

The genomics of the gelatinous: Genomic insights into major evolutionary transitions within the Cnidaria

By

© 2018

Elizabeth S. Chang

M.A., University of Kansas, 2013

B.A., Swarthmore College, 2011

Submitted to the graduate degree program in Ecology and Evolutionary Biology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chair: Paulyn Cartwright

Nadine Folino-Rorem

John K. Kelly

Stuart J. Macdonald

Maria E. Orive

Date Defended: 26 October 2018

The dissertation committee for Elizabeth S. Chang certifies that this is the approved version of the following dissertation:

The genomics of the gelatinous: Genomic insights into major evolutionary transitions within the Cnidaria

Chair: Paulyn Cartwright

Date Approved: 31 October 2018

Abstract

Cnidaria is a marine phylum of over 13,000 species that possess an astounding diversity of habitats, body plans and life cycles. As early-diverging animals that are sister to all of the Bilateria, the study of cnidarians can help us reconstruct the evolutionary histories of traits in common to all bilaterians. The extreme diversity of cnidarian life histories, as well as their important phylogenetic position, makes Cnidaria an excellent group for the study of the drivers of diversity and the evolution of complexity and novelty. In my dissertation, I use phylogenetic, genomic, and population genetic approaches to study genome-scale and population-level changes associated with the evolution of major life history transitions. Chapter 1 is a brief introduction to cnidarian evolutionary genomics. In Chapter 2, I investigate the cnidarian transition to endoparasitism by contributing to phylogenomic analyses to place these parasites within the cnidaria, and characterizing major genome-scale changes, such as gene loss and genome size change, as compared with other non-parasitic cnidarians. In Chapter 3, I assess the population genetic consequences of the re-evolution of coloniality from a solitary ancestor in the hydrozoan *Ectopleura larynx*, which creates colonies by fusion of offspring rather than budding. My research tests whether or not these colonies are genetically chimeric, and therefore may be subject to evolutionary conflict between polyps. In Chapter 4, I gain insight into the cnidarian transition to freshwater by studying the euryhaline, invasive hydrozoan *Cordylophora*. I use phylogenomic and population genomic techniques in order to study the relationship between salinity level and population structure, reconstruct the evolution of salinity tolerance within *Cordylophora*, and to clarify some taxonomic uncertainty within the genus. My work contributes to the important and growing field of cnidarian evolutionary genomics and hopefully paves the way for future evolutionary genomic work on these and other intriguing cnidarian systems.

Acknowledgements

First, I would like to acknowledge the great privilege that has allowed me to complete this scientific work on a subject of my own choosing. This work could not have been done if not for the particular set of circumstances I happen to inhabit, including access to technology, relative stability throughout my life and the presence of a wonderful system of support. To quote one of the scientists I most admire, Stephen Jay Gould, *"I am somehow less interested in the weight and convolutions of Einstein's brain than in the near certainty that people of equal talent have lived and died in cotton fields and sweatshops."* I am extremely thankful for the opportunity to conduct this research and I dedicate this work to those of equal or greater talent who did not have this opportunity.

This dissertation has been a labor of love for over five years, and truly represents a collaborative effort – it really takes a village to raise a Doctor of Philosophy. First, thank you to those teachers throughout high school and college who helped to set me on this path towards a doctorate. Thank you especially to Dr. Sloe, who not only inspired me with her summers spent in residence at a research lab, but also assigned me to read Matt Ridley's "The Genome", which I can pinpoint as the first time I thought about how cool it might be to study genomics. At Swarthmore College, I had the opportunity to interact with many professors whose lessons I still think about today. Many thanks to Dr. Jason Downes and Dr. Steve Wang, who taught fantastic courses and let me get a first taste of independent research, as well as Dr. Rachel Merz, who gave me my first taste of how strange and wonderful marine invertebrates could be.

I have a large amount of gratitude as well for all of my colleagues who have made up my academic community over the past decade. Thank you to the Swarthmore Science Center Crew, who really taught me the value of persistence, companionship and keeping your sense of humor

intact no matter the circumstance. Thank you for your continuing friendship and I am so excited to spend this New Year's and all New Year's in the foreseeable future with you. Particular gratitude goes to the Lady Scientist Hangout crew, and even more especially to Althea Gaffney, (Dr.) Angela Wu and Jen Tinsman, who have been so good at staying in touch and being supportive of me even while working hard on their own doctorates.

It probably goes without saying (but I will say it anyway) that I am extremely grateful to my fellow University of Kansas EEB graduate students. I have been at KU long enough to see many of you come and go, and you have all inspired me in one way or another (or at least made it a fun time in the mean time!). Graduate school can be an arduous journey at times but your presence really smoothed the road. It would take another dissertation to list you all, but I have so much gratitude for the continued companionship and support of Lucas Hemmer, Alex Erwin, Kaila Colyott, Boryana Koseva, Emily Arsenault, Kaylee Herzog, Andrew Mongue, Az Klymiuk, Keely Brown, Desiree Harpel, Tori Pocius, Rachel Bowesm, Chris Hensz, Jack Colicchio, Logan Luce, Elise Krueger, and many, many others, including all past and present participants of EEB Genetics. Thank you also to Jaime Keeler, Aagje Ashe and Dorothy Johanning, who really helped with the logistics of the graduate school journey, and Jenny Hackett, who not only is my friend but also really helped me out in the Genome Sequencing Core. I would also to acknowledge that this work was supported by the HPC facilities operated by, and the staff of, the University of Kansas Center for Research Computing who offered invaluable technical advice throughout my dissertation.

Thank you also to friends that I have outside of graduate school as well, including Ellen Glock, Kelly Kluthe and Jon Lane, my Nerd Nite Co-Bosses and entire NNLFK community, as

well as the PBR Book Club and the Lawrence Trail Hawks. There are no words to express how helpful it was to be able to step away from my research and have a community there for me.

Thank you to my graduate committee, you have contributed so much to my intellectual growth and it has been a pleasure to work with you. There were definitely times when all of you were more convinced of my ability to finish this work than I was. I want to extend a special thanks to Dr. Maria E. Orive for her contributions to my growth as a scientist and who I had the great pleasure of collaborating with on a publication.

Thank you to the Cartwright Lab, past and present, for making the lab and office such a wonderful place to work. Thank you to Annalise Nawrocki, Steve Sanders, Bastian Bentlage and Mariya Shcheglovitova for being here at the beginning of it all when I was a fresh, innocent Master's student, and for Anna Klompen and Matthew Traver for being here to put up with me at as a very stressed Ph.D. candidate at the end of it all.

Of course, thank you to my advisor, Dr. Pauly Cartwright. There is no adequate way to express how much your mentorship has meant to me and contributed to me not only as a scientist but as a person. Even if I was not going to pursue a future scientific career, the time spent under your mentorship will have been well worth it. I will simply sum it up by saying that joining your lab was one of the best decisions of my life.

Finally, thank you to my family. Thank you to my mother in law, Teri Pope, who has been a wonderful presence in my life and is currently putting up with this grumpy Ph.D. student living in her basement. Thank you to my parents for always encouraging my interests in strange marine creatures, even when that involved taking a bunch of third graders to the Baltimore Aquarium, and for somehow knowing exactly when I needed a word of encouragement or a care package. Thank you to my sister, Sophie, for always being there for a serious conversation or

hilarious chat – I wonder if we will ever solve the mystery of whether it is more difficult to get a Ph.D. or start a Suzuki cello studio. And of course, thank you to my extended family in Maryland, Arkansas, California and everywhere else – I am so excited to see you more now that this had been turned in.

Last but not least, thank you for the continued love and support of my husband, Michael Pope. You have been so solid and stable throughout all of the turbulence of a graduate degree, and have always reminded me not to take myself too seriously, which may have been the most important key to surviving this process. You make every day such a joy and I absolutely cannot wait for our next adventure!

TABLE OF CONTENTS

CHAPTER 1 A BRIEF INTRODUCTION TO CNIDARIAN EVOLUTIONARY GENOMICS	1
BRIEF HISTORY OF CNIDARIAN EVOLUTIONARY GENOMICS	2
SUMMARY OF THE CONTRIBUTION OF THIS WORK	5
<i>Chapter 2: The evolution of cnidarian endoparasitism</i>	5
<i>Chapter 3: The re-evolution of coloniality in cnidarians</i>	6
<i>Chapter 4: The evolution of salinity tolerance in an invasive hydrozoan</i>	7
FUTURE APPLICATIONS OF GENOME SEQUENCING TO CNIDARIANS	8
LITERATURE CITED	11
CHAPTER 2 GENOMIC INSIGHTS INTO THE EVOLUTIONARY ORIGIN OF MYXOZOA WITHIN CNIDARIA	13
ABSTRACT	14
INTRODUCTION	15
MATERIALS AND METHODS	18
<i>Genome and Transcriptome Assemblies</i>	19
<i>Host contaminant filtering</i>	19
<i>Phylogenetic reconstruction</i>	20
<i>Analysis of assembly completeness</i>	22
<i>Estimation of genome size</i>	22
<i>Genome Annotation</i>	24
<i>Gene enrichment analysis of transcriptomes</i>	24
<i>Gene enrichment analysis of predicted genes from annotated genomes</i>	25
<i>Analysis of gene pathways and candidate genes</i>	25
<i>OrthoMCL Analysis</i>	26
RESULTS	27
<i>Phylogenetic position</i>	27
<i>Estimation of the completeness of genome and transcriptome assemblies</i>	28
<i>General characteristics of genomes</i>	28
<i>Characteristics of transcriptomes: Comparisons of gene ontology and gene orthology</i>	29
<i>Analyses of gene pathways and candidate genes in the assemblies</i>	31
DISCUSSION	32
<i>Note about data accessibility</i>	33
TABLES	34
FIGURES	38
LITERATURE CITED	41
CHAPTER 3 NON-CLONAL COLONIALITY: GENETICALLY CHIMERIC COLONIES THROUGH FUSION OF SEXUALLY PRODUCED POLYPS IN THE HYDROZOAN <i>ECTOPLEURA LARYNX</i>	46
ABSTRACT	47
INTRODUCTION	48
<i>Sampling of Ectopleura larynx</i>	52
<i>DNA extraction/Library construction/Sequencing</i>	52
<i>Read quality filtering/reconstruction of polymorphic loci using Stacks</i>	53
<i>Colony-level and population-level filtering of individuals and polymorphic loci</i>	54
<i>Determination of within-colony relationships using genetic distances</i>	54
<i>Evaluating patterns of intracolony allelic segregation</i>	55
<i>Assessing the effects of read-depth and allelic dropout on our results</i>	56
<i>Calculating population-wide estimates of diversity</i>	57

<i>Genomic estimates of relatedness between colony-mates</i>	58
RESULTS	59
<i>Determination of relationships via simulations</i>	59
<i>Comparison of genetic distances</i>	60
<i>Examination of individual loci</i>	61
<i>Population-level genetic diversity</i>	63
<i>Genomic measures of allele-sharing within and between colonies</i>	64
DISCUSSION	65
<i>Genetic composition of and levels of relatedness within <i>Ectopleura larynx</i> colonies</i>	65
<i>Low genetic diversity and self/non-self recognition</i>	67
<i>The potential for genetic conflict within colonies of <i>Ectopleura larynx</i></i>	69
<i>Conclusions</i>	69
TABLES	71
FIGURES	72
LITERATURE CITED	76
CHAPTER 4 SPECIES DELIMITATION AND THE EVOLUTION OF FRESHWATER TOLERANCE IN THE INVASIVE HYDROZOAN <i>CORDYLOPHORA</i> USING PHYLOGENETIC, POPULATION GENOMIC AND ENVIRONMENTAL EVIDENCE	81
ABSTRACT	82
INTRODUCTION	83
MATERIALS AND METHODS	88
<i>Sample Collection Methods</i>	88
<i>DNA Extraction and Sanger-Marker PCR</i>	89
<i>Generation and analysis of Sanger-sequenced markers</i>	89
<i>RAD-sequencing and phylogenetic analyses</i>	91
<i>Trait evolution and ancestral state reconstruction</i>	95
<i>Tests for gene flow within and between major clades of <i>Cordylophora</i></i>	96
<i>Species Delimitation Analyses</i>	97
RESULTS	98
<i>Collection of <i>Cordylophora</i> from across North American Range</i>	98
<i>Results of phylogenetic analyses using Sanger-sequenced markers</i>	99
<i>RAD-sequencing Data Set</i>	101
<i>Reconstruction of RAD-sequencing trees</i>	102
<i>Ancestral Character State Reconstruction using RAD-sequencing trees</i>	105
<i>Population Genetics Analyses of Clade 1</i>	105
<i>Hybrid Index analysis of Clade 1</i>	107
<i>Bayes Factor Delimitation analysis</i>	108
DISCUSSION	108
<i>RAD-seq phylogeny of the <i>Cordylophora</i> species complex</i>	108
<i>Evidence for multiple lineages of <i>Cordylophora</i> in estuaries</i>	109
<i>Geographic patterning in the <i>Cordylophora</i> species complex</i>	110
<i>Patterns of Salinity Tolerance Evolution in <i>Cordylophora</i></i>	112
<i>Integrative efforts for species delimitation in <i>Cordylophora</i></i>	113
<i>Taxonomic recommendations</i>	115
<i>Conclusions</i>	116
TABLES	117
FIGURES	126
LITERATURE CITED	134
APPENDIX	140

LIST OF TABLES

TABLE 2.1: TABLE 1 ASSEMBLY STATISTICS FOR SEQUENCED GENOMES AND TRANSCRIPTOMES.....	34
TABLE 2.2 ESTIMATED GENOME CHARACTERISTICS	35
TABLE 2.3 GENE ONTOLOGY CATEGORIES FROM TRANSCRIPTOMES SHOWING DEPLETION IN MYXOZOANS COMPARED TO OTHER CNIDARIANS.....	36
TABLE 2.4 PRESENCE (X) OR ABSENCE (-) OF GENES AND KEGG PATHWAYS THAT HAVE BEEN CHARACTERIZED IN OTHER CNIDARIANS.....	37
TABLE 3.1 PER-COLLECTING-LOCATION GENETIC DIVERSITY STATISTICS FOR <i>E. LARYNX</i>	71
TABLE 4.1 DESCRIPTION OF LOCALITIES SAMPLED FOR CORDYLOPHORA	117
TABLE 4.2 DESCRIPTION OF DATA SETS USED TO ESTIMATE PHYLOGENIES OF THE CORDYLOPHORA SPECIES COMPLEX	119
TABLE 4.3 MEASURES OF PHYLOGENETIC SUPPORT AND DIAGNOSABILITY FOR MAJOR CLADES IN 16S PHYLOGENY	120
TABLE 4.4 DESCRIPTIVE STATISTICS OF SELECTED RAD-SEQUENCING ASSEMBLIES	121
TABLE 4.5 MEASURES OF PHYLOGENETIC SUPPORT AND DIAGNOSABILITY FOR MAJOR CLADES RECOVERED IN RAD-SEQUENCE PHYLOGENY	122
TABLE 4.6 PER-INDIVIDUAL RESULTS OF HYBRID INDEX CALCULATIONS USING CLADE 1A AND CLADE 1C AS PARENTAL POPULATIONS.....	123
TABLE 4.7 PER-INDIVIDUAL RESULTS OF HYBRID INDEX CALCULATIONS USING CLADE 1A AND CLADE 1B AS PARENTAL POPULATIONS.....	124
TABLE 4.8 RESULTS OF BAYES FACTOR DELIMITATION ANALYSIS	125

LIST OF FIGURES

FIGURE 2.1 LIFE CYCLES OF <i>MYXOBOLUS CEREBRALIS</i> AND <i>POLYPODIUM HYDRIFORME</i>	38
FIGURE 2.2 PHYLOGENETIC TREE GENERATED FROM A MATRIX OF 51,940 AMINO-ACID POSITIONS AND 77 TAXA USING BAYESIAN INFERENCE UNDER THE CAT MODEL.....	39
FIGURE 2.3 GO ANNOTATION OF UNIGENES IN TRANSCRIPTOMES.....	40
FIGURE 3.1 <i>ECTOPLEURA LARYNX</i>	72
FIGURE 3.2 GENETIC RELATIONSHIPS WITHIN SELECTED COLONIES OF <i>E. LARYNX</i>	73
FIGURE 3.3 DENSITY DISTRIBUTIONS OF THE COMBINED PAIRWISE GENETIC DISTANCES AND RESULTS OF K-MEANS CLUSTERING FOR THE THREE MAJOR RELATIONSHIP TYPES	74
FIGURE 3.4 FIGURE 4. DISTRIBUTIONS OF VALUES FOR SEVERAL MEASURES OF RELATEDNESS FOR WITHIN- AND BETWEEN-COLONY COMPARISONS BETWEEN INDIVIDUALS.....	75
FIGURE 4.1 FIGURE 1. SCHEMATIC OF COLLECTING SITES ALONG THE SAN FRANCISCO BAY	126
FIGURE 4.2 PHYLOGENETIC TREE OF THE CORDYLOPHORA SPECIES COMPLEX BASED ON MAXIMUM- LIKELIHOOD ANALYSIS OF 16S SEQUENCES	127
FIGURE 4.3 BOXPLOT OF SALINITY VS. CLADE BASED ON 16S PHYLOGENY	128
FIGURE 4.4 PRESENCE-ABSENCE MATRIX OF LOCI RETAINED IN DEFAULT (A) AND FINAL (B) RAD-SEQ ASSEMBLIES.....	129
FIGURE 4.5 RAD-SEQ MAXIMUM-LIKELIHOOD PHYLOGENY OF CORDYLOPHORA SPECIES COMPLEX BASED ON ALIGNMENT OF 23852 UNLINKED SNPs.....	130
FIGURE 4.6 BOXPLOT OF SALINITY VS. CLADE BASED ON RAD-SEQ PHYLOGENY.....	131
FIGURE 4.7 BAYESIAN ANCESTRAL CHARACTER STATE RECONSTRUCTION OF NATIVE SALINITY BASED ON RAD-SEQUENCING PHYLOGENY	132
FIGURE 4.8 INFERRED POPULATION STRUCTURE FOR INDIVIDUALS MAKING UP CLADE 1, EXCLUDING SUBCLADE 1B (A) AND INCLUDING SUBCLADE 1B (B).....	133

