 2017 HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Res. 27: 801–812.

Ellis, E. A., and T. H. Oakley, 2016 High Rates of Species Accumulation in Animals with Bioluminescent Courtship Displays. Curr. Biol. 26: 1916–1921.

Fairclough, S. R., Z. Chen, E. Kramer, Q. Zeng, S. Young *et al.*, 2013 Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate Salpingoeca rosetta. Genome Biol. 14: R15.

Fernández, R., and T. Gabaldón, 2020 Gene gain and loss across the metazoan tree of life. Nat Ecol Evol 4: 524–533.

Fischer, A., and U. Fischer, 1995 On the Life-Style and Life-Cycle of the Luminescent Polychaete Odontosyllis enopla (Annelida: Polychaeta). Invertebr. Biol. 114: 236–247.

Francis, W. R., L. M. Christianson, R. Kiko, M. L. Powers, N. C. Shaner *et al.*, 2013 A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly. BMC Genomics 14: 167.

Francis, W. R., N. C. Shaner, L. M. Christianson, M. L. Powers, and S. H. D. Haddock, 2015 Occurrence of Isopenicillin-N-Synthase Homologs in Bioluminescent Ctenophores and Implications for Coelenterazine Biosynthesis. PLoS One 10: e0128742.

Freeman, G., 1977 The establishment of the oral-aboral axis in the ctenophore embryo. Development 42: 237–260.

Gaiti, F., A. D. Calcino, M. Tanurdžić, and B. M. Degnan, 2017 Origin and evolution of the metazoan non-coding regulatory genome. Dev. Biol. 427: 193–202.

Garrison, E., and G. Marth, 2012 Haplotype-based variant detection from short-read sequencing. arXiv [q-bio.GN].

Gaston, G. R., and J. Hall, 2000 Lunar periodicity and bioluminescence of swarming Odontosyllis luminosa (Polychaeta: Syllidae) in Belize. Gulf Caribb. Res. 12: 47–51.

Gouveneaux, A., and J. Mallefet, 2013 Physiological control of bioluminescence in a deep-sea planktonic worm, Tomopteris helgolandica. J. Exp. Biol. 216: 4285–4289.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson *et al.*, 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29: 644–652.

Guo, L., A. Accorsi, S. He, C. Guerrero-Hernández, S. Sivagnanam *et al.*, 2018 An adaptable chromosome preparation methodology for use in invertebrate research organisms. BMC Biol. 16: 25.

Haddock, S. H. D., M. A. Moline, and J. F. Case, 2010 Bioluminescence in the sea. Ann. Rev. Mar. Sci. 2: 443–493.

Haddock, S. H., T. J. Rivers, and B. H. Robison, 2001 Can coelenterates make coelenterazine? Dietary requirement for luciferin in cnidarian bioluminescence. Proc. Natl. Acad. Sci. U. S. A. 98: 11148–11151.

Hahn, C., L. Bachmann, and B. Chevreux, 2013 Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads--a baiting and iterative mapping approach. Nucleic Acids Res. 41: e129.

Harris, R. S., 2007 Improved pairwise alignment of genomic DNA: The Pennsylvania State University.

Harvey, E. N., 1921 STUDIES ON BIOLUMINESCENCE. Biol. Bull. 41: 280–287.

Heinz, S., L. Texari, M. G. B. Hayes, M. Urbanowski, M. W. Chang *et al.*, 2018 Transcription Elongation Can Affect Genome 3D Structure. Cell 174: 1522–1536.e22.

Henikoff, S., M. A. Keene, K. Fechtel, and J. W. Fristrom, 1986 Gene within a gene: nested

Drosophila genes encode unrelated proteins on opposite DNA strands. Cell 44: 33–42.

Hentschel, U., K. M. Usher, and M. W. Taylor, 2006 Marine sponges as microbial fermenters. FEMS Microbiol. Ecol. 55: 167–177.