LIST OF APPENDICES

APPENDIX 1: SAMPLE INFORMATION AND ACCESSION NUMBERS	140
APPENDIX 2: PHYLOGENETIC TREE GENERATED FROM A MATRIX OF 41,237 AMINO ACID POSITIONS, WHICH EXCLUDES RIBOSOMAL GENES, AND 77 TAXA USING BAYESIAN INFERENCE UNDER THE CAT MODEL	141
APPENDIX 3: PHYLOGENETIC RECONSTRUCTIONS WITH ONLY CNIDARIAN TAXA	142
APPENDIX 4: SEQUENCE SIZE DISTRIBUTION OF THE ASSEMBLED TRANSCRIPTOME SEQUENCES	143
APPENDIX 5: GO ANNOTATION OF UNIGENES IN GENOMES AND TRANSCRIPTOMES	144
APPENDIX 6: COMPARISON OF OGS IN MYXOZOAN AND OTHER CNIDARIAN TRANSCRIPTOMES.....	145
APPENDIX 7: COLLECTION SITES OF <i>ECTOPLEURA LARYNX</i> COLONIES FROM THE COAST OF MAINE AND IRELAND	146
APPENDIX 8: . SUMMARY OF LOCI FOR EACH COLONY-LEVEL DATA SET	146
APPENDIX 9: DESCRIPTION OF DATA SETS USED IN EACH ANALYSIS IN CHAPTER 3	147
APPENDIX 10: GENETIC RELATIONSHIPS WITHIN SELECTED COLONIES OF <i>E. LARYNX</i> NOT ALREADY INCLUDED IN FIGURE 3.2	148
APPENDIX 11: RESULTS OF DIFFERENT METHODS FOR CHOOSING A BEST K (NUMBER OF CLUSTERS).....	149
APPENDIX 12: NUMBER OF CLUSTERS (K) FOR THE FOR THE PROPORTION OF SITES WHICH DIFFER IN A WITHIN-COLONY COMPARISON REPORTED FROM DIFFERENT ALGORITHMS.....	150
APPENDIX 13: COMPARISON OF READ-DEPTH DISTRIBUTIONS BETWEEN SNPS WHERE POLYPS WERE ALTERNATIVE HOMOZYGOTES VS. ALL OTHER SNPS.....	151
APPENDIX 14: HISTOGRAM OF THE NUMBER OF SITES AT WHICH POLYPS DIFFER IN COMPARISONS BETWEEN COLONIES AT A GIVEN COLLECTING SITE.....	152
APPENDIX 15: BETWEEN-COLONY PAIRWISE FST VALUES FOR COLLECTING SITES	153
APPENDIX 16: SUMMARY OF MAJOR PER-COLONY RESULTS.....	153
APPENDIX 17: PHYLOGENETIC TREE OF <i>CORDYLOPHORA</i> BASED ON MAXIMUM-LIKELIHOOD ANALYSIS OF 28S MARKERS	155
APPENDIX 18: PHYLOGENETIC TREE OF <i>CORDYLOPHORA</i> BASED ON MAXIMUM-LIKELIHOOD ANALYSIS OF CO1 SEQUENCES	156
APPENDIX 19: PHYLOGENETIC TREE OF <i>CORDYLOPHORA</i> BASED ON MAXIMUM-LIKELIHOOD ANALYSIS OF CONCATENATED CO1+16S SEQUENCES	157
APPENDIX 20: PHYLOGENETIC TREE OF <i>CORDYLOPHORA</i> BASED ON MAXIMUM-LIKELIHOOD ANALYSIS OF CONCATENATED 16S+28S+CO1 SEQUENCES.....	158
APPENDIX 21: PHYLOGENETIC TREE OF <i>CORDYLOPHORA</i> BASED ON MAXIMUM-LIKELIHOOD ANALYSIS COMBINED RAD SEQUENCE AND SANGER DATA.....	159
APPENDIX 22: RAD-SEQ MAXIMUM-LIKELIHOOD PHYLOGENY OF <i>CORDYLOPHORA</i> BASED ON AN ALIGNMENT OF UNLINKED NUCLEOTIDE SITES FOR THREE COLONIES PER LOCALITY	162
APPENDIX 23: RAD-SEQ MAXIMUM-LIKELIHOOD PHYLOGENY OF <i>CORDYLOPHORA</i> BASED ON AN ALIGNMENT OF UNLINKED NUCLEOTIDE SITES FOR THREE COLONIES PER LOCALITY, 50% MISSING DATA	163
APPENDIX 24: RAD-SEQ MAXIMUM-LIKELIHOOD PHYLOGENY OF <i>CORDYLOPHORA</i> WITH NODE NUMBERS FOR INTERPRETATION OF APPENDIX 25.....	164
APPENDIX 25: SUMMARY OF BAYESIAN ANCESTRAL CHARACTER STATE RECONSTRUCTION FOR EACH INTERNAL NODE OF THE RAD-SEQ PHYLOGENY.....	165
APPENDIX 26: SUMMARY OF EVANNO METHOD FOR CLADE 1, NOT INCLUDING SUBCLADE 1B.....	170
APPENDIX 27: SUMMARY OF EVANNO METHOD FOR CLADE 1, INCLUDING SUBCLADE 1B.....	170

Chapter 1
A Brief Introduction to Cnidarian Evolutionary Genomics

Brief history of cnidarian evolutionary genomics

Cnidaria is a diverse phylum of over 13,000 species, all possessing some form of the nematocyst, a complex stinging organelle. The phylum is divided into well-supported clades: the anthozoans (sea anemones and corals) and medusozoans, the latter of which contains the hydrozoans (model organism *Hydra*, colonial forms and the Portuguese man-of-war), the scyphozoans (the “true jellies”), cubozoans (box jellies) and staurozoans (stalked jellies). The cnidarians are one of the earliest diverging animal groups that are sister to Bilateria and are thus an excellent system for studying evolutionary patterns and processes. In addition, they have an astounding diversity of habitat usage, body plans and life cycles. Medusozoans have a life cycle stage that includes the free-swimming medusa and the sessile polyp, or both, with the prominence of each of these stages differing across the phylum. Further, during the polyp stage, they can either be solitary or colonial, and colonies contain either monotypic or morphologically distinct and functionalized polymorphic polyps. Cnidarian taxa also display a variety in life histories, with regard to relative contribution of sexual and asexual reproduction. Finally, they have managed to colonize an amazing amount of marine habitat – from coral reefs, to the deep sea, to the rocky intertidal and within other organisms as parasites.

This extreme diversity of form and life cycle gives us an opportunity to study the drivers of diversity, the repeated evolution of certain features and the evolution of complexity and novelty. Despite having an extensive natural history literature and hundreds of years of examination from notable biologists such as Huxley, Agassiz, and Haeckel in order to understand phenomena like regeneration, asexual reproduction, and coloniality (Cartwright et al. 1999; Chapman et al. 2010), many open, tantalizing questions remain about the Cnidaria. For instance, understanding how the major cnidarian groups are related to one another, the genetic

basis behind their life history diversity, and whether their genome size and content reflects or contributes to this diversity, are all open areas of research. Further, as early-diverging animals that are sister to all of the Bilateria, the study of cnidarians can help us reconstruct the evolutionary histories of traits in common to all bilaterians and whether particular genetic toolkits or developmental pathways existed early in animal evolution, in the common ancestor to bilaterians and cnidarians.

The publication of two cnidarian genomes, the model organism *Nematostella* in 2007 (Putnam et al. 2007) and *Hydra* in 2010 (Chapman et al. 2010) immediately gave us insight into the origins of novelty in this group, and paved the way for comparative genomic work (reviewed in Steele et al. 2011). One observation was that the genomes were both unexpectedly complex and remarkably different in both genome size and characteristics, mirroring their millions of years of divergence from one another. Despite this length of divergence between each other and from Bilateria, it is clear that many genomic elements originally thought to be characteristic of bilaterians, are actually shared by both cnidarian representatives and bilaterians and therefore likely predate the cnidarian/bilaterian split. For example, the cnidarians possess what appear to be typical animal telomeres, despite not appearing to senesce (Traut et al. 2007), and possess almost all proteins known to facilitate cell-cell contact in bilaterians (Chapman et al. 2010). Strikingly, both *Hydra* and *Nematostella* contain functional Hox codes, which had long been assumed to be a bilaterian innovation (Ryan et al. 2007). In fact, cnidarian genomes appear overall to have high conservation of linkage with other metazoan genomes, and have higher synteny and possession of certain gene groups in common with humans than do traditional model organisms *Drosophila* and *C. elegans* (Steele et al. 2011). In addition, study of these genomes also yielded the

discovery of gene families specific to cnidarians such as those that code for parts of the nematocyst (Steele et al. 2011).

Since the publication of the first two cnidarian genomes, the rapid development of genome sequencing technologies and related bioinformatics tools have given researchers the opportunity to apply genomics to a variety of evolutionary and population-level questions. In particular, the development of “next generation” sequencing technologies allowed for the rapid growth of comparative and evolutionary genomics in Cnidaria beyond *Hydra* and *Nematostella* by allowing for the relatively rapid and inexpensive sequencing of all or parts of the genome of an organism or set of expressed sequences (transcriptome). The ease and affordability of sequencing has provided for more investigation of genomic topics in Cnidaria, such as: determining the relationships between major cnidarian taxa (Zapata et al. 2015; Kayal et al. 2018; Munro et al. 2018), discovery of the genes involved in development of polymorphic polyp types from a common genetic background (Sanders et al. 2014) and different life cycle stages (Sanders and Cartwright 2015), determining the genes involved in responses to coral bleaching (Thomas and Palumbi 2017), and characterization and evolutionary history of genes specific to the stinging cells (David et al. 2008; Shpirer et al. 2014; Shpirer et al. 2018).

In this dissertation I use short-read genome scale data sets to focus on specific evolutionary transitions in cnidarians and what to determine genomics can inform us about their natural history: 1) The evolution of parasitism and the accompanying changes to genome size and content, 2) The genotypic composition and population structure in a hydrozoan that forms genetically chimeric colonies, and 3) Population genomics and species delimitation in a freshwater hydrozoan. Through the characterization of cnidarian genomes and genome-scale data sets, I provide insight into some of the most remarkable evolutionary transitions that have

occurred within Cnidaria and demonstrate the power of genomic techniques to uncover new knowledge about complex events in animal evolution. These topics are introduced below.

Summary of the contribution of this work

Chapter 2: The evolution of cnidarian endoparasitism

The evolution from a free-living form to a parasitic form often requires many extreme adaptations involved in harnessing resources from the host and evading the host immune system, which can be accompanied by reductions in body size and complexity. Within the cnidarians, there are several endoparasitic taxa such as the monospecific *Polypodium*, which parasitizes the Acipenseriform fishes and possesses a cnidarian-like body plan with tentacles and a gut, and the diverse Myxozoa, which possess a more complex life cycle and a much further reduced body plan. We compared the very distinct forms of parasitism in the myxozoans and *Polypodium* in order to understand whether simplicity in body plan necessarily relates to genome simplicity and also assessed their phylogenetic placement within the cnidaria to investigate whether endoparasitism evolved more than once.

To address these questions, we sequenced, assembled and assigned function to genomes and transcriptomes of *Polypodium* and several myxozoans. This work allowed us to conduct phylogenomic analyses across animals with the best cnidarian taxon sampling to date in order to place these groups firmly within Cnidaria, and to demonstrate that there is probably just one origin of endoparasitism within the cnidarians (Chang et al. 2015). This comprehensive set of phylogenomic sequence data, including our new assemblies, have since been used and cited in subsequent studies since 2015, such as Whelan et al. (2017), Kayal et al. (2018), Richter et al. (2018). Further, we use the genome and transcriptome assemblies to show massive reduction of

both genome size and loss of genes thought crucial for animal development (i.e. Hox and Wnt family genes) in the morphologically reduced myxozoans but not in the more cnidarian-like *Polypodium*. This chapter was published as part of a collaborative effort with multiple contributors (Chang et al. 2015) I served as lead author and my contributions were generating and assembling a large portion of the genomic and transcriptomic data, leading the efforts to annotate, characterize and estimate sizes for the genomes, and spearheading the analyses of the presence/absence of gene families and genetic pathways. Initial characterization of the *Polypodium* and myxozoan data was done as part of my Master's project (Chang 2013).

Chapter 3: The re-evolution of coloniality from a solitary ancestor

Coloniality is traditionally defined as conspecific organisms living close to or physically attached to one another for mutual benefit. In cnidarians a key evolutionary innovation is to form colonies through clonal reproduction, which allows them to rapidly grow in size and compete for space in substrate-limited environments. In hydrozoans this coloniality takes on a very extreme form, wherein new polyps are created through asexual budding and remain attached, creating physiologically integrated colonies that share digestive systems and epithelia. However, a group of hydrozoans called Aplanulata, which includes *Hydra* and consists primarily of solitary polyps, have lost coloniality, and subsequently their ability to grow the colony through asexual budding alone, in their evolution (Nawrocki et al. 2013). Several species in the aplanulata genus *Ectopleura* have re-evolved large colony size, but do not achieve this large colony size through asexual budding, but instead through the aggregation and fusion of sexually (non-clonally) produced polyps (Nawrocki and Cartwright 2012). The apparent chimerism within a physiologically integrated colony presents a potential source of conflict between distinct genetic

lineages, which may vary in their ability to access the germline. In order to determine the extent to which the potential for genetic conflict exists, we characterized the types of genetic relationships present between polyps within colonies of *Ectopleura larynx*, using a reduced-representation sequencing approach allowing us to sequence material from many individuals at once. We find that indeed these colonies are chimeric, comprising close familial individuals, but surprisingly, also involve fusion of unrelated individuals. We also find that background genetic diversity in *E. larynx* populations is low. This suggests that an evolutionary compromise between large colony size and genetic homogeneity may be possible when genetic diversity is relatively low. This gives insight into the genetic consequences of the re-evolution of a complex trait which was previously lost and the conflict and compromises therein. This work has been published and is part of a collaborative effort (Chang et al. 2018). I was principal author and I analyzed and interpreted all of the data and led the writing of the publication.

Chapter 4: The evolution of salinity tolerance in an invasive hydrozoan

While mostly a marine phylum, a few cnidarians have managed to make the physiologically challenging transition to freshwater, allowing for expansion into new habitats. Most cnidarians tolerant of lower salinities have evolved to be obligately freshwater, such as the model organism *Hydra*. The invasive hydrozoan *Cordylophora* is unique in that it is tolerant of a wide range of salinities and so gives us an opportunity to study a potentially transitional system. Previous studies suggest it may be a species complex comprised of lineages of exclusively brackish, freshwater and euryhaline colonies (Folino-Rorem et al. 2009). Taxonomic uncertainty exists, as it is not known if these lineages rise to the level of independently evolving species. We used the same reduced-representation genome sequencing technique as in Chapter 3 to generate a

well-sampled phylogeny, including newly collected samples from estuaries where the diverged lineages may still come back into contact with one another. We use this data to analyze patterns of salinity habitat trait evolution within the complex. We conclude that there are at least two major lineages of *Cordylophora* separated by both salinity regime and geography, and suggest resurrecting a previously synonymized species to represent the primarily freshwater lineage. This work provides a fuller understanding of species boundaries and phylogenetic patterns, as well as providing initial genomic resources for *Cordylophora*, all of which will provide the basis for future work on the cnidarian evolution of freshwater tolerance and how that tolerance can lead to invasiveness.

A publication of the work is in preparation on which I will serve as lead author, which is part of a collaborative effort. I assisted in collecting some of the newly sequenced samples, isolated their DNA, led the effort for library preparation and sequencing, analyzed and interpreted all of the data and wrote the initial drafts.

Future applications of genome sequencing to cnidarians

My work and the work of many other researchers highlights the fruitful nature of genomic investigations into cnidarian biology and the promise for more valuable research using recently developed and emerging technologies. One way in which the cnidarian evolutionary genomic field is expanding is through the development of more cnidarian model systems. One system which is the focus of an ongoing, large-scale genome sequencing project is that of *Hydractinia*, a colonial hydrozoan. This genome sequencing will further the development of *Hydractinia* as a model for allorecognition, development and and regeneration (Rosengarten and Nicotra 2011; Bradshaw et al. 2015; Gahan et al. 2016). Other cnidarians are being developed

into models for other phenomena, such as understanding the effects of climate change on coral and their symbionts by using *Aiptasia*, an anemone that is can be cultured in lab (Weis et al. 2008; Rentzsch and Technau 2016). Additionally, ease of sequencing and the development of state-of-the-art phylogenetic and bioinformatics tools are allowing for ever broader taxon sampling of the cnidarians for phylogenomic and comparative genomic studies (i.e. Zapata et al. 2015; Kayal et al. 2018).

Further development of genome-related technologies will allow us to understand genome content and function more completely than ever. For example, many current genome sequencing projects mentioned above are incorporating long-read sequencing to assemble more contiguous and accurate genome assemblies than those using short-read next generation sequencing data alone. These more completely assembled genomes will be important for many applications such as assessing gene organization, large-scale synteny between different species, and building better references for population genomics (Shendure et al. 2017)

Additionally, the introduction of sophisticated developmental biology assays and genome editing methods such as the CRISPR/Cas-9 are opening the door for researchers to hone in on the function of specific genes. This will allow for researchers to study the evolution of that function across Cnidaria and the conservation of that function between cnidarians and bilaterians. Much work has been done on *Nematostella*, building on earlier functional and genomic work to understand the role of certain genes on developmental biology (Rentzsch and Technau 2016). CRISPR has also being utilized successfully in *Hydractinia*, the first example of the germline transmission of a CRISPR/Cas9 inserted transgene into *Hydractinia* (Sanders et al. 2018). In general, gene knockout and knock-in techniques have proven useful for uncovering developmental mechanisms and comparing them across taxa. For instance, using a short-hairpin

RNA technique, it was recently demonstrated that despite lacking an apparent anterior-posterior axis, which is established by the Hox code in bilaterians, this code is employed in *Nematostella* to control segmentation of the larval endoderm (He et al. 2018). Additionally, single-cell transcriptomics are allowing researchers to generate extremely fine-scale maps of gene function in entire organisms, which has been used to investigate cell type diversity and developmental patterning in thus far in *Nematostella* (Sebe-Pedros et al. 2018) and *Hydra* (Siebert et al. 2018), and will likely be an important tool in emerging model systems.

Over the past decade, we have managed to learn a great deal about the underpinnings of cnidarian biology and natural history by using genomic approaches. Genomic research has thus far helped to link genotypes to phenotypes and to address questions about the relationship between genomic complexity and novelty and morphological complexity and novelty. Through my dissertation, I have successfully contributed to this growing field by using genome-scale data sets to investigate the genomic contributions to and consequences of evolutionary novelty, paving the way for future evolutionary genomic work on these and other intriguing cnidarian systems.

LITERATURE CITED

- Bradshaw, B., K. Thompson, and U. Frank. 2015. Distinct mechanisms underlie oral vs aboral regeneration in the cnidarian *Hydractinia echinata*. *eLife* 4:e05506.
- Cartwright, P., J. Bowsher, and L. W. Buss. 1999. Expression of a Hox gene, *Cnox-2*, and the division of labor in a colonial hydroid. *Proceedings of the National Academy of Sciences of the United States of America* 96:2183-2186.
- Chang, E. S. 2013. Transcriptomic evidence that enigmatic parasites *Polypodium hydriforme* and myxozoa are cnidarians. Pp. 53. *Ecology and Evolutionary Biology*. University of Kansas, Lawrence, KS
- Chang, E. S., M. Neuhof, N. D. Rubinstein, A. Diamant, H. Philippe, D. Huchon, et al. 2015. Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. *Proceedings of the National Academy of Sciences* 112:14912.
- Chang, E. S., M. E. Orive, and P. Cartwright. 2018. Nonclonal coloniality: Genetically chimeric colonies through fusion of sexually produced polyps in the hydrozoan *Ectopleura larynx*. *Evolution Letters* 2:442-455.
- Chapman, J. A., E. F. Kirkness, O. Simakov, S. E. Hampson, T. Mitros, T. Weinmaier, et al. 2010. The dynamic genome of *Hydra*. *Nature* 464:592.
- David, C. N., S. Özbek, P. Adamczyk, S. Meier, B. Pauly, J. Chapman, et al. 2008. Evolution of complex structures: minicollagens shape the cnidarian nematocyst. *Trends in Genetics* 24:431-438.
- Folino-Rorem, N. C., J. A. Darling, and C. A. D'Ausilio. 2009. Genetic analysis reveals multiple cryptic invasive species of the hydrozoan genus *Cordylophora*. *Biological Invasions* 11:1869-1882.
- Gahan, J. M., B. Bradshaw, H. Flici, and U. Frank. 2016. The interstitial stem cells in *Hydractinia* and their role in regeneration. *Current Opinion in Genetics & Development* 40:65-73.
- He, S., F. del Viso, C.-Y. Chen, A. Ikmi, A. E. Kroesen, and M. C. Gibson. 2018. An axial Hox code controls tissue segmentation and body patterning in *Nematostella vectensis*. *Science* 361:1377.
- Kayal, E., B. Bentlage, M. Sabrina Pankey, A. H. Ohdera, M. Medina, D. C. Plachetzki, et al. 2018. Phylogenomics provides a robust topology of the major cnidarian lineages and insights on the origins of key organismal traits. *BMC Evolutionary Biology* 18:68.
- Munro, C., S. Siebert, F. Zapata, M. Howison, A. Damian-Serrano, S. H. Church, et al. 2018. Improved phylogenetic resolution within Siphonophora (Cnidaria) with implications for trait evolution. *Molecular Phylogenetics and Evolution* 127:823-833.
- Nawrocki, Annalise M. and P. Cartwright. 2012. A Novel Mode of Colony Formation in a Hydrozoan through Fusion of Sexually Generated Individuals. *Current Biology* 22:825-829.
- Nawrocki, A. M., A. G. Collins, Y. M. Hirano, P. Schuchert, and P. Cartwright. 2013. Phylogenetic placement of *Hydra* and relationships within Aplanulata (Cnidaria: Hydrozoa). *Molecular Phylogenetics and Evolution* 67:60-71.
- Putnam, N. H., M. Srivastava, U. Hellsten, B. Dirks, J. Chapman, A. Salamov, et al. 2007. Sea Anemone Genome Reveals Ancestral Eumetazoan Gene Repertoire and Genomic Organization. *Science* 317:86.

- Rentzsch, F. and U. Technau. 2016. Genomics and development of *Nematostella vectensis* and other anthozoans. *Current Opinion in Genetics & Development* 39:63-70.
- Richter, D. J., P. Fozouni, M. B. Eisen, and N. King. 2018. Gene family innovation, conservation and loss on the animal stem lineage. *eLife* 7:43.
- Rosengarten, R. D. and M. L. Nicotra. 2011. Model Systems of Invertebrate Allorecognition. *Current Biology* 21:R82-R92.
- Ryan, J. F., M. E. Mazza, K. Pang, D. Q. Matus, A. D. Baxevanis, M. Q. Martindale, et al. 2007. Pre-Bilaterian Origins of the Hox Cluster and the Hox Code: Evidence from the Sea Anemone, *Nematostella vectensis*. *PLOS ONE* 2:e153.
- Sanders, S. M. and P. Cartwright. 2015. Interspecific Differential Expression Analysis of RNA-Seq Data Yields Insight into Life Cycle Variation in Hydractiniid Hydrozoans. *Genome Biology and Evolution* 7:2417-2431.
- Sanders, S. M., Z. Ma, J. M. Hughes, B. M. Riscoe, G. A. Gibson, A. M. Watson, et al. 2018. CRISPR/Cas9-mediated gene knockin in the hydroid *Hydractinia symbiolongicarpus*. *BMC Genomics* 19:649.
- Sanders, S. M., M. Shcheglovitova, and P. Cartwright. 2014. Differential gene expression between functionally specialized polyps of the colonial hydrozoan *Hydractinia symbiolongicarpus* (Phylum Cnidaria). *BMC Genomics* 15:406.
- Sebe-Pedros, A., B. Saudemont, E. Chomsky, F. Plessier, M. P. Mailhe, J. Renno, et al. 2018. Cnidarian Cell Type Diversity and Regulation Revealed by Whole-Organism Single-Cell RNA-Seq. *Cell* 173:1520-1534.e1520.
- Shendure, J., S. Balasubramanian, G. M. Church, W. Gilbert, J. Rogers, J. A. Schloss, et al. 2017. DNA sequencing at 40: past, present and future. *Nature* 550:345.
- Shpirer, E., E. S. Chang, A. Diamant, N. Rubinstein, P. Cartwright, and D. Huchon. 2014. Diversity and evolution of myxozoan minicollagens and nematogalectins. *BMC Evolutionary Biology* 14:205.
- Siebert, S., J. A. Farrell, J. F. Cazet, Y. Abeykoon, A. S. Primack, C. E. Schnitzler, et al. 2018. Stem cell differentiation trajectories in *Hydra* resolved at single-cell resolution. *bioRxiv*.
- Shpirer, E., A. Diamant, P. Cartwright, and D. Huchon. 2018. A genome wide survey reveals multiple nematocyst-specific genes in Myxozoa. *BMC Evolutionary Biology* 18:138.
- Steele, R. E., C. N. David, and U. Technau. 2011. A genomic view of 500 million years of cnidarian evolution. *Trends in genetics : TIG* 27:7-13.
- Thomas, L. and S. R. Palumbi. 2017. The genomics of recovery from coral bleaching. *Proceedings of the Royal Society B: Biological Sciences* 284.
- Traut, W., M. Szczepanowski, M. Vitkova, C. Opitz, F. Marec, and J. Zrzavy. 2007. The telomere repeat motif of basal Metazoa. *Chromosome Res.* 15:371-382.
- Weis, V. M., S. K. Davy, O. Hoegh-Guldberg, M. Rodriguez-Lanetty, and J. R. Pringle. 2008. Cell biology in model systems as the key to understanding corals. *Trends in Ecology & Evolution* 23:369-376.
- Whelan, N. V., K. M. Kocot, T. P. Moroz, K. Mukherjee, P. Williams, G. Paulay, et al. 2017. Ctenophore relationships and their placement as the sister group to all other animals. *Nat. Ecol. Evol.* 1:1737-1746.
- Zapata, F., F. E. Goetz, S. A. Smith, M. Howison, S. Siebert, S. H. Church, et al. 2015. Phylogenomic Analyses Support Traditional Relationships within Cnidaria. *PLOS ONE* 10:e0139068.

Chapter 2
Genomic Insights into the Evolutionary Origin of Myxozoa Within Cnidaria

ABSTRACT

The Myxozoa comprise over 2,000 species of microscopic obligate parasites that utilize both invertebrate and vertebrate hosts as part of their life cycle. Although the evolutionary origin of myxozoans has been elusive, a close relationship with cnidarians, a group that includes corals, sea anemones, jellyfish, and hydroids, is supported by some phylogenetic studies and the observation that the distinctive myxozoan structure, the polar capsule, is remarkably similar to the stinging structures (nematocysts) in cnidarians. In order to gain insight into the extreme evolutionary transition from a free-living cnidarian to a microscopic endoparasite, we analyzed genomic and transcriptomic assemblies from two distantly related myxozoan species, *Kudoa iwatai* and *Myxobolus cerebralis*, and compared these to the transcriptome and genome of the less reduced cnidarian parasite, *Polypodium hydriforme*. A phylogenomic analysis, using for the first time, a taxonomic sampling that represents the breadth of myxozoan diversity, including four newly generated myxozoan assemblies, confirms that myxozoans are cnidarians and are a sister taxon to *Polypodium hydriforme*. Estimations of genome size reveal that myxozoans have one of the smallest reported animal genomes. Gene enrichment analyses show depletion of expressed genes in categories related to development, cell differentiation, and cell-cell communication. In addition, a search for candidate genes indicates that myxozoans lack key elements of signaling pathways and transcriptional factors important for multicellular development. Our results suggest that the degeneration of the myxozoan body plan from a free-living cnidarian to a microscopic parasitic cnidarian, was accompanied by extreme reduction in genome size and gene content.

INTRODUCTION

Obligate parasitism can lead to dramatic reduction of body plans and associated morphological structures (Hoeg 1995; Kobayashi et al. 1999). One of the most spectacular examples are the microscopic Myxozoa, which spend most of their parasitic life cycle as just a few cells in size (Kent et al. 2001). The most conspicuous myxozoan cell type houses a polar capsule, which is a complex structure with an eversible tube (or filament) that is thought to facilitate attachment to the host. The polar capsule bears remarkable similarity to the stinging structures (nematocysts) of cnidarians (corals, sea anemones, jellyfish, and hydroids), suggesting that nematocysts and polar capsules are homologous and that myxozoans are related to cnidarians (Štolc 1899; Weill 1938).

Myxozoa are a diverse group of obligate endoparasites that comprise over 2,180 species (Lom and Dykova 2006). The vast majority of myxozoan species alternate between a fish and annelid host. In *Myxobolus cerebralis* (Wolf and Markiw 1984), the causative agent for whirling disease in rainbow trout (Hedrick et al. 1998), the annelid host *Tubifex tubifex*, releases infective actinospores (Fig. 2.1A) which subsequently anchor on to the fish, injecting the sporoplasm into the host tissue (El-Matbouli et al. 1995). Infective myxospores develop within the fish and are eventually ingested by the annelid where they develop into an actinospore, hereby completing the parasitic life cycle (Fig. 2.1A). Although the vast majority of myxozoan species are comprised of just a few cells, some malacosporean myxozoans, such as *Buddenbrockia plumatellae*, have a complex vermiform life cycle stage (Monteiro et al. 2002), which is a derived trait that has been lost and re-gained several times within this particular lineage (Hartikainen et al. 2014).

Another obligate cnidarian endoparasite, *Polypodium hydriforme*, in contrast to myxozoans, does not have a degenerate body form and instead displays conventional cnidarian-like features, including tentacles, a gut, and mouth (Fig. 2.1B). *P. hydriforme* lies dormant as a binucleate cell within the oocytes of female acipenseriform fishes (paddlefish and sturgeon) (Fig 1B), eventually developing into an elongated stolon, which emerges from the host's oocyte upon spawning. Once freely living, the stolon begins to fragment into multiple individuals, each developing a mouth to feed with (Fig 2.1B). Adult *P. hydriforme* infect juvenile female fish to repeat the life cycle (Raikova et al. 1979).

Classification of Myxozoa and *P. hydriforme* has been controversial (Evans 2010; Foox and Siddall 2015; Okamura and Gruhl 2015). From the first descriptions in the 1880s until relatively recently, myxozoans were considered to be protists, largely due to their highly reduced, microscopic construction (Bütschli 1881). Unlike Myxozoa, the placement of the monotypic species *P. hydriforme* as a cnidarian has long been proposed based on morphology (Lipin 1925; Raikova 1988). With the advent of molecular phylogenetics, it was discovered that myxozoans are not protists, but instead are metazoans (Smothers et al. 1994; Siddall et al. 1995). The first studies, mainly based on analyses of 18S rDNA, usually recovered Myxozoa as the sister taxon to *P. hydriforme*. However, the position of this clade was unstable. It was either placed as the sister clade to Bilateria, or nested within Cnidaria, depending on taxon sampling, alignment, optimization method, and the characters considered (Siddall et al. 1995; Siddall and Whiting 1999; Zrzavý and Hypša 2003; Evans et al. 2008; Evans 2010). Recent phylogenomic studies support a position of Myxozoa within Cnidaria, as the sister clade to Medusozoa (Jiménez-Guri et al. 2007; Nesnidal et al. 2013; Feng et al. 2014). However, in these studies *P.*

hydriforme and representatives of major lineages of myxozoan and non-myxozoan cnidarians were notably absent. Thus, the precise phylogenetic position within Cnidaria remains uncertain.

In recent studies, myxozoans were found to possess the cnidarian-specific minicollagen and nematogalactin genes (Holland et al. 2011; Feng et al. 2014; Shpirer et al. 2014), each of which has been shown to play important roles in the nematocyst structure in *Hydra* (Adamczyk et al. 2008; Hwang et al. 2010). These studies support previous morphology-based assertions, that myxozoan polar capsules are homologous to cnidarian nematocysts (Štolc 1899; Weill 1938; Reft and Daly 2012; Okamura and Gruhl 2015) and thus indirectly suggest a close evolutionary relationship between these two groups.