Hernandez-Nicaise, M. L., 1973 The nervous system of ctenophores. III. Ultrastructure of synapses. J. Neurocytol. 2: 249–263.

Herring, P. J., 1987 Systematic distribution of bioluminescence in living organisms. J. Biolumin. Chemilumin. 1: 147–163.

Hestetun, J. T., J. Vacelet, N. Boury-Esnault, C. Borchiellini, M. Kelly *et al.*, 2016 The systematics of carnivorous sponges. Mol. Phylogenet. Evol. 94: 327–345.

Hoencamp, C., O. Dudchenko, A. M. O. Elbatsh, S. Brahmachari, J. A. Raaijmakers *et al.*, 2021 3D genomics across the tree of life reveals condensin II as a determinant of architecture type. Science 372: 984–989.

Hoff, K. J., S. Lange, A. Lomsadze, M. Borodovsky, and M. Stanke, 2016 BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32: 767–769.

Hoff, K. J., A. Lomsadze, M. Borodovsky, and M. Stanke, 2019 Whole-Genome Annotation with BRAKER. Methods Mol. Biol. 1962: 65–95.

Hou, C., L. Li, Z. S. Qin, and V. G. Corces, 2012 Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. Mol. Cell 48: 471–484.

Huelsenbeck, J. P., and F. Ronquist, 2001 MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754–755.

Huntsman, A. G., 1948 Odontosyllis at Bermuda and Lunar Periodicity. J. Fish. Res. Bd. Can. 7b: 363–369.

Huson, D. H., A. F. Auch, J. Qi, and S. C. Schuster, 2007 MEGAN analysis of metagenomic data. Genome Res. 17: 377–386.

Kajitani, R., K. Toshimoto, H. Noguchi, A. Toyoda, Y. Ogura *et al.*, 2014 Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 24: 1384–1395.

Kaskova, Z. M., A. S. Tsarkova, and I. V. Yampolsky, 2016 1001 lights: luciferins, luciferases, their mechanisms of action and applications in chemical analysis, biology and medicine. Chem. Soc. Rev. 45: 6048–6077.

Katoh, K., K. Misawa, K.-I. Kuma, and T. Miyata, 2002 MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30: 3059–3066.

Kenny, N. J., W. R. Francis, R. E. Rivera-Vicéns, K. Juravel, A. de Mendoza *et al.*, 2020 Tracing animal genomic evolution with the chromosomal-level assembly of the freshwater sponge Ephydatia muelleri. Nat. Commun. 11: 3676.

Kent, W. S., 1880 *A Manual of the Infusoria: Including a Description of All Known Flagellate, Ciliate, and Tentaculiferous Protozoa, British and Foreign, and an Account of the Organization and the Affinities of the Sponges*. David Bogue.

Kerpedjiev, P., N. Abdennur, F. Lekschas, C. McCallum, K. Dinkla *et al.*, 2018 HiGlass: web-based visual exploration and analysis of genome interaction maps. Genome Biol. 19: 125.

Kirkpatrick, M., 2010 How and why chromosome inversions evolve. PLoS Biol. 8.:

Kohn, A. B., M. R. Citarella, K. M. Kocot, Y. V. Bobkova, K. M. Halanych *et al.*, 2012 Rapid evolution of the compact and unusual mitochondrial genome in the ctenophore, Pleurobrachia bachei. Mol. Phylogenet. Evol. 63: 203–207.

Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman *et al.*, 2017 Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27: 722–736.

Korneliussen, T. S., A. Albrechtsen, and R. Nielsen, 2014 ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics 15: 356.

Krogh, A., B. Larsson, G. von Heijne, and E. L. Sonnhammer, 2001 Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305: 567–580.

Laetsch, D. R., and M. L. Blaxter, 2017 BlobTools: Interrogation of genome assemblies. F1000Res. 6: 1287.

Lang, B. F., C. O'Kelly, T. Nerad, M. W. Gray, and G. Burger, 2002 The closest unicellular relatives of animals. Curr. Biol. 12: 1773–1778.