For this study, we analyzed genomic and transcriptomic assemblies from two distantly related myxosporean myxozoans, *Kudoa iwatai* and *M. cerebralis*, as well as the cnidarian parasite *P. hydriforme*, in order to gain insight into the evolutionary transition to parasitism and extreme reduction of body plans from a free-living cnidarian. First, we used these newly generated data, in conjunction with publicly available data, to determine the phylogenetic relationships between all major lineages of myxozoans (Fiala and Bartošová 2010), *P. hydriforme*, and other cnidarians, in order to reconstruct the evolutionary history of endoparasitism in Cnidaria. Second, we compared genome size, gene number, gene content, and enrichment of expressed genes between myxozoans, *P. hydriforme*, and published cnidarian genomes, in order to determine if the degeneration of the cnidarian body plan displayed in Myxozoa (but not in *P. hydriforme*) was accompanied by genome reduction and gene loss. Our findings re-affirm a cnidarian origin for myxozoans and recover them as the sister group to *P. hydriforme*. Analysis of genome and transcriptome assemblies reveal that the highly degenerate body plan of myxozoans coincided with extreme reduction in genome size and gene loss, while

retaining some genes necessary to function as an obligate parasite. By contrast, *P. hydriforme*, which displays many cnidarian-like morphological features, has a genome size and gene content similar to that of published cnidarian genomes.

MATERIALS AND METHODS

Specimen Collection

Locality information is given in Appendix 1. Actinospores of *Myxobolus cerebralis* (kindly provided by Ron Hedrick at UC Davis) were collected and flash frozen as they emerged from the annelid host (Fig. 2.1). For *Kudoa iwatai* plasmodia was collected from the fish host. *Kudoa iwatai* plasmodia form pseudocysts encapsulated by host cells. Inside the plasmodia, the myxospores are at various stages of maturation (Diamant et al. 2005). Since our RNA extractions were based on several cysts pooled together, we can assume that our RNA data represent all *Kudoa* life stages present in the fish. Unfortunately, the annelid host of *K. iwatai* is unknown. *Polypodium hydriforme* was collected and flash frozen 3-5 days after emerging from the host's oocytes, after it has fragmented into free-living individuals (Fig. 2.1).

Illumina sequencing

For the *M. cerebralis* transcriptome assembly, RNA extraction, library preparation, and sequencing was performed as described for the *P. hydriforme* transcriptome as in (Shpirer et al. 2014). Library preparations and genome sequencing of *P. hydriforme* was carried out at the Genome Sequence Facility at the University of Kansas Medical School (GSF-KUMC). *P. hydriforme* gDNA was sheared to a size of 350bp, and 100pb paired-end (PE) sequencing was performed on an Illumina HiSeq 2000. Library preparation and Illumina 100bp PE HiSeq 2000

sequencing of *K. iwatai* (transcriptome and genome), *S. zaharoni* (genome), and *E. leei* (genome) was described in Shpirer et al. (2014). In addition, to the Illumina HiSeq sequencing, the *K. iwatai* genomic library was also independently sequenced with an Illumina Genome Analyzer IIX platform, which produced 95,434,687 paired reads (76 bp).

Genome and Transcriptome Assemblies

The *M. cerebralis* transcriptome was filtered for read quality and assembled following protocols described for the *P. hydriforme* transcriptome (Shpirer et al. 2014). Genome *de novo* assemblies for *P. hydriforme*, *E. leei*, *S. zaharoni*, and *K. iwatai* were performed with ABySS v. 1.3.6 (Simpson et al. 2009) and the transcriptome *de novo* assembly of *K. iwatai* was performed with Trinity (Haas et al. 2013). Contigs shorter than 500bp and 300bp were removed from the genomic and transcriptomic assemblies, respectively. The accession of the different assemblies and short read data are indicated in Appendix 1.

Host contaminant filtering

To filter the assemblies from fish contaminants, genomic sequences were obtained for *Sparus aurata* (*Enteromyxum leei* and *K. iwatai* host) and *Pterois miles* (*Sphaeromyxa zaharoni* host) as described in (Dray et al. 2016a, b). Blast searches were conducted to eliminate contaminating fish sequences from the genomic assemblies. Specifically, BLASTN (version 2.2.27+) searches were performed for each of the three myxozoans, using the genomic assembly sequences as query against a database of their respective fish host DNA contigs. Sequences of *S. aurata* available in the NCBI dbEST (<http://www.ncbi.nlm.nih.gov/dbEST/>) were also included. The BLASTN parameters that were used are: -e-value 1e-75 and -perc_identity 85 (Altschul et

al. 1997). All sequences which passed this threshold were considered to be contaminants. Further, we performed additional BLASTN searches against the NCBI non-redundant nucleotide (nt) database (e.g. to remove other contaminant such as bacterial sequences).

In order to filter the *K. iwatai* RNA assembly from contaminants (either host RNA or other sample contaminants), we ran RSEM (Li and Dewey 2011) with the default parameters, and filtered the low abundance transcripts using the `filter_fasta_by_rsem_values.pl` script supplied by Trinity (Haas et al. 2013) with default parameters (`--fpkm_cutoff=1200 --isopct_cutoff=1.00`). We then ran BLASTN with `-e-value 1e-75` and `-perc_identity 85` against the sequences of the contaminant database described above (mainly *S. aurata* DNA contigs and ESTs available in the NCBI database). We also added the sequence of *S. aurata* mitochondrial genome (Dray et al. 2016b). All sequences identified were removed. We then performed BLASTN against NCBI nt database on the remaining sequences with `-e-value 1e-75` and `-perc_identity 80`. We removed all the contigs that matched any euteleost sequence with over 80% identity, and other taxa (e.g., fungi, *Drosophila*) with over 90% identity. Finally, we then performed a BLASTN search against the two filtered *K. iwatai* DNA assemblies (HiSeq and GIIx assemblies) with `-e-value 1e-5` threshold, and removed sequences which could not align to any of the DNA assemblies. *P. hydriforme* genomic and transcriptomic sequences were filtered using sequence material from the paddlefish (*Polyodon spathula*) oocyte transcriptome as described in (Shpirer et al. 2014).

Phylogenetic reconstruction

Phylogenetic reconstructions based on the Bayesian and maximum likelihood criteria were performed for different gene and species combinations: 1) A dataset that includes 77

species representative of the animal diversity with their closest outgroups and 200 ribosomal and non-ribosomal protein genes (51,940 amino-acids): 2) A dataset that includes 77 species representative of the animal diversity with their closest outgroups and 128 non-ribosomal protein genes (41,237 amino-acids): 3) A dataset that includes 30 cnidarian species and 200 ribosomal and non-ribosomal protein genes (51,940 amino-acids) and: 4) A dataset that includes 30 cnidarian species and 128 non-ribosomal protein genes (41,237 amino-acids). Additional analyses were also performed excluding either Myxozoa or *P. hydriforme*. For all datasets, Bayesian tree reconstructions were conducted under the CAT model (Lartillot and Philippe 2004) as implemented in Phylobayes MPI vs.1.5 (Lartillot et al. 2013). For the third dataset, the CAT-GTR model, which is more computationally intensive, was also used. For the analyses datasets 1 and 3, two independent chains were run for 10,000 cycles and trees were saved every ten cycles. The first 2,000 trees were discarded (burn-in). For the analysis of the second dataset, the two chains were run for 6,000 generations, sampled every 10 trees, and the first 2,000 trees were discarded. For the analysis of the fourth dataset, the two chains were run for 20,000 generations, sampled every 10 trees, and the first 5,000 trees were discarded. Chain convergence was evaluated with the bpcomp and tracecomp programs of the Phylobayes software. The maximum and average differences, observed at the end of each run were lower than 0.0005 for all analyses. Similarly, the effsize and rel_diff parameters were always higher than 30 and lower than 0.3, respectively, which indicates a correct chain convergence since our analyses investigate the topology rather than branch length and all relevant posterior probabilities = 1. The ML analyses were conducted for each dataset under the PROTGAMMAGTR as implemented in RAxML 8.1.3 (Stamatakis 2006). Bootstrap support was computed after 250 rapid bootstrap

replicates. Alignments and Bayesian tree have been deposited in the TreeBASE repository (Piel et al. 2009) (<http://purl.org/phylo/treebase/phyloids/study/TB2:S17743>).

Analysis of assembly completeness

For each transcriptome and genome assembly, relative completeness was assessed using the Core Eukaryotic Genes Mapping Approach (CEGMA) that searches for the presence of 248 ultra-conserved core eukaryotic genes (CEGs) (Parra et al. 2007). *P. hydriforme* assemblies were run using default settings. Since our evidence indicates that myxozoans have unusually small intron sizes (see below), the `-max_intron_size` parameter for *M. cerebralis* and *K. iwatai* genomic CEGMA runs were adjusted to match the maximum intron size of the *M. cerebralis* intron size distribution (2,630bp).

Estimation of genome size

Output from the CEGMA runs was used for coverage-based estimates of the genome size for *P. hydriforme* and *K. iwatai*. We used the “dna” output files from the genomic CEGMA runs, which include the raw sequence for each region identified by CEGMA as a partial or complete core gene, as well as the 2,000bp of sequence on each side. Because CEGs were chosen to minimize in-paralogy and therefore should be largely single-copy, mapping reads to these regions provides a simple unbiased estimate of genome coverage. For each of the genomes, the complete set of raw reads was mapped to the CEGMA output files using the Burrows-Wheeler Aligner, BWA-MEM (v.0.7.8) with default settings (Li and Durbin 2010). Coverage was calculated using QualiMap 2 (García-Alcalde et al. 2012). Under the assumption that the coverage estimated from the conserved CEGMA-identified regions is representative of genome-

wide coverage, the number of base pairs used in the whole read set for each assembly was divided by the calculated coverage for that species, thus providing an estimate of the genome size for each species (Table 2.2). In addition, an independent approach based on coverage of all contigs was used to estimate the genome size of *K. iwatai* (below). We also conducted independent estimates of *K. iwatai* genome size based on assembly coverage using reads and contig sequences from two independent sequencing runs. We ran CAP3 (Huang and Madan 1999) with the parameters -o 300 -p 90 (300 bp overlap between contigs and 90% identity in the overlapping sequence) on the GIIx *K. iwatai* DNA assembly in order to remove redundant contigs. 952 redundant contigs were removed using CAP3. In order to filter the *K. iwatai* HiSeq DNA reads from contaminants, we created a contaminants database consisting of the *Kudoa* sequences identified as contaminants, the *S. aurata* DNA contigs obtained, as well as all *S. aurata* ESTs available in the NCBI EST database on May 2014. Bowtie2 was used with default settings to align the *K. iwatai* HiSeq DNA reads to the contaminants database. The --un-conc flag was used to save the reads which did not map to the contaminants database to separate paired-end FASTQ files. A total of 164284209 reads remained after this step. We then used Bowtie2 (Langmead and Salzberg 2012) to align the filtered paired-end HiSeq reads to the GIIx assembly, with default parameters. The coverage was calculated using bedtools genomecov -d -ibam (Quinlan and Hall 2010) on the output of Bowtie2. The average coverage-per-position of the GIIx assembly was estimated to be (1391.6, SD: 1095.4, SE=0.2577). The genome size was then estimated by dividing the number of base pair sequenced (filtered reads) by the coverage according to the following formula: $((\text{number of paired reads}) \times (\text{read length}) \times 2) / (\text{average coverage-per-position}) = (163,897,028 \times 100 \times 2) / 1391.6 = 23,555,192$. Using this method, the genome size was thus estimated to be ~23.5 Mbp.

Genome Annotation

Genome annotation was conducted using MAKER2 (Holt and Yandell 2011), incorporating the Semi-HMM based Nucleic Acid Parser (SNAP) gene predictor software, in order to assess gene content of the *K. iwatai* and *P. hydriforme* genome assemblies. For each assembly, MAKER2 was first run using the assembled transcriptome for each species (EST evidence), a file of the core CEGMA proteins, and a random precompiled eukaryotic HMM profile in order to train SNAP. The output of this training was a species-specific HMM profile for each assembly created by SNAP. In the next round, MAKER2 was used to annotate the *K. iwatai* and *P. hydriforme* genomes by using EST evidence, protein evidence for each species, and the species-specific HMM files generated in the training round. The number of genes found by MAKER2 and other annotation statistics were tabulated using the gene-stats function of the SNAP package (Table 2). Mean intron and exon sizes for *Hydra magnipapillata* and *Nematostella vectensis* were calculated from the annotated scaffolds from the Joint Genome Institute (JGI). An independent estimate of intron size for *K. iwatai* was performed by mapping the RNA contigs onto the DNA contigs (a total of 23,393 introns were evaluated). This method revealed a similar mean intron length estimate (i.e. 85.4bps).

Gene enrichment analysis of transcriptomes

Before conducting the gene enrichment analysis, duplicate sequences and alternative transcripts were removed from the transcriptomes of *P. hydriforme* and *M. cerebralis*. No redundant contigs were found in the transcriptome of *K. iwatai*. We used the Trinotate pipeline (Bryant et al. 2017), with default parameters, to annotate the *K. iwatai*, *M. cerebralis*, and *P.*

hydriforme unigenes. In particular, Trinotate searched the Gene Ontology (GO) database (Ashburner et al. 2000) and recovered for each transcript its relevant GO terms. The GO terms of each transcriptome were reduced using the GOSlim list in CateGORizer (Zhi-Liang et al. 2008). The database of *H. magnipapillata* and *N. vectensis* protein sequences was downloaded from the Metazome website (<http://metazome.net/>). GO terms were also assigned to *H. magnipapillata* and *N. vectensis* using Trinotate as described above. The GO terms categories of Myxozoa and non-myxozoan Cnidaria were compared for depletion or enrichment using Fisher's exact tests. The significance level was corrected for multiple testing using a Bonferroni correction (specifically $\alpha=0.05$ was corrected to $\alpha=0.000446$).

Gene enrichment analysis of predicted genes from annotated genomes

The gene contigs longer than 300bps, predicted by MAKER2, for the *K. iwatai* and *P. hydriforme* assembly, were annotated using the Trinotate pipeline. The GO terms provided by the Trinotate annotation were then analyzed and compared to those assigned to the transcriptome of *K. iwatai* and *P. hydriforme* and to the *H. magnipapillata* and *N. vectensis* protein sequences, as described for the gene enrichment analysis of transcriptomes.

Analysis of gene pathways and candidate genes

For each candidate gene, a sequence from *H. magnipapillata* (Supplemental Data Set) was used as a query for performing a tblastx search (e-value cutoff: $1e-03$) against the genome and transcriptome assemblies. In order to confirm their cnidarian identity, assembly sequences with significant hits were then BLAST-searched against the NCBI NR sequence database using the blastx algorithm. In some cases, an additional step using HMMer3 (Eddy 2011), was taken to

double check for conserved elements corresponding to the candidate gene families against the Pfam databases (Finn et al. 2014). We also assessed the completeness of candidate signaling pathways in our assemblies using the Kyoto Encyclopedia of Genes and Genomes (KEGG). Genomic and transcriptomic material was combined into one file per species, and sent through the KEGG Automatic Annotation Server (KAAS) for ortholog assignment and pathway mapping (<http://www.genome.jp/tools/kaas/>). The KAAS assigned KEGG orthology (KO) terms for each species data set using the single-directional best-hit method against a representative eukaryotic data set. After assignment of KO terms, completeness of the candidate pathways as compared to their canonical pathway was assessed and visualized in each species using the KEGG Mapper tool (http://www.genome.jp/kegg/tool/map_pathway1.html).

OrthoMCL Analysis

The OrthoMCL database (Chen et al. 2006) was used in order to determine the number of orthologous groups identified in the transcriptome assemblies of *K. iwatai* and *P. hydriforme*, compared to published predicted proteins in *H. magnipapillata* and *N. vectensis*. Open reading frames were predicted and sequences translated using the OrfPredictor server (Min et al. 2005). Predicted proteins for *H. magnipapillata* were downloaded from NCBI. Each protein FASTA file was uploaded to the OrthoMCL Groups web server (<http://www.orthomcl.org/orthomcl/proteomeUpload.do>). OrthoMCL analysis results in lists of OrthoGroup assignments for each protein in an assembly. These were parsed using R to create lists of unique ortholog group IDs (OGs) found in each assembly. These lists were used as input for the jvenn web-server (<http://bioinfo.genotoul.fr/jvenn/example.html>) (Bardou et al. 2014),

which calculated the overlaps between all combinations of the lists of OGs and created a four-way Venn Diagram to visualize these overlaps (Appendix 6).

RESULTS

Phylogenetic position

A phylogenomic analysis was performed using the newly generated transcriptome assemblies from *K. iwatai*, *M. cerebralis*, and *P. hydriforme*, as well as genomic data of *Enteromyxum leei* and *Sphaeromyxa zaharoni*, in conjunction with published sequences from three additional myxozoans (*Buddenbrockia plumatellae*, *Tetracapsuloides bryosalmonae*, and *Thelohanellus kitauei*), altogether encompassing 22 cnidarians, 38 representatives of the Metazoan diversity, and 9 unicellular opisthokont taxa. Both Bayesian analyses using the CAT model (Lartillot and Philippe 2004) and a Maximum likelihood (ML) analysis using the GTR model recovered *P. hydriforme* as sister to a monophyletic Myxozoa with maximal support (Bayesian posterior probability (PP) of 1.0, ML bootstrap percentage (BP) of 100). Within a monophyletic Cnidaria (PP=1.0, BP=100), the Myxozoa + *P. hydriforme* clade was recovered as sister to the medusozoan clade with maximal support (PP=1.0, BP=100) (Fig. 2.2). The Bayesian and ML phylogeny topologies only differed in the position of two taxa (*Porites* and *Strigamia*). Several analyses were conducted to evaluate the robustness of the position of Myxozoa + *P. hydriforme* clade within Cnidaria. Because it has been claimed that ribosomal genes can contain a different signal from non-ribosomal genes (Nosenko et al. 2013), phylogenetic analyses were conducted on a dataset of 41,237 amino acids, excluding ribosomal genes (Appendix 2). Additionally, because taxon sampling can affect phylogenetic inferences, phylogenetic reconstructions were performed with either only cnidarian taxa (Appendix 3) or after removing

either Myxozoa or *P. hydriforme*. None of these analyses affected the position of Myxozoa and *P. hydriforme*, or its position as the sister clade to Medusozoa.

Estimation of the completeness of genome and transcriptome assemblies

RNA libraries from *M. cerebralis*, *K. iwatai*, and *P. hydriforme*, and DNA libraries from the latter two were sequenced using a short read *Illumina* platform. Data were deposited in NCBI archives (Appendix 1). Previously generated *M. cerebralis* genomic data (Nesnidal et al. 2013) were downloaded from NCBI (SRX208206). Assembly statistics are shown in Table 2.1 and size distribution of the transcriptome sequences are shown in Appendix 4. Completeness of the genome and transcriptome assemblies were estimated by determining the presence of the 248 ultra-conserved core eukaryotic genes (CEGs), obtained from the CEGMA database (Parra et al. 2009) (Table 2.1). The *K. iwatai* genome and transcriptome assemblies recovered over 70% of the CEGs, with over 1,000X estimated mean base-pair coverage. The *M. cerebralis* transcriptome was less complete, recovering only 39% of the CEGs. We were unable to recover any CEGs for *M. cerebralis* from its published genomic data, most likely due to their low coverage. Although the *P. hydriforme* genome assembly recovered very few CEGs due to low coverage, its transcriptome assembly recovered 90% of complete CEGs.

General characteristics of genomes

Genome size estimates based on overall coverage of known individual genes are shown in Table 2.2. These estimates suggest that the myxozoan genome of *K. iwatai* (22.5 Mb) is one of the smallest reported animal genomes, comparable to the genome of the recently reported parasitic nematode (~20 Mb) (Burke et al. 2015). The *K. iwatai* genome is more than 20-fold

smaller than the estimated size of the *P. hydriforme* genome (561 Mb) and the published genome of the cnidarian *Nematostella vectensis* (450Mb) (Putnam et al. 2007), and more than 40-fold smaller than the published estimated genome size of the cnidarian, *Hydra magnipapillata* (1,005Mb) (Chapman et al. 2010). Although the published estimated genome size of the myxozoan *Thelohanellus kitauei* (188.5 Mb), which was based on K-mer distribution (Yang et al. 2014), is 8-fold larger than our estimated *K. iwatai* genome size, it is still significantly smaller than the non-myxozoan cnidarian genomes. As an independent test of the accuracy of genome-size estimation, we compared genome size based on overall assembly coverage's of the *K. iwatai* genome, from two independent sequencing runs. This revealed a very similar genome-size estimate (23.5 Mb). Due to the low coverage of the published *M. cerebralis* genomic read data, it was not possible to estimate its genome size.

The number of protein coding genes and average intron and exon sizes were estimated from the genome assemblies using the MAKER2 genome annotation package (Holt and Yandell 2011). These analyses revealed that the number of protein coding genes in the *K. iwatai* genome (5,533) is less than 30% of those estimated in *P. hydriforme* (17,440), *H. magnipapillata* (16,839), and *N. vectensis* (18,000) (Table 2.2). In addition, the myxozoan genome appears to be much more compact, with a mean intron size of 82bps in *K. iwatai*, compared to 1,163bps in *P. hydriforme*, 2,673bps in *H. magnipapillata*, and 799bps in *N. vectensis* (Table 2.2).

Characteristics of transcriptomes: Comparisons of gene ontology and gene orthology

To identify the biological pathways that have gained or lost a significant fraction of expressed genes in the myxozoans *M. cerebralis*, *K. iwatai*, and *P. hydriforme* when compared to the cnidarian model species, *H. magnipapillata* and *N. vectensis*, Fisher's exact tests were used

to infer enrichment and depletion in the proportion of genes present in 112 Gene Ontology (GO) categories as defined by the GOSlim list of CateGORizer (Zhi-Liang et al. 2008) (Table 2.3, Supplemental Datasets). Since the GO terms of *P. hydriforme* were more similar to *N. vectensis* and *H. magnipapillata* than to the myxozoans *M. cerebralis* and *K. iwatai*, the most informative comparison was between *M. cerebralis* and *K. iwatai* versus *P. hydriforme*, *H. magnipapillata* and *N. vectensis* (Supplemental Dataset). Of the top 20 GO categories with the highest occurrences of GO terms (Fig. 2.3), the expressed myxozoan genes appear to be significantly depleted (by comparison with other cnidarians) in categories that are related to development, cell differentiation, and cell-to-cell communication (Table 2.3), consistent with lack of a complex multi-cellular body in myxosporean myxozoans. By contrast, myxozoan expressed genes have an abundance of categories such as cellular function, for which the number of genes does not differ significantly from the number observed in cnidarians (e.g., a similar number of nucleoplasm genes was found in both, Supplemental Dataset). Although these analyses were from transcriptomes, the general patterns likely reflect overall genome content as multiple life cycle stages are represented in the combined transcriptome of *K. iwatai* and *M. cerebralis*. To confirm this, we also performed a GO comparison analysis of the genes predicted based on genomic sequences, which revealed the same general patterns for *K. iwatai*, but not for *P. hydriforme* whose transcriptome assembly was of better quality than its genome assembly (Appendix 5).

Using the OrthoMCL database (Chen et al. 2006), we determined the number of orthologous groups (OG) that could be identified from our transcriptome assemblies of *P. hydriforme* and *K. iwatai*, compared to published predicted proteins from *H. magnipapillata* (Chapman et al. 2010) and *N. vectensis* (Putnam et al. 2007) (Appendix 6). A total of 8,021 unique OGs were recovered from *H. magnipapillata*, 11,162 from *N. vectensis*, 5,451 from *P.*

hydriforme and 2,735 from *K. iwatai*. Although there was more overlap in the number of OGs between *H. magnipapillata*, *N. vectensis* and *P. hydriforme* than between *K. iwatai* and the other three, this is consistent with the significantly lower number of total OGs in *K. iwatai* (Appendix 6).

Analyses of gene pathways and candidate genes in the assemblies

We searched for several candidate genes and gene pathways that have been previously characterized as important for cnidarian cell signaling and development, in the transcriptome and genome assemblies of *M. cerebralis*, *K. iwatai*, and *P. hydriforme*. Using BLAST searches and KEGG pathway analyses we determined presence/absence of representatives within a gene family, but not precise orthology, within the particular families. Myxozoan genomes appear to lack key genes and signaling pathways that are present in *P. hydriforme* and other cnidarians (Table 2.4). Specifically, the conserved transcriptional factors belonging to the Hox and Runx gene families, that have been shown to be important for cnidarian patterning and cellular differentiation, respectively (Ryan et al. 2006; Sullivan et al. 2008), were found in neither the genomes nor transcriptomes of myxozoans, but were nevertheless present in *P. hydriforme* (Table 2.4). In addition, myxozoans appear to have lost the ligands, receptors, and most downstream elements of the Wnt and Hedgehog signaling pathways, which have been shown to be important for axial patterning (Hensel et al. 2014), and cell signaling (Matus et al. 2008), respectively; whereas nearly all components of these pathways were recovered in *P. hydriforme* (Table 2.4). By contrast, myxozoans and *P. hydriforme* possess orthologs to the stem-cell markers FoxO (Boehm et al. 2012) and Piwi (Juliano et al. 2014) and *P. hydriforme* and *K. iwatai* appear to have the gene Hap2, which was shown to be involved in gamete fusion in *Hydra*

(Steele and Dana 2009). (Table 2.4, Supplemental Dataset). The two myxozoans and *P. hydriforme* were also found to have key elements of the Notch signaling pathway, and *M. cerebralis* possessed some elements of the TGF β pathway (Table 2.4, Supplemental Dataset). Notch is reported to have an important role in differentiation of stem-cell lineages (Käsbauer et al. 2007), whereas TGF β appears to play a more general role in cell signaling (Hobmayer et al. 2001).

DISCUSSION

Our analyses of transcriptomic and genomic assemblies of myxozoans have yielded significant insight into the evolution of these microscopic parasites from free-living cnidarians. We report for the first time a broad phylogenomic sampling of myxozoans, including representatives from the malacosporean clade, and the fresh water and marine myxosporean clades (Fiala and Bartošová 2010), as well as the only phylogenomic study to date to include *P. hydriforme*. In addition, we have a more comprehensive sampling of cnidarians than previous phylogenomic studies, addressing the placement of myxozoans (Jimenez-Guri et al. 2007, Nesnidal et al. 2013, Feng et al. 2014). We recover *P. hydriforme* as the sister taxon to Myxozoa and can confirm, with an increased sampling and thus a higher degree of confidence, the placement of this clade as the sister taxon to medusozoan cnidarians. These results are consistent with those of other molecular phylogenetic studies (Siddall et al. 1995; Zrzavý and Hypša 2003), although these have been criticized as possible artifacts of long-branch attraction (Evans et al. 2008). The monophyly of Myxozoa + *P. hydriforme* is also supported by endoparasitism in fish, a unique cell-within-cell developmental stage, possession of a single similar type of nematocysts (Adamczyk et al. 2008), and similarity in minicollagen sequences (Shpirer et al. 2014). This

phylogenomic pattern suggests that endoparasitism in Cnidaria was a single event that occurred at the base of Myxozoa + *P. hydriforme*, but that the dramatic reduction in body plan occurred following the divergence of *P. hydriforme* from myxozoans, as *P. hydriforme* retains many cnidarian-features.

The Myxozoa represent an extreme example of degeneration of body plans due to parasitism. Genome and transcriptome analyses reveal this degeneration was accompanied by massive genome reduction, with myxozoans having one of the smallest reported animal genomes. Genome size reduction included loss of many genes considered hallmarks of metazoan development, yet retention of genes necessary to function as obligate parasites, such as nematocyst specific genes. In contrast to myxozoans, *P. hydriforme* has a genome similar in size, gene number, and gene content to the model system *Hydra*. This finding is not surprising given that although *P. hydriforme* is an obligate parasite, it has maintained its cnidarian-like body plan, including epithelia, mouth, gut and tentacles. Our study provides a robust phylogenetic hypothesis for myxozoan placement within Cnidaria, as the sister taxon to *P. hydriforme*, and a framework for comparative genomic studies, which should be valuable for future phylogenetic and genomic investigations of Cnidaria *sensu lato*.

Note about data accessibility

Items marked “Supplemental Dataset” are either too large or in an inappropriate format for inclusion in the dissertation itself. Please find these data sets as supplemental material to Chang et al. 2015: <https://doi.org/10.1073/pnas.1511468112>.

TABLES

Table 2.1: Assembly statistics for sequenced genomes and transcriptomes.

	<i>Kudoa iwatai</i>		<i>Myxobolus cerebralis</i>		<i>Polypodium hydriforme</i>	
	Genome	Transcriptome	Genome ¹	Transcriptome	Genome	Transcriptome
Raw Reads	167917062	154253215	N/A	312202378	229917588	467431688
Contigs	1637	6528	N/A	52972	83415	24523
N50	40195	1662	N/A	994	3865	1475
CEGs(c)/% ²	179/72	190/77		97/39	14/6	223/90
CEGs(p)/% ²	188/76	208/84		164/66	56/23	232/94

¹Published ESTS. ²Number/percentage out of 248 ultra-conserved core eukaryotic genes (CEGs); (c) = complete, (p)=complete+partial.

Table 2.2: Estimated Genome Characteristics

	<i>Kudoa iwatai</i>	<i>Polypodium</i>	<i>Hydra</i> ₁	<i>Nematostella</i> ²
Genome size (Mb)	22.5	561	1005	450
No. protein genes	5,533	17,440	16,839	18,000
GC content (%)	28	47	29	41
Mean intron size (bps)	82	1,163	2,673	799
Mean exon size (bps)	102	216	218	208

¹Chapman et al. (2010)

²Putnam et al. (2007)

Table 2.3: Gene Ontology categories from transcriptomes showing depletion in myxozoans compared to other cnidarians

Category	Myxozoa ¹ /%	Cnidaria ² /%
cell differentiation	548/0.0234	1763/0.0302
development	1052/0.0450	3892/0.0667
morphogenesis	456/.0195	1693/0.0290
receptor activity	46/0.00197	310/0.00531
signal transducer activity	49/0.0002	328/0.0056

¹Myxozoa = *Kudoa iwatai* + *Myxobolus cerebralis*. ²Cnidaria = *Hydra magnipillata* + *Nematostella vectensis* + *Polypodium hydriforme*. Myxozoan total number = 23,388. Cnidarian total number = 53,354. All categories had a P value = 0.0001.

Table 2.4: Presence (X) or absence (-) of genes and KEGG pathways that have been characterized in other cnidarians

	Genes					KEGG pathways			
	Hox-like	Runx	Piwi	FoxO	Hap2	Wnt	Hedgeh	TGFB	Notch
<i>K. iwatai</i>			X	X	X				X
<i>M. cerebralis</i>			X	X				X	X
<i>Polypodium</i>	X	X	X	X	X	X	X	X	X

FIGURES

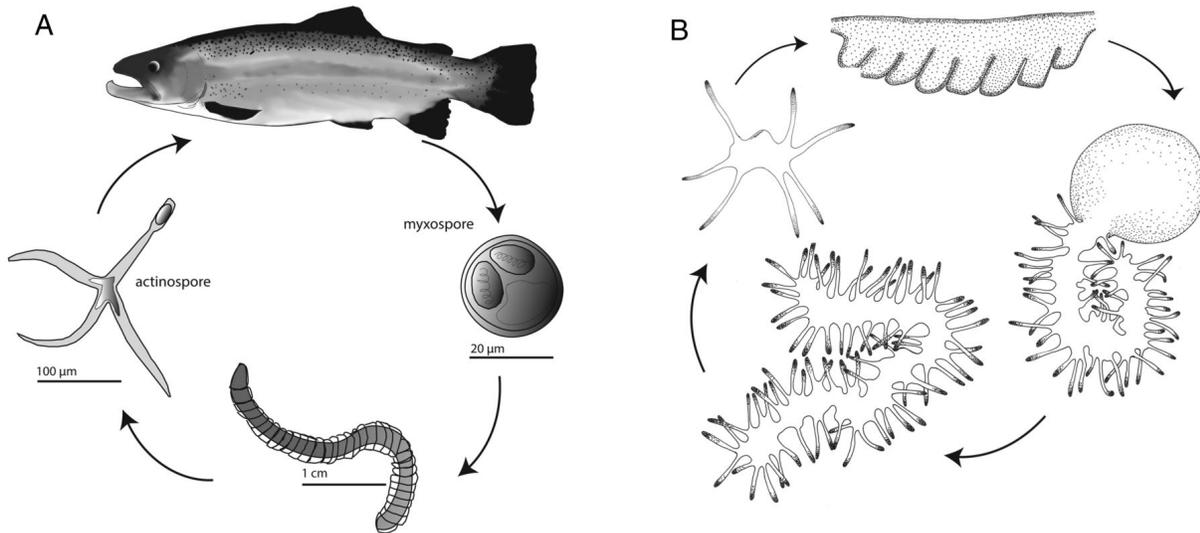


Figure 2.1: Life cycles of *Myxobolus cerebralis* and *Polypodium hydriforme*.

(A) *M. cerebralis* alternates its development between a fish (salmonid) host and an annelid (*Tubifex tubifex*) host. The myxospore is produced in the fish (right) and the actinospore is produced in the annelid (left). Both stages consist of just a few cells, including those housing polar capsules. (B) In *P. hydriforme*, the stolon stage (top) develops inside of the ovaries of its host (acipenseriform fish). Upon host spawning, *P. hydriforme* emerges from the host's oocyte (right), fragments and lives as a free-living stage with a mouth (left) before infecting its host.