Laslett, D., and B. Canback, 2008 ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. Bioinformatics 24: 172–175.

Laumer, C. E., R. Fernández, S. Lemer, D. Combosch, K. M. Kocot *et al.*, 2019 Revisiting metazoan phylogeny with genomic sampling of all phyla. Proc. Biol. Sci. 286: 20190831.

Lau, E. S., and T. H. Oakley, 2021 Multi‐level convergence of complex traits and the evolution of bioluminescence. Biol. Rev. Camb. Philos. Soc.

Leadbeater, B. S. C., 1983 Life-history and ultrastructure of a new marine species of Proterospongia (Choanoflagellida). J. Mar. Biol. Assoc. U. K. 63: 135–160.

Leffler, E. M., K. Bullaughey, D. R. Matute, W. K. Meyer, L. Ségurel *et al.*, 2012 Revisiting an old riddle: what determines genetic diversity levels within species? PLoS Biol. 10: e1001388.

Lemer, S., G. Y. Kawauchi, S. C. S. Andrade, V. L. González, M. J. Boyle *et al.*, 2015 Re-evaluating the phylogeny of Sipuncula through transcriptomics. Mol. Phylogenet. Evol. 83: 174–183.

Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv.

Li, H., 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987–2993.

Li, H., 2020 auN: a new metric to measure assembly contiguity.

Li, H., 2017 Minimap2: pairwise alignment for nucleotide sequences. arXiv.

Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy *et al.*, 2009 Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326: 289–293.

Li, Y., L. Gao, Y. Pan, M. Tian, Y. Li *et al.*, 2020 Chromosome-level reference genome of the jellyfish Rhopilema esculentum. Gigascience 9.:

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

Li, Y., K. M. Kocot, N. V. Whelan, S. R. Santos, D. S. Waits *et al.*, 2017 Phylogenomics of tubeworms (Siboglinidae, Annelida) and comparative performance of different reconstruction methods. Zool. Scr. 46: 200–213.

Li, Y., X.-X. Shen, B. Evans, C. W. Dunn, and A. Rokas, 2021 Rooting the animal tree of life. Mol. Biol. Evol.

Lloyd, J. E., 1965 Aggressive Mimicry in Photuris: Firefly Femmes Fatales. Science 149: 653–654.

Loman, N. J., and A. R. Quinlan, 2014 Poretools: a toolkit for analyzing nanopore sequence data. Bioinformatics 30: 3399–3401.

Lomsadze, A., P. D. Burns, and M. Borodovsky, 2014 Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res. 42: e119.

Lowe, T. M., and S. R. Eddy, 1997 tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25: 955–964.

Lv, J., P. Havlak, and N. H. Putnam, 2011 Constraints on genes shape long-term conservation of macro-synteny in metazoan genomes. BMC Bioinformatics 12 Suppl 9: S11.

Marçais, G., and C. Kingsford, 2011 A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27: 764–770.

Martini, S., D. T. Schultz, L. Lundsten, and S. H. D. Haddock, 2020 Bioluminescence in an Undescribed Species of Carnivorous Sponge (Cladorhizidae) From the Deep Sea. Frontiers in Marine Science 7: 1041.

Matthews, B. J., and L. B. Vosshall, 2020 How to turn an organism into a model organism in 10 “easy” steps. J. Exp. Biol. 223.:

McArthur, E., and J. A. Capra, 2021 Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. Am. J. Hum. Genet. 108: 269–283.

McBroome, J., D. Liang, and R. Corbett-Detig, 2020 Fine-Scale Position Effects Shape the Distribution of Inversion Breakpoints in Drosophila melanogaster. Genome Biol. Evol. 12: 1378–1391.

de Mendoza, A., H. Suga, J. Permanyer, M. Irimia, and I. Ruiz-Trillo, 2015 Complex transcriptional regulation and independent evolution of fungal-like traits in a relative of

animals. Elife 4: e08904.