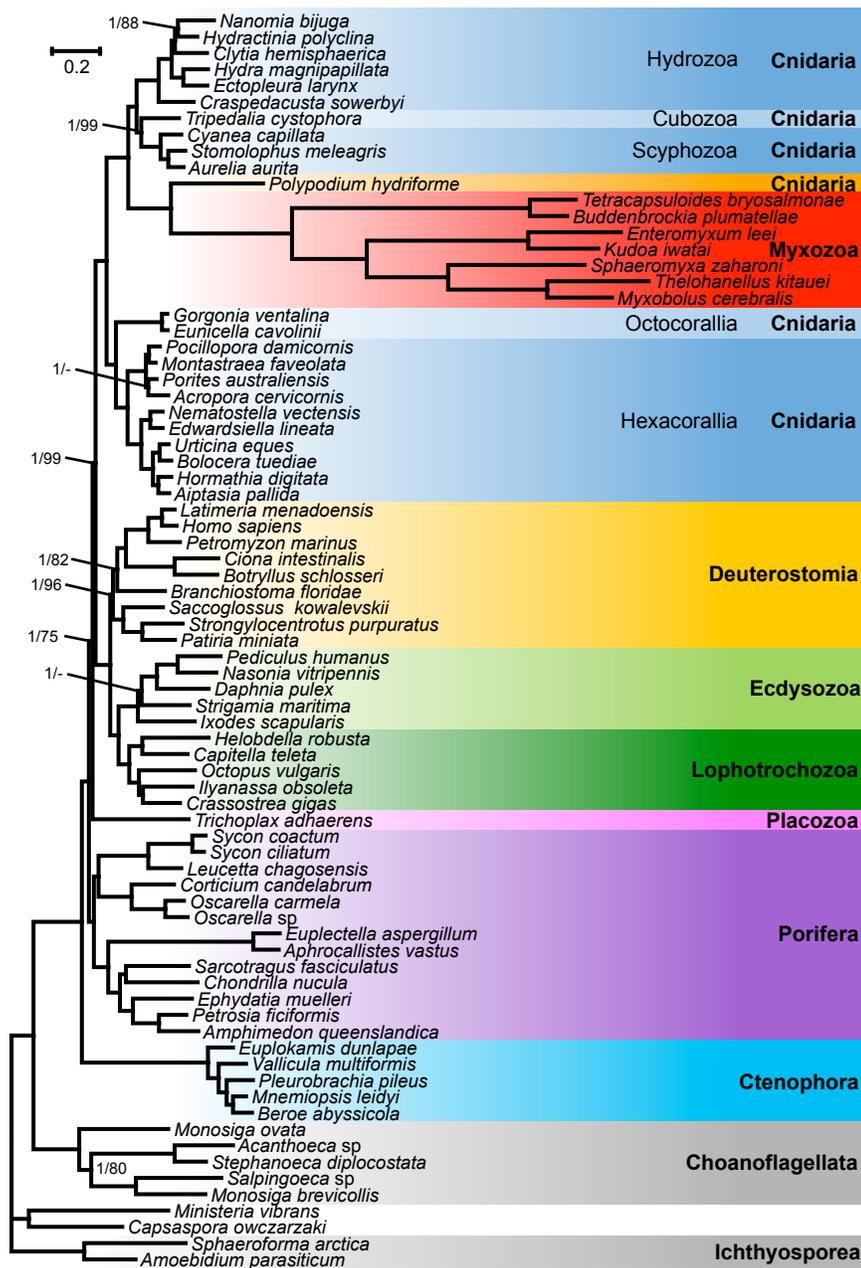


Figure 2.2: Phylogenetic tree generated from a matrix of 51,940 amino-acid positions and 77 taxa using Bayesian inference under the CAT model.

Support values are only indicated for nodes that did not received maximal support. Bayesian posterior probabilities/ML bootstrap supports under the PROTGAMMAGTR ML bootstrap supports are given near the corresponding node. “-” indicates that the corresponding node is absent from the ML bootstrap consensus tree.

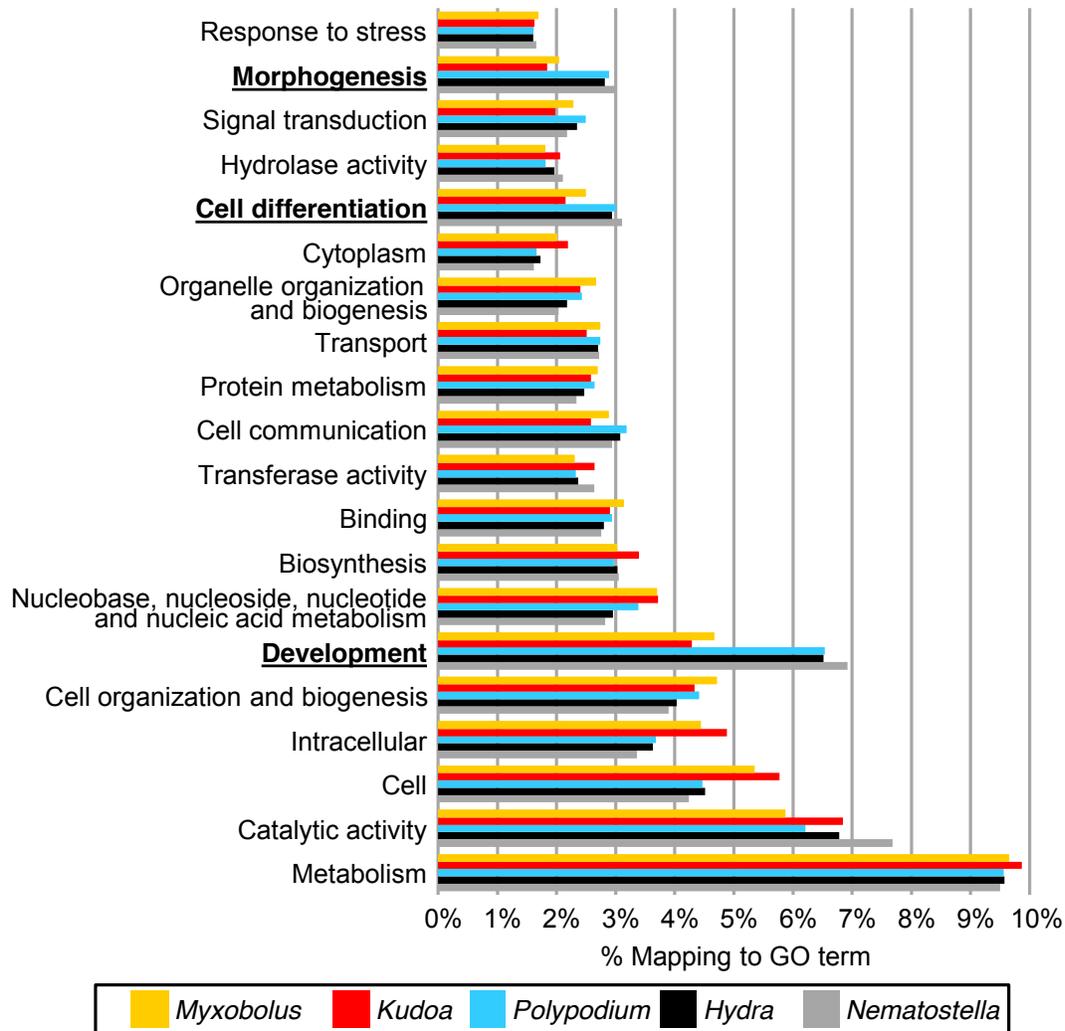


Figure 2.3: GO annotation of unigenes in transcriptomes.

The top 20 gene ontology (GO) categories are shown as a percentage of total GO terms from the transcriptome assemblies of *M. cerebralis*, *K. iwatai*, *P. hydriforme*, and the published protein sequences of *H. magnipapillata* and *N. vectensis*. Categories for which myxozoans present significantly less GO terms than other cnidarians are indicated in bold.

LITERATURE CITED

- Adamczyk, P., S. Meier, T. Gross, B. Hobmayer, S. Grzesiek, H. P. Bächinger, et al. 2008. Minicollagen-15, a Novel Minicollagen Isolated from Hydra, Forms Tubule Structures in Nematocysts. *Journal of Molecular Biology* 376:1008-1020.
- Altschul, S. F., T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25:3389-3402.
- Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, et al. 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics* 25:25-29.
- Bardou, P., J. Mariette, F. Escudie, C. Djemiel, and C. Klopp. 2014. jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics* 15:293.
- Boehm, A.-M., K. Khalturin, F. Anton-Erxleben, G. Hemmrich, U. C. Klostermeier, J. A. Lopez-Quintero, et al. 2012. FoxO is a critical regulator of stem cell maintenance in immortal Hydra. *Proceedings of the National Academy of Sciences* 109:19697-19702.
- Bryant, D. M., K. Johnson, T. DiTommaso, T. Tickle, M. B. Couger, D. Payzin-Dogru, et al. 2017. A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. *Cell Reports* 18:762-776.
- Burke, M., E. H. Scholl, D. M. Bird, J. E. Schaff, S. Coleman, R. Crowell, et al. 2015. The plant parasite *Pratylenchus coffeae* carries a minimal nematode genome. Brill.
- Bütschli, O. 1881. Myxosporidien. *Zoologischer Jahrbuch für 1*:162-164.
- Chapman, J. A., E. F. Kirkness, O. Simakov, S. E. Hampson, T. Mitros, T. Weinmaier, et al. 2010. The dynamic genome of Hydra. *Nature* 464:592-596.
- Chen, F., A. J. Mackey, C. J. Stoeckert, and D. S. Roos. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Research* 34:D363-D368.
- Diamant, A., M. Ucko, I. Paperna, A. Colorni, and A. Lipshitz. 2005. *Kudoa iwatai* (myxosporidia: multivalvulida) in wild and cultured fish in the red sea: redescription and molecular phylogeny. *Journal of Parasitology* 91:1175-1189.
- Dray, L., M. Neuhof, A. Diamant, and D. Huchon. 2016a. The complete mitochondrial genome of the devil firefish *Pterois miles* (Bennett, 1828) (Scorpaenidae). *Mitochondrial DNA A DNA Mapp Seq Anal.* 27:783-784.
- Dray, L., M. Neuhof, A. Diamant, and D. Huchon. 2016b. The complete mitochondrial genome of the gilthead seabream *Sparus aurata* L. (Sparidae). *Mitochondrial DNA A DNA Mapp Seq Anal.* 27:781-782.
- Eddy, S. R. 2011. Accelerated Profile HMM Searches. *PLoS Comput Biol* 7:e1002195.
- El-Matbouli, M., R. W. Hoffmann, and C. Mandok. 1995. Light and electron-microscopic observations on the route of the triactinomyxon sporoplasm of the myxobolus cerebralis from epidermis into rainbow trout cartilage. *Journal of Fish Biology* 46:919-935.
- Evans, N., A. Lindner, E. Raikova, A. Collins, and P. Cartwright. 2008. Phylogenetic placement of the enigmatic parasite, *Polypodium hydriforme*, within the Phylum Cnidaria. *BMC Evolutionary Biology* 8:139.
- Evans, N. M., Holder, M.T., Barbeitos, M.S., Okamura, B., and Cartwright, P. 2010. The phylogenetic position of Myxozoa: Exploring conflicting signals in phylogenomic and ribosomal datasets. *Molecular Biology and Evolution* 19:968-971.

- Feng, J.-M., J. Xiong, J.-Y. Zhang, Y.-L. Yang, B. Yao, Z.-G. Zhou, et al. 2014. New phylogenomic and comparative analyses provide corroborating evidence that Myxozoa is Cnidaria. *Molecular Phylogenetics and Evolution* 81:10-18.
- Fiala, I. and P. Bartošová. 2010. History of myxozoan character evolution on the basis of rDNA and EF-2 data. *BMC Evolutionary Biology* 10:228.
- Finn, R. D., A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, et al. 2014. Pfam: the protein families database. *Nucleic Acids Research* 42:D222-D230.
- Foxx, J. and M. E. Siddall. 2015. The Road To Cnidaria: History of Phylogeny of the Myxozoa. *Journal of Parasitology* 101:269-274.
- García-Alcalde, F., K. Okonechnikov, J. Carbonell, L. M. Cruz, S. Götz, S. Tarazona, et al. 2012. Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* 28:2678-2679.
- Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protocols* 8:1494-1512.
- Hartikainen, H., A. Gruhl, and B. Okamura. 2014. Diversification and repeated morphological transitions in endoparasitic cnidarians (Myxozoa: Malacosporea). *Molecular Phylogenetics and Evolution* 76:261-269.
- Hedrick, R. P., M. El-Matbouli, M. A. Adkison, and E. MacConnell. 1998. Whirling disease: re-emergence among wild trout. *Immunol. Rev.* 166:365-376.
- Hensel, K., T. Lotan, S. M. Sanders, P. Cartwright, and U. Frank. 2014. Lineage-specific evolution of cnidarian Wnt ligands. *Evolution & Development* 16:259-269.
- Hobmayer, B., F. Rentzsch, and T. Holstein. 2001. Identification and expression of HySmad1, a member of the R-Smad family of TGF β signal transducers, in the diploblastic metazoan Hydra. *Dev Genes Evol* 211:597-602.
- Hoeg, J. T. 1995. The biology and life cycle of the Rhizocephala (Cirripedia). *Journal of the Marine Biological Association of the United Kingdom* 75:517-550.
- Holland, J. W., B. Okamura, H. Hartikainen, and C. J. Secombes. 2011. A novel minicollagen gene links cnidarians and myxozoans. *Proceedings of the Royal Society B-Biological Sciences* 278:546-553.
- Holt, C. and M. Yandell. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *Bmc Bioinformatics* 12.
- Huang, X. and A. Madan. 1999. CAP3: A DNA Sequence Assembly Program. *Genome Research* 9:868-877.
- Hwang, J. S., Y. Takaku, T. Momose, P. Adamczyk, S. Lñzbek, K. Ikeo, et al. 2010. Nematogalectin, a nematocyst protein with GlyXY and galectin domains, demonstrates nematocyte-specific alternative splicing in Hydra. *Proceedings of the National Academy of Sciences* 107:18539-18544.
- Jiménez-Guri, E., H. Philippe, B. Okamura, and P. W. H. Holland. 2007. Buddenbrockia Is a Cnidarian Worm. *Science*.
- Juliano, C. E., A. Reich, N. Liu, J. G \ddot{u} tzfried, M. Zhong, S. Uman, et al. 2014. PIWI proteins and PIWI-interacting RNAs function in Hydra somatic stem cells. *Proceedings of the National Academy of Sciences* 111:337-342.
- Käsbauer, T., P. Towb, O. Alexandrova, C. N. David, E. Dall'Armi, A. Staudigl, et al. 2007. The Notch signaling pathway in the cnidarian Hydra. *Developmental Biology* 303:376-390.

- Kent, M. L., A. K. B., J. L. Bartholomew, M. El-Matbouli, S. S. Desser, R. H. Devlin, et al. 2001. Recent advances in our knowledge of the Myxozoa. *J. of Euk. Micro.* 48:395–413.
- Kobayashi, M., H. Furuya, and P. W. H. Holland. 1999. Evolution: Dicyemids are higher animals. *Nature* 401:762-762.
- Langmead, B. and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Meth* 9:357-359.
- Lartillot, N. and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* 21:1095-1109.
- Lartillot, N., N. Rodrigue, D. Stubbs, and J. Richer. 2013. PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. *Systematic Biology* 62:611-615.
- Li, B. and C. Dewey. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323.
- Li, H. and R. Durbin. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26:589-595.
- Lipin, A. 1925. Geschlechtliche Form, Phylogenie und systematische Stellung von Polypodium hydriforme Ussov. *Zool Jahrb Anat*:541-635.
- Lom, J. and I. Dykova. 2006. Myxozoan genera: definition and notes on taxonomy, life-cycle terminology and pathogenic species. *Folia Parasitologica* 53:1-36.
- Matus, D. Q., C. R. Magie, K. Pang, M. Q. Martindale, and G. H. Thomsen. 2008. The Hedgehog gene family of the cnidarian, *Nematostella vectensis*, and implications for understanding metazoan Hedgehog pathway evolution. *Developmental Biology* 313:501-518.
- Min, X. J., G. Butler, R. Storms, and A. Tsang. 2005. OrfPredictor: predicting protein-coding regions in EST-derived sequences. *Nucleic Acids Research* 33:W677-W680.
- Monteiro, A. S., B. Okamura, and P. W. H. Holland. 2002. Orphan Worm Finds a Home: *Buddenbrockia* is a Myxozoan. *Mol. Biol. Evol.* 19: 968-971 19:968-971.
- Nesnidal, M. P., M. Helmkampf, I. Bruchhaus, M. El-Matbouli, and B. Hausdorf. 2013. Agent of Whirling Disease Meets Orphan Worm: Phylogenomic Analyses Firmly Place Myxozoa in Cnidaria. *PLoS ONE* 8:e54576.
- Nosenko, T., F. Schreiber, M. Adamska, M. Adamski, M. Eitel, J. Hammel, et al. 2013. Deep metazoan phylogeny: When different genes tell different stories. *Molecular Phylogenetics and Evolution* 67:223-233.
- Okamura, B. and A. Gruhl. 2015. Myxozoan Affinities and Route to Endoparasitism. Pp. 23-44. *Myxozoan Evolution, Ecology and Development*. Springer.
- Parra, G., K. Bradnam, and I. Korf. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23:1061-1067.
- Parra, G., K. Bradnam, Z. Ning, T. Keane, and I. Korf. 2009. Assessing the gene space in draft genomes. *Nucleic Acids Research* 37:289-297.
- Piel, W. H., L. Chan, M. J. Dominus, J. Ruan, R. A. Vos, and V. Tannen. 2009. TreeBASE v. 2: a database of phylogenetic knowledge.
- Putnam, N. H., M. Srivastava, U. Hellsten, B. Dirks, J. Chapman, A. Salamov, et al. 2007. Sea Anemone Genome Reveals Ancestral Eumetazoan Gene Repertoire and Genomic Organization. *Science* 317:86-94.
- Quinlan, A. R. and I. M. Hall. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841-842.