Meyer, M., and M. Kircher, 2010 Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc. 2010: db.prot5448.

Mills, D. B., W. R. Francis, S. Vargas, M. Larsen, C. P. Elemans *et al.*, 2018 The last common ancestor of animals lacked the HIF pathway and respired in low-oxygen environments. Elife 7.:

Miron, M.-J., L. LaRivière, J.-M. Bassot, and M. Anctil, 1987 Immunohistochemical and radioautographic evidence of monoamine-containing cells in bioluminescent elytra of the scale-worm Harmothoe imbricata (Polychaeta). Cell Tissue Res. 249: 547–556.

Morin, J. G., 2019 Luminaries of the reef: The history of luminescent ostracods and their courtship displays in the Caribbean. J. Crustacean Biol. 39: 227–243.

Moroz, L. L., 2015 Convergent evolution of neural systems in ctenophores. J. Exp. Biol. 218: 598–611.

Moroz, L. L., K. M. Kocot, M. R. Citarella, S. Dosung, T. P. Norekian *et al.*, 2014 The ctenophore genome and the evolutionary origins of neural systems. Nature 510: 109–114.

Nishimura, O., Y. Hara, and S. Kuraku, 2017 gVolante for standardizing completeness assessment of genome and transcriptome assemblies. Bioinformatics 33: 3635–3637.

Nong, W., J. Cao, Y. Li, Z. Qu, J. Sun *et al.*, 2020 Jellyfish genomes reveal distinct homeobox gene clusters and conservation of small RNA processing. Nat. Commun. 11: 3051.

Oba, Y., and D. T. Schultz, 2014 Eco-evo bioluminescence on land and in the sea. Adv. Biochem. Eng. Biotechnol. 144: 3–36.

Okada, Y. K., 1925 LUMINESCENCE IN SPONGES. Science 62: 566–567.

Ou, S., W. Su, Y. Liao, K. Chougule, J. R. A. Agda *et al.*, 2019 Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 20: 275.

Pagenstecher, H. A., 1881 *Allgemeine zoologie, oder Grundgesetze des thierischen baus und lebens*. Wiegandt, Hempel & Parey.

Paps, J., and I. Ruiz-Trillo, 2010 Animals and Their Unicellular Ancestors, in Encyclopedia of Life Sciences (Wiley Online Library).

Patry, W. L., M. K. Bubel, C. Hansen, and T. Knowles, 2019 Diffusion tubes: a method for the mass culture of ctenophores and other pelagic marine invertebrates. bioRxiv 751099.

Patterson, M., T. Marschall, N. Pisanti, L. van Iersel, L. Stougie *et al.*, 2015 WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. J. Comput. Biol. 22: 498–509.

Pertea, M., G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell *et al.*, 2015 StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33: 290–295.

Pett, W., and D. V. Lavrov, 2015 Cytonuclear Interactions in the Evolution of Animal Mitochondrial tRNA Metabolism. Genome Biol. Evol. 7: 2089–2101.

Petushkov, V. N., M. A. Dubinnyi, A. S. Tsarkova, N. S. Rodionova, M. S. Baranov *et al.*, 2014 A novel type of luciferin from the Siberian luminous EarthwormFridericia heliota: Structure elucidation by spectral studies and total synthesis. Angew. Chem. Weinheim Bergstr. Ger. 126: 5672–5674.

Philippe, H., R. Derelle, P. Lopez, K. Pick, C. Borchiellini *et al.*, 2009 Phylogenomics revives traditional views on deep animal relationships. Curr. Biol. 19: 706–712.

Picard Toolkit, 2016 Broad institute, GitHub repository.

Picelli, S., O. R. Faridani, Å. K. Björklund, G. Winberg, S. Sagasser *et al.*, 2014 Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. 9: 171–181.

Presnell, J. S., and W. E. Browne, 2019 Krüppel-like factor gene function in the ctenophore *Mnemiopsis* suggests an ancient role in promoting cell proliferation in metazoan stem cell niches. bioRxiv 527002.

Putnam, N. H., T. Butts, D. E. K. Ferrier, R. F. Furlong, U. Hellsten *et al.*, 2008 The amphioxus genome and the evolution of the chordate karyotype. Nature 453: 1064–1071.

Putnam, N. H., B. L. O'Connell, J. C. Stites, B. J. Rice, M. Blanchette *et al.*, 2016 Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 26: 342–350.

Ramírez, F., V. Bhardwaj, L. Arrigoni, K. C. Lam, B. A. Grüning *et al.*, 2018 High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat. Commun. 9: 189.

Ranallo-Benavidez, T. R., K. S. Jaron, and M. C. Schatz, 2020 GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat. Commun. 11: 1432.

Ranallo-Benavidez, T. R., K. S. Jaron, and M. C. Schatz, 2019 GenomeScope 2.0 and Smudgeplots: Reference-free profiling of polyploid genomes. BioRxiv.

Rice, E. S., and R. E. Green, 2019 New Approaches for Genome Assembly and Scaffolding. Annu Rev Anim Biosci 7: 17–40.

Rio, D. C., M. Ares Jr, G. J. Hannon, and T. W. Nilsen, 2010 Purification of RNA using TRIzol (TRI reagent). Cold Spring Harb. Protoc. 2010: db.prot5439.

Roach, M. J., S. A. Schmidt, and A. R. Borneman, 2018 Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics 19: 460.

Robinson, J. T., H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander *et al.*, 2011 Integrative genomics viewer. Nat. Biotechnol. 29: 24.

Ronquist, F., and J. P. Huelsenbeck, 2003 MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574.

Ruan, J., and H. Li, 2019 Fast and accurate long-read assembly with wtdbg2. bioRxiv 530972.

Ryan, J. F., and M. Chiodin, 2015 Where is my mind? How sponges and placozoans may have lost neural cell types. Philos. Trans. R. Soc. Lond. B Biol. Sci. 370: 20150059–.

Ryan, J. F., K. Pang, C. E. Schnitzler, A.-D. Nguyen, R. T. Moreland *et al.*, 2013 The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. Science 342: 1242592.

Sacerdot, C., A. Louis, C. Bon, C. Berthelot, and H. Roest Crollius, 2018 Chromosome evolution at the origin of the ancestral vertebrate genome. Genome Biol. 19: 166.

Sambrook, J., and D. W. Russell, 2006 Purification of nucleic acids by extraction with phenol:chloroform. CSH Protoc. 2006.:

Saremi, N. F., M. A. Supple, A. Byrne, J. A. Cahill, L. L. Coutinho *et al.*, 2019 Puma genomes from North and South America provide insights into the genomic consequences of inbreeding. Nat. Commun. 10: 4769.

Schlining, B. M., and N. J. Stout, 2006 MBARI's Video Annotation and Reference System, pp. 1–5 in *OCEANS 2006*, ieeexplore.ieee.org.

Schultz, D. T., W. R. Francis, J. D. McBroome, L. M. Christianson, S. H. D. Haddock *et al.*, 2021 A chromosome-scale genome assembly and karyotype of the ctenophore Hormiphora californensis. G3 Genes|Genomes|Genetics.

Schultz, D. T., A. A. Kotlobay, R. Ziganshin, A. Bannikov, N. M. Markina *et al.*, 2018

Luciferase of the Japanese syllid polychaete Odontosyllis umdecimdonta. Biochem. Biophys. Res. Commun. 502: 318–323.

Sebé-Pedrós, A., C. Ballaré, H. Parra-Acero, C. Chiva, J. J. Tena *et al.*, 2016 The Dynamic Regulatory Genome of Capsaspora and the Origin of Animal Multicellularity. Cell 165: 1224–1237.