- Raikova, E. V. 1988. On the systematic position of *Polypodium hydriforme* Ussov, (Coelenterata). Pp. 116-122 in V. M. Koltum, and S. D. Stepanjants, eds. Porifera and Cnidaria. Contemporary state and perspectives of investigations. Zoological Institute of Academy of Sciences of USSR, Leningrad.
- Raikova, E. V., V. C. Suppes, and G. L. Hoffmann. 1979. The parasitic coelenterate, *Polypodium hydriforme* Ussov, from the eggs of the American acipenseriform *Polyodon spathula*. *Journal of Parasitology* 65:804-810.
- Reft, A. J. and M. Daly. 2012. Morphology, distribution, and evolution of apical structure of nematocysts in hexacorallia. *Journal of Morphology* 273:121-136.
- Ryan, J. F., P. M. Burton, M. E. Mazza, G. K. Kwong, J. C. Mullikin, and J. R. Finnerty. 2006. The cnidarian-bilaterian ancestor possessed at least 56 homeoboxes: evidence from the starlet sea anemone, *Nematostella vectensis*. *Genome Biology* 7.
- Shpirer, E., E. S. Chang, A. Diamant, N. Rubinstein, P. Cartwright, and D. e. Huchon. 2014. Diversity and evolution of myxozoan minicollagens and nematogalectins. *BMC Evolutionary Biology* 14.
- Siddall, M. E., D. S. Martin, D. Bridge, D. M. Cone, and S. S. Desser. 1995. The demise of a phylum of protists. Phylogeny of Myxozoa and other parasitic cnidaria. *J. Parasitol.* 81:961-967.
- Siddall, M. E. and M. F. Whiting. 1999. Long-branch abstractions. *Cladistics* 15:9-24.
- Simpson, J. T., K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, and Í. Birol. 2009. ABySS: A parallel assembler for short read sequence data. *Genome Research* 19:1117-1123.
- Smothers, J. F., C. D. von Dohlen, L. H. Smith, and R. D. Spall. 1994. Molecular evidence that the myxozoan protists are metazoans. *Science*:1719-1721.
- Stamatakis, A. 2006. RAxML-VI-HPC: Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. *Bioinformatics* 22:2688-2690.
- Steele, R. E. and C. E. Dana. 2009. Evolutionary History of the HAP2/GCS1 Gene and Sexual Reproduction in Metazoans. *PLoS ONE* 4:e7680.
- Štolc, A. 1899. Actinomyxidies, nouveau groupe de Mesozoaires parent des Myxosporidies. *Bull. Intl. Acad. Sci. Boheme* 22:1-12.
- Sullivan, J., D. Sher, M. Eisenstein, K. Shigesada, A. Reitzel, H. Marlow, et al. 2008. The evolutionary origin of the Runx/CBFBeta transcription factors - Studies of the most basal metazoans. *BMC Evolutionary Biology* 8:228.
- Weill, R. 1938. L'interpretation des Cnidosporidies et la valeur taxonomique de leur cnidome. Leur cycle comparé à la phase larvaire des Narcomeduses Cuninides. *Travaux de la Station Zoologique de Wimereaux* 13:727-744.
- Wolf, K. and M. E. Markiw. 1984. Biology Contravenes Taxonomy in the Myxozoa: New Discoveries Show Alternation of Invertebrate and Vertebrate Hosts. *Science* 225:1449-1452.
- Yang, Y., J. Xiong, Z. Zhou, F. Huo, W. Miao, C. Ran, et al. 2014. The Genome of the Myxosporean *Thelohanellus kitauei* Shows Adaptations to Nutrient Acquisition within Its Fish Host. *Genome Biology and Evolution* 6:3182-3198.
- Zhi-Liang, H., B. Jie, and J. M. Reecy. 2008. CateGORizer: A Web-Based Program to Batch Analyze Gene Ontology Classification Categories. *Online Journal of Bioinformatics* 9:108-112.

Zrzavý, J. and V. Hypša. 2003. Myxozoa, *Polypodium*, and the origin of the Bilateria: The phylogenetic position of "Endocnidozoa" in the light of the rediscovery of Buddenbrockia. *Cladistics* 19:164-169.

Chapter 3

Non-clonal coloniality: Genetically chimeric colonies through fusion of sexually produced polyps in the hydrozoan *Ectopleura larynx*

ABSTRACT

Hydrozoans typically develop colonies through asexual budding of polyps. Although colonies of *Ectopleura* are similar to other hydrozoans in that they consist of multiple polyps physically connected through continuous epithelia and shared gastrovascular cavity, *Ectopleura larynx* does not asexually bud polyps indeterminately. Instead, after an initial phase of limited budding in a young colony, *E. larynx* achieves its large colony size through the aggregation and fusion of sexually (non-clonally) produced polyps. The apparent chimerism within a physiologically integrated colony presents a potential source of conflict between distinct genetic lineages, which may vary in their ability to access the germline. In order to determine the extent to which the potential for genetic conflict exists, we characterized the types of genetic relationships between polyps within colonies, using a RAD-Seq approach. Our results indicate that *E. larynx* colonies are indeed comprised of polyps that are clones and sexually reproduced siblings and offspring, consistent with their life history. In addition, we found that colonies also contain polyps that are less genetically related, and that estimates of genome-wide relatedness suggests a potential for conflict within a colony. Taken together, our data suggests that there are distinct categories of relationships in colonies of *E. larynx*, likely achieved through a range of processes including budding, regeneration and fusion of progeny and unrelated polyps, with the possibility for a genetic conflict resolution mechanism. Together these processes contribute to the re-evolution of the ecologically important trait of coloniality in *E. larynx*.

INTRODUCTION

Coloniality is a key evolutionary innovation which confers a strong advantage over solitary organisms in substrate-limited marine environments by allowing for rapid colonization and spread over available substratum (Jackson 1977; Coates 1985). Amongst most hydrozoans (phylum Cnidaria) coloniality is achieved through asexual budding of polyps which remain physically attached by continuous epithelia and a shared gastrovascular cavity. Recent findings reported that the hydrozoan *Ectopleura larynx* Ellis and Solander, 1786 does not asexually bud polyps indeterminately, but rather achieves large colony size through the aggregation and fusion of sexually produced offspring (Nawrocki and Cartwright 2012) (Figure 3.1). Not only do the sexually-produced offspring settle upon established colonies, Nawrocki and Cartwright (2012) demonstrated that the epithelia of the polyp and the colony eventually become fused such that the gastrovascular cavity is shared throughout the entire colony. The end result is indistinguishable from other hydrozoans that achieve this level of integration through asexual budding. The formation of colonies through an amalgamation of sexually reproduced polyps in *E. larynx* has important evolutionary and genetic consequences. If individual colonies of *Ectopleura larynx* are mixtures of genotypes, then a potential source of conflict exists between distinct genetic lineages, which may vary in their ability to access the germline and achieve representation in the gametes produced by the colony. The most extreme form of this conflict is germline parasitism where one lineage monopolizes the reproductive output while contributing only partially to the somatic functioning of the colony (Buss 1982). Experimental evidence of successful germline parasitism by a particular lineage has been noted in the colonial tunicate *Botryllus schlosseri* (Stoner and Weissman 1996; Stoner et al. 1999) and in the social slime mold *Dictyostelium* (Buss 1982; Noce and Takeuchi 1985; Ennis et al. 2000).

Alternatively, the germline of a colony formed by fusion of polyps may display germline chimerism in which multiple germline lineages persist and potentially produce gametes, leading to conflict over allocation of reproductive resources to gametes formed from different genotypes. There is evidence of the existence of multiple genotypes in gonad tissue after the fusion of two colonies in *Botryllus schlosseri* (Pancer et al. 1995; Carpenter et al. 2011) and of the long-term persistence of chimerism in colonies of some corals (Puill-Stephan et al. 2009). Given that *E. larynx* colonies are formed by a process that includes polyp fusion, either of the above scenarios of conflict may exist within an *E. larynx* colony.

There are many examples of selection for mechanisms to keep colonies genetically homogenous in order to prevent competition between genetic lineages (i.e. to not allow multiple genotypes to produce gametes), especially for benthic, sessile organisms where different colonies may unavoidably come in contact with one another and need to prevent fusion (Buss 1987). Indeed, genetically-encoded allorecognition systems, which only allow colony fusion within certain bounds of relatedness, are well-defined in the hydrozoan *Hydractinia* and in *B. schlosseri* (reviewed in Rosengarten and Nicotra 2011). Likewise, there is theoretical evidence for the effectiveness of within-organism selection between cell lineages and forms of somatic growth in decreasing within-colony genetic differentiation in the face of somatic mutations (Otto and Orive 1995; Otto and Hastings 1998; Orive 2001). Recent empirical work in both plants and animals has demonstrated mechanisms for both an increased opportunity for within-organism selection (Burian et al. 2016) and a reduction in the opportunity for between-genotype competition for reproduction (Barfield et al. 2016). In coalescing red algae, which exhibit high levels of genetic chimerism from fusion (González and Santelices 2017), differences between growth rates between lineages of cells may serve to segregate the lineages into different axes, thus reversing

chimerism in upright axes or branches after fusion as one cell lineage outcompetes the others (Santelices et al. 2016).

In *E. larynx*, whose ancestors were solitary and lost the ability to form extensive colonies through asexual budding (Cartwright and Nawrocki 2010; Nawrocki and Cartwright 2012), the fierce competition for space on marine surfaces, and other size-related mortality factors (Jackson 1977; Coates 1985), may have driven the re-evolution of colonies that are a compromise as far as genetic homogeneity. Colonies of *E. larynx* may represent an evolutionary “kluge” of sorts where the pressure to form large colonies takes precedence over preventing germline-soma or germline-germline competition.

Ectopleura larynx belongs to the hydrozoan clade Aplanulata (Collins et al. 2005; Nawrocki et al. 2013). Most members of Aplanulata lack a free-living planula larvae stage and instead brood offspring inside the gonophores of the mother until the juvenile polyp (actinulae) stage. Within Aplanulata, coloniality has been lost and most members of this clade, including the model organism *Hydra*, are solitary. In the lineage leading to *Ectopleura*, the ability to continually bud asexually in an adult colony, leading to indeterminate growth, was also lost, although they display a remarkable ability to regenerate tissues when injured (Nawrocki and Cartwright 2012; Nawrocki et al. 2013).

It is within this evolutionary context that *Ectopleura* re-evolved coloniality via polyp fusion. This is the only known instance of re-evolution of a fully-integrated colonial phenotype in the Hydrozoa (Cartwright and Nawrocki 2010), with different species within *Ectopleura* displaying varying levels of coloniality. In *Ectopleura larynx*, male colonies release sperm and fertilize local female colonies which then brood their offspring. If a new polyp lands on available substratum, as opposed to an established colony, it will branch 4-6 polyps off its apical end to

establish the new colony in an initial phase of determinate growth, but will not proceed beyond this size using the asexual budding process (Petersen 1990; Schuchert 2006; Nawrocki and Cartwright 2012). New polyps can also form as the result of an apparent tissue-damage response when the colony is preyed upon by nudibranchs (P. Cartwright, pers. comm.). Apart from during early development and in the regenerative response to tissue damage, *E. larynx* colonies have never been observed to spontaneously asexually bud polyps (Nawrocki and Cartwright 2012). Instead it appears that *E. larynx* colonies achieve their large size (dozens to hundreds of polyps per colony) through the aforementioned fusion of sexually-produced juvenile polyps from the local population of colonies, ultimately producing colonies which may contain multiple genetic lineages.

Here we characterize the extent to which these different processes (initial budding, regeneration, fusion, and any potential genetic homogenization mechanism) contribute to the formation of *E. larynx* colonies. Using a RAD-seq approach, we determined the genetic diversity within and between colonies and characterized the types of genetic relationships (i.e. familial, clonal) present within colonies of *E. larynx*. Further, to estimate the potential for genetic conflict, we made genome-scale estimates of relatedness to serve as a proxy for the probability of matching or mismatching at a genetic conflict locus. This represents one of the first genome-scale studies of diversity within single colonies of an animal. This unique system, which potentially decouples the effects of coloniality and strict asexual reproduction, will allow us to investigate the effects of coloniality and clonality on the evolution and genetics of a species, as well as the interplay between selection for spatial competition and selection for genetically homogenous colonies.

MATERIALS AND METHODS

Sampling of *Ectopleura larynx*

Specimens of *Ectopleura larynx* were collected from eight locations along the coasts of Maine and Northern Ireland. Colonies were removed from the edges of docks or from submerged rocks in shaded subtidal areas. Sex was determined morphologically for each colony and 5-15 polyps per colony were chosen at random from colonies composed of approximately 25-100 polyps (Figure 3.1C). The polyps (Figure 3.1D) were stored individually in ethanol for DNA isolation.

DNA extraction/Library construction/Sequencing

DNA was extracted from each polyp using the QIAGEN DNeasy Blood and Tissue kit. To obtain genome-wide sequencing representing loci across the entire genome of *E. larynx*, we took a restriction site-associated-digest (RAD-Seq) approach (Miller et al. 2007; Baird et al. 2008) using a modified multiplex-shotgun-genotyping method (Andolfatto et al. 2011) as implemented in Monnahan et al. (2015), with the following modifications and barcoding of samples as outlined: Use of the NdeI restriction enzyme (NEB Biolabs, Ipswich, MA, USA) for DNA digestion, a total of 14 rounds of PCR on the pooled products, and the addition of a 10% PhiX spike-in to increase library complexity for sequencing. To allow for demultiplexing of sequences from different samples, each sample was ligated to one of 48 unique barcode adaptors, and each such set of 48 samples was pooled independently. Each of these sublibraries was then combined with different Illumina indices during the PCR step of the library preparation (Monnahan et al. 2015). Ultimately, one hundred ninety-two samples were prepared in two 96-sample batches for sequencing, with assistance from the University of Kansas Genome

Sequencing Core (GSC; Lawrence, KS). Illumina sequencing of the two libraries was performed by the GSC in one lane each of high-output paired-end 100bp reads on an Illumina HiSeq 2500 System, although only forward reads were used in further analyses.

Read quality filtering/reconstruction of polymorphic loci using Stacks

After sequencing, the raw Illumina sequence data was quality-filtered and demultiplexed into sample-specific FASTQ files using the `process_radtags` program of Stacks v.1.44 (Catchen et al. 2013) on default settings. A series of consistently low-quality bases had to be omitted from the middle of reads from one of the two libraries, so the same 14bp were trimmed from the second library as well, leaving 86bp reads for further analysis. At this stage, sequences for several polyps that had fewer than <10,000 retained reads after the `process_radtags` step were removed from further analysis.

We utilized the Stacks pipeline (Catchen et al. 2013) for *de novo* assembly of restriction-site-associated loci and identification of SNPs for further analysis. The pipeline was run twice for the purposes of comparison of major results: Once with entirely default parameters for each program except those enabling parallel execution, and once with additional adjustments made in the UStacks module to increase the minimum stack depth required to build a locus (-m 6) and to enable the resolution of overmerged tags (-d). These adjustments were chosen to prevent both loss of alleles due to insufficient read depth, and false merging of multiple loci into a single stack.

Colony-level and population-level filtering of individuals and polymorphic loci

For investigation of within-colony relationships without any effects of missing data, we retained for each colony through the Stacks “populations” module only those polymorphic sites present in all polyps in a colony (-r 1.0) and with at least eight reads per site/per individual (-m 8). For within-colony investigations of relationships and genetic diversity, only colonies that contained more than four polyps which successfully passed previous filtering steps and had at least 100 associated SNPs were retained, resulting in a data set containing 19 colonies from two populations in Maine and two in Northern Ireland (Appendices 7 and 8). Details about the data sets used for population-level calculations are available in Appendix 9.

Determination of within-colony relationships using genetic distances

A major goal of this study was to determine whether polyps within a colony are part of the same genetic lineage (i.e. matching multi-locus lineages with differences likely only due to sequencing error or somatic mutation) or if there were more distant types of relationships present within the colony resulting from fusion, which, to our knowledge, has never been assessed at a genome-scale for a hydrozoan colony of any species. To do so, we compared genetic distances to those expected under sexual reproduction within a given colony of *E. larynx* both with and without selfing using the R package RClone (Bailleul et al. 2016). In our case, because colonies are a single sex, and therefore within-colony reproduction is not actually possible, the results of these simulations represent a hypothetical distribution of genetic distances to which we can compare the actual observed within-colony distances, allowing for identification of putative clones. Although the selfing simulations do not represent a biologically feasible mode of reproduction given single-sex colonies, they served as an important visual contrast for the lowest

limits of possible non-clonal diversity. We categorized the relationships between polyps in a given colony into three overall categories based on the relation of their genetic distances to those simulated as products of reproduction within a colony: (Type I) comparisons that produced distances less than expected due to sexual reproduction; (Type II) comparisons with distances within the range expected due to sexual reproduction; and (Type III) comparisons corresponding to genetic distances greater than expected due to theoretical within-colony sexual reproduction. Per-colony results of these simulations are presented in Figure 3.2 and Appendix 10.

To visualize these within-colony relationships, Structure-formatted output files from Stacks for each colony were processed into `genind` files using `Adegenet` R package v2.0.1 (Jombart 2008; Jombart and Ahmed 2011). Using the `genet_dist()` function of `Rclone` all possible pairwise allelic distances between polyps in a colony were calculated, and then, using the actual within-colony data, the distribution of those distances with and without selfing (occurring at a rate determined by the number of clonal replicates in the sample) were simulated using the `genet_dist_sim()` function with 1000 simulations each for each colony separately. Density plots of the actual data and simulated distributions were made using the base R graphics library. Neighbor-joining trees using the distance matrices created by `RClone` were created for each colony using the `nj()` function from the `ape` R package v4.1 (Paradis et al. 2004).

Evaluating patterns of intracolony allelic segregation

To further understand the types of relationships present within a colony, we examined patterns of allelic segregation between polyps at all SNPs in the given colony's data set using custom R scripts. The three genetic relationship categories we discriminated between were: (A) genotypes whose only genetic differences were ones that could be explained solely by

sequencing error or somatic mutation within a clonal genotype; (B) genotypes that can be explained by Mendelian segregation of alleles found within the main clonal lineage; and (C) genotypes containing alleles not found in other polyps in the colony, giving possible evidence of gene flow from outside the main clonal lineage. We further divided this last category into whether or not they differed by one or both alleles from the genotypes of other polyps in the colony. Mismatching at both alleles as compared to other polyps was taken as evidence of fusion from unrelated polyps.

Because our data (proportion of sites which differ in a given comparison) may violate assumptions of an ANOVA (namely those of normality and independence), we conducted k-means clustering to test for the presence of distinct clusters that matched with the three relationship types we described previously. Assessment of optimal number of clusters was carried out in `fviz_nclust()` function from R package `factoextra` v.1.05 (Kassambara and Mundt 2017) and the `NbClust()` function from the `NbClust` package (Charrad et al. 2014) (Appendices 11 and 12). Actual k-means clustering of relationships by genetic distance was carried out using the `kmeans()` function from base R. Density plots to visualize these distances were created using the `ggplot2` R package (Wickham 2009), using the same data set as for the k-means clustering process (Figure 3.3).

Assessing the effects of read-depth and allelic dropout on our results

One potential issue with RAD-sequencing based techniques, as opposed to Sanger-sequenced genetic markers such as microsatellites, is loss of one allele in a true heterozygote, creating false homozygotes in the resultant data set. This “allelic dropout” can either be due to low sequencing depth of one of the two alleles, meaning that during the SNP-calling phase these

sites are called as homozygotes, or due to a genuine mutation in one of the two alleles such that it no longer has a restriction-enzyme cut site (Gautier et al. 2013, Davey et al. 2014). It might be expected that the former issue will be most prevalent when sequencing depth is low, while the latter can be diagnosed by the presence of low read depth due to the loss of the RAD-sequences associated with the null allele (Gautier et al. 2013). Given this, it was important to assess whether or not read-depth was roughly the same for all loci, particularly for sites which had the alternate homozygote in comparison to the other polyps in the colony, by comparing distributions of read depth at those sites both graphically and statistically using functions from the `vcfR` package v1.4 (Knaus and Grünwald 2017) and the `t.test()` and `Wilcox.test()` functions from Base R v.3.3.3 (Team 2017).

Calculating population-wide estimates of diversity

For the data sets used to calculate collecting-location-level genetic diversity statistics, every individual available at a given site was incorporated into the analyses, even if that polyp's colony was not included in colony-level analyses. Minimum read depth was decreased to six (-m 6) and coverage across all polyps in the whole location (-r 1.0) was required for a SNP to be retained. Genetic diversity statistics reported in Table 3.1 were taken directly from the output of the Stacks "populations" module for each relevant collecting location. Additionally, these location-level data sets were used to calculate pairwise between-colony F_{ST} values also using the Stacks "populations" module. For a brief summary of data sets and filtering parameters for each analysis conducted in this publication, see Appendix 9.

Genomic estimates of relatedness between colony-mates

To further understand the potential for either cooperation or conflict between polyps in a given colony of *E. larynx*, we determined whether polyps within a colony are more closely related to one another than polyps from different colonies. Such an analysis will determine whether different alleles at loci involved in conflict would likely interact with one another in a colony, even given the low amount of background genetic diversity observed in these samples. One traditional measure of the potential for such conflict is Hamilton's r (relatedness), defined as the probability that a recipient individual carries an allele identical by descent with an allele sampled randomly from a donor (Charlesworth and Charlesworth 2010). For a relevant locus, the higher the relatedness between individuals, the less potential for genetic conflict between individuals with respect to this locus.

For large SNP data sets, one proxy for estimating relatedness of two individuals is to estimate the average probability of a match between alleles drawn at random from each individual, under the assumption that sequence identity implies identity by descent (approach reviewed in Speed and Balding 2015). The probability of identity by descent for two alleles sampled from two different individuals (i.e. a kinship coefficient) is directly related to the potential for genetic conflict as it should be half of Hamilton's r in the absence of inbreeding. One algorithm that can calculate this probability for SNP data is KING-robust (Manichaikul et al. 2010). This algorithm has the added benefit of being robust to unknown background population structure, which we have little knowledge of in *E. larynx*. Using the KING-robust algorithm as implemented in the “-relatedness2” option of VCFtools v. 0.1.15 (Danecek et al. 2011), we calculated ϕ (kinship coefficient) for all possible pairs of polyps in a collecting location, both within and between colonies.

To further quantify the probability that individuals in a given colony will mismatch or not at potential allorecognition loci in the absence of knowledge of specific loci, we utilized the “—genome” option in the software package PLINK (Purcell et al. 2007) to calculate probabilities of identity by descent at a random SNP for 0 alleles, 1 allele, and both alleles for every possible pair of two individuals. In order to create a larger data set of independent SNPs for estimating relatedness between colonies, we relaxed some SNP filtering and turned on the —write_single_snp option in the Stacks populations module (see Appendix 9). Although this increased the amount of missing data, it allowed for an increased number of loci to consider genome-scale, between-colony estimates of relatedness. Measures of relatedness could be affected by the presence of unknown inbreeding, but we find that there is limited evidence for inbreeding in this system (see estimates of F_{IS} in Results section).

RESULTS

Sequencing, read filtering and subsequent loci reconstruction resulted in a data set containing nineteen colonies from four of the sampled locations, two in Ireland and two on the coast of Maine, and hundreds of loci per colony that had no missing data and sufficient depth of coverage (Appendix 8).

Determination of relationships via simulations

We generated and visualized simulated distributions of genetic distances (as number of loci which vary in a pairwise comparison) between offspring resulting from hypothetical within-colony sexual reproduction, and then plotted our actual within-colony, between-polyp pairwise genetic distance data on the same graphs (Figure 3.2). Our hypothetical distributions served as reference for the possible amount of genetic variation among offspring that could be produced by

within-colony genetic variation alone. We categorized the relationships between polyps in a given colony into three categories based on the relationship of their genetic distances to those simulated as described in the Methods section.

Some combination of these relationships/distance classes were present in each given colony (Figure 3.2, Appendix 10). Notably, 18 out of 19 colonies examined possessed Type I (clonal) comparisons and 16 contained some combination of Type II and Type III. There were no significant positive or negative associations between the possession of Type II and III relationships (P-value for Fisher's exact test of association between two categorical variables is 0.6649). Initial examination of the presence of these distance classes in the sets of loci resulting from our "default" and "conservative" Stacks loci reconstruction runs yielded identical results, so the "default" data set was not investigated further.

Comparison of genetic distances

Visually, the three types of within-colony relationships described above had distinct, nearly non-overlapping genetic distances (Figure 3.3). A majority of methods to select an optimal number of clusters (K) selected only two clusters as the minimal K to explain most variation in the data (Appendices 11 and 12), although some measures did select values of K greater than two. K-means clustering confirmed that grouping the comparison data into only two groups explained >88% of the variation in this data set, although expanding the analysis to three clusters does explain more of the data (93.2% total). Type III relationships (those that are larger than expected from within-colony sexual reproduction), fall in a distinct, non-overlapping cluster from Types I and II (Figure 3.4B).

The average Phred score for our data set after quality-filtering of reads was 37.79, corresponding to a Q-score, or the probability that a particular base is incorrectly called, of just .00017. The Type I and some Type II genetic distances are greater than our Q-score error estimates and most are less than published empirical estimates (Minoche et al. 2011; Wall et al. 2014; Schirmer et al. 2016) of Illumina HiSeq sequencing error rates from studies using similar sequencing approaches (HiSeq 2000/2500, looking at error rates from R1 (forward) reads, including pre-analysis quality filtering). All Type III relationships are greater than both of these estimates (Figure 3.3). Although all of Type I and most of the Type II relationships fall below published estimates of sequencing error, the existence of two distinct peaks for Type I and Type II relationships and our further examination of individual SNPs (below) indicate that Type II comparisons are distinct from the between-clone comparisons with sequencing error that comprise the Type I comparisons. Taken together, the visual and statistical evidence suggests the presence of at least two, and likely three, distinct classes of relationships amongst our within-colony genetic distance data. Type III relationships particularly (those greater than expected according to a simulated model of sexual reproduction), represent a unique class of within-colony genetic relationships.

Examination of individual loci

For each site that was polymorphic amongst Type I relationships in a colony, we examined whether it appeared to be evidence of single, unique sequencing errors, somatic mutation or other forms of actual divergence between putative clones. We found that 79.3% of all SNPs with the Type I clonal category are found in just a single within-colony comparison, suggesting that they are random sequencing error. Of sites polymorphic between clones, 19.3%

of them contained two repeated allelic configurations that are one mutational/error step apart, suggesting the possibility of colony-specific somatic mutation. Finally, just 1.3% of sites polymorphic between clones are not fully explicable by either of the processes above (more than two alleles per site or one polyp being separated by more than one mutational/error step from the others).

Next, we examined the loci polymorphic in the other two types of relationships to determine if fusion of polyps from other colonies or from polyps not sampled in this study was responsible for some of the genetic divergence between polyps in these comparisons. In particular, we identified a class of loci where polyps had a different homozygous genotype than the other polyps in the same colony. That is, the divergence in these sites could not be explained as solely as products of Mendelian segregation of the clonal alleles, a single step of somatic mutation, or sequencing error from other genotypes in the colony. Given the extremely low probability of somatic mutation occurring twice at a given site (Orive 2001), the existence of loci where polyps had a different homozygous genotype can be taken as evidence of the presence of multiple distinct genotypes in a given colony that are not consistent with mosaicism through random somatic mutations, and so are likely the products of chimeric fusion (Schweinsberg et al. 2015; Schweinsberg et al. 2016; Schweinsberg et al. 2017). All Type III comparisons and some Type II comparisons possessed sites which differed in this manner.

To rule out the possibility that this pattern of divergence was generated by dropout of one allele in either genotype in a comparison (i.e. either polyp could be a false homozygote) due to low sequencing depth, we compared the distributions of per-colony read depth between sites at which this pattern occurred vs. all other sites. We found that the shape and location of the distributions of read depths were nearly identical (Appendix 13) and did not have a significantly

different mean ($t = 1.6338$, $df = 4748.8$, $p\text{-value} = 0.1024$) or overall location/shape (Mann-Whitney U Test, $U = 0$, $W = 56071000$, $p\text{-value} = 0.6064$).

Population-level genetic diversity

At each collecting location, the distributions of between-colony and within-colony genetic distances are largely overlapping, indicating that there are some within-colony comparisons that are just as large as between-colony comparisons, and that there are some extremely similar genotypes present in multiple distinct colonies (Appendix 14). The presence of these nearly identical genotypes in different colonies may represent fusion of polyps produced through matings between close relatives or may also simply be an artifact of having few polymorphic loci with which to distinguish individuals.

Calculations of relevant diversity parameters considering all sites, both variant and invariant, are presented in Table 3.1. Overall, populations of *E. larynx* display low levels of polymorphism and allelic diversity (π) when compared to other location-level, RADseq-based estimates of genomic diversity of marine invertebrates, even considering other clonal and/or colonial cnidarians (Bellis et al. 2016; Drury et al. 2016; Drury et al. 2017; Gleason and Burton 2017; Xu et al. 2017). The F_{IS} values, which are effectively zero for each collecting location (Table 1), indicate that inbreeding is limited within *E. larynx*, at least at the scale of a whole locale. Despite this low diversity, our calculated between-colony pairwise F_{ST} values within each locale (Appendix 15) appear elevated and are higher or comparable to population-level comparisons in the above studies on other marine invertebrates, and comparable to microsatellite-based F_{ST} values for distinct populations of the hydrozoan *Macrorhynchia*

phoenicea (Postaire et al. 2017). This suggests that different colonies of *E. larynx* have distinct allele frequencies for the few sites that are polymorphic in a given collecting location.

Genomic measures of allele-sharing within and between colonies

For both within- and between-colony comparisons between polyps, SNP-based measures of relatedness, as estimated through the probability of allele-sharing between individuals at many independent sites across their genome, are summarized in Figure 3.4, and full results for each pairwise comparison for each analysis are available in Supplemental Table 7 (available as supplemental material: <https://doi.org/10.1002/evl3.68>). Considering relatedness (ϕ) as calculated using the KING-robust algorithm (Fig. 3.4A), the degree to which polyps share alleles is significantly higher in within-colony comparisons (mean within = 0.264, mean between = 0.067, $t = 33.174$, $df = 569.38$, $p\text{-value} < 2.2e-16$). For comparison, for a parent-offspring or full-sib relationship, ϕ is expected to be 0.25, and ranges from 0 (negative values are treated as 0) for unrelated individuals and 0.5 for monozygotic twins or, in our case, identical clones (Manichaikul et al. 2010). This comparison suggests that, on average, polyps in a given colony are approximately as related as parent-offspring pairs or siblings. However, within most colonies there are also pairwise comparisons between polyps that have ϕ of roughly 0, indicating that there are in fact fused polyps in these colonies that have no familial relationship whatsoever to the other polyps in the colony.

Calculations of identity-by-descent (IBD) probabilities using PLINK (Purcell et al. 2007) confirms these findings (Figure 3.4B-D). Notably, the distributions of these measures for between- and within-colony comparisons between polyps had very different shapes but overlapped in their ranges. For example, for between-colony comparisons, the probability that a

random locus is not IBD at either allele for a pair of polyps (i.e. $\Pr(Z = 0)$) is very close to 1.0, with a mean $\Pr(Z = 0)$ of 0.945. For within-colony comparisons, however, the distribution for IBD probabilities are much more uniform and have slight increases in density close to 0 and 1.0 for both $\Pr(Z = 0)$ and $\Pr(Z = 2)$ (the probabilities that the number of alleles that are IBD at a random locus for a pair of polyps is 0 or 2, respectively; Fig. 3.4B and 3.4D), further suggesting the presence of at least two classes of relatedness within colonies.

Given these measures of relatedness in a given colony, it appears that on average, polyps within the same colony are more related than those from different colonies and may therefore encounter less opportunity for genetic conflict due to allelic mismatches at loci associated with genetic conflict. However, another distinct class of less-closely related polyps also exist within a given a colony, with similar estimates of relatedness as compared to between-colony comparisons, and which tend to share no alleles at polymorphic sites. This indicates