Sebé-Pedrós, A., E. Chomsky, K. Pang, D. Lara-Astiaso, F. Gaiti *et al.*, 2018 Early metazoan cell type diversity and the evolution of multicellular gene regulation. Nat Ecol Evol 2: 1176–1188.

Sebé-Pedrós, A., B. M. Degnan, and I. Ruiz-Trillo, 2017 The origin of Metazoa: a unicellular perspective. Nat. Rev. Genet. 18: 498–512.

Sebé-Pedrós, A., A. de Mendoza, B. F. Lang, B. M. Degnan, and I. Ruiz-Trillo, 2011 Unexpected repertoire of metazoan transcription factors in the unicellular holozoan Capsaspora owczarzaki. Mol. Biol. Evol. 28: 1241–1254.

Shen, X.-X., C. T. Hittinger, A. Rokas, B. Q. Minh, and E. L. Braun, 2017 Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nature Ecology & Evolution 1: 0126.

Shimomura, O., 2006 *Bioluminescence: Chemical Principles and Methods*. World Scientific.

Shimomura, O., J. R. Beers, and F. H. Johnson, 1964 The cyanide activation of Odontosyllis luminescence. J. Cell. Comp. Physiol. 64: 15–21.

Shimomura, O., F. H. Johnson, and Y. Saiga, 1963 Partial purification and properties of the Odontosyllis luminescence system. J. Cell. Comp. Physiol. 61: 275–292.

Simakov, O., J. Bredeson, K. Berkoff, F. Marletaz, T. Mitros *et al.* Genome tectonics and the evolution of metazoan chromosomes. Submitted.

Simakov, O., F. Marlétaz, J.-X. Yue, B. O'Connell, J. Jenkins *et al.*, 2020 Deeply conserved

synteny resolves early events in vertebrate evolution. Nat Ecol Evol.

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

Simion, P., H. Philippe, D. Baurain, M. Jager, D. J. Richter *et al.*, 2017 A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. Curr. Biol. 27: 958–967.

Simister, R. L., S. Schmitt, and M. W. Taylor, 2011 Evaluating methods for the preservation and extraction of DNA and RNA for analysis of microbial communities in marine sponges. J. Exp. Mar. Bio. Ecol. 397: 38–43.

Skinner, M. E., A. V. Uzilov, L. D. Stein, C. J. Mungall, and I. H. Holmes, 2009 JBrowse: a next-generation genome browser. Genome Res. 19: 1630–1638.

Srivastava, M., E. Begovic, J. Chapman, N. H. Putnam, U. Hellsten *et al.*, 2008 The Trichoplax genome and the nature of placozoans. Nature 454: 955–960.

Srivastava, M., O. Simakov, J. Chapman, B. Fahey, M. E. A. Gauthier *et al.*, 2010 The Amphimedon queenslandica genome and the evolution of animal complexity. Nature 466: 720–726.

Stamatakis, A., 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 1312–1313.

Stanke, M., R. Steinkamp, S. Waack, and B. Morgenstern, 2004 AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 32: W309–12.

Suga, H., Z. Chen, A. de Mendoza, A. Sebé-Pedrós, M. W. Brown *et al.*, 2013 The Capsaspora genome reveals a complex unicellular prehistory of animals. Nat. Commun. 4: 2325.

Tange, O., and Others, 2011 Gnu parallel-the command-line power tool. The USENIX Magazine 36: 42–47.

Tikhonenkov, D. V., E. Hehenberger, A. S. Esaulov, O. I. Belyakova, Y. A. Mazei *et al.*, 2020 Insights into the origin of metazoan multicellularity from predatory unicellular relatives of animals. BMC Biol. 18: 39.

Torruella, G., A. de Mendoza, X. Grau-Bové, M. Antó, M. A. Chaplin *et al.*, 2015 Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi. Curr. Biol. 25: 2404–2410.

Trainor, G. L., 1979 STUDIES ON THE ODONTOSYLLIS BIOLUMINESCENCE SYSTEM [PhD]: Harvard University.

Tsuji, F. I., and E. Hill, 1983 REPETITIVE CYCLES OF BIOLUMINESCENCE AND SPAWNING IN THE POLYCHAETE, ODONTOSYLLIS PHOSPHOREA. Biol. Bull. 165: 444–449.

Verdes, A., and D. F. Gruber, 2017 Glowing Worms: Biological, Chemical, and Functional Diversity of Bioluminescent Annelids. Integr. Comp. Biol. 57: 18–32.

Wainright, P. O., G. Hinkle, M. L. Sogin, and S. K. Stickel, 1993 Monophyletic origins of the metazoa: an evolutionary link with fungi. Science 260: 340–342.

Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel *et al.*, 2014 Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement (J. Wang, Ed.). PLoS One 9: e112963.

Wang, X., X. Fan, H. C. Schröder, and W. E. G. Müller, 2012 Flashing light in sponges through their siliceous fiber network: A new strategy of "neuronal transmission" in animals. Chin. Sci. Bull. 57: 3300–3311.

Wang, Y., F. Hammes, M. Düggelin, and T. Egli, 2008 Influence of size, shape, and flexibility

on bacterial passage through micropore membrane filters. Environ. Sci. Technol. 42: 6749–6754.

Wang, J., L. Zhang, S. Lian, Z. Qin, X. Zhu *et al.*, 2020 Evolutionary transcriptomics of metazoan biphasic life cycle supports a single intercalation origin of metazoan larvae. Nat Ecol Evol 4: 725–736.

Weigert, A., C. Helm, M. Meyer, B. Nickel, D. Arendt *et al.*, 2014 Illuminating the base of the annelid tree using transcriptomics. Mol. Biol. Evol. 31: 1391–1401.

Weisenfeld, N. I., V. Kumar, P. Shah, D. M. Church, and D. B. Jaffe, 2017 Direct determination of diploid genome sequences. Genome Res. 27: 757–767.

Whelan, N. V., K. M. Kocot, L. L. Moroz, and K. M. Halanych, 2015 Error, signal, and the placement of Ctenophora sister to all other animals. Proc. Natl. Acad. Sci. U. S. A. 112: 5773–5778.

Whelan, N. V., K. M. Kocot, T. P. Moroz, K. Mukherjee, P. Williams *et al.*, 2017 Ctenophore relationships and their placement as the sister group to all other animals. Nat Ecol Evol 1: 1737–1746.

Wiens, M., X. Wang, A. Unger, H. C. Schröder, V. A. Grebenjuk *et al.*, 2010 Flashing light signaling circuit in sponges: endogenous light generation after tissue ablation in Suberites domuncula. J. Cell. Biochem. 111: 1377–1389.

Wilkens, L. A., and J. J. Wolken, 1981 Electroretinograms from Odontosyllis enopla (polychaeta; syllidae): Initial observations on the visual system of the bioluminescent fireworm of Bermuda. Mar. Behav. Physiol. 8: 55–66.

Xu, G.-C., T.-J. Xu, R. Zhu, Y. Zhang, S.-Q. Li *et al.*, 2019 LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. Gigascience 8.:

Yang, J., R. Yan, A. Roy, D. Xu, J. Poisson *et al.*, 2015 The I-TASSER Suite: protein

structure and function prediction. Nat. Methods 12: 7–8.

Yu, P., D. Ma, and M. Xu, 2005 Nested genes in the human genome. Genomics 86: 414–422.

Zimmermann, B., S. M. C. Robb, G. Genikhovich, W. J. Fropf, L. Weilguny *et al.*, 2020 Sea anemone genomes reveal ancestral metazoan chromosomal macrosynteny. bioRxiv 2020.10.30.359448.

Zook, J. M., D. Catoe, J. McDaniel, L. Vang, N. Spies *et al.*, 2016 Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data 3: 160025.

掘井直二郎, 1982 富山湾産発光ゴカイの観察. Science Report of the Yokosuka City Museum 1–3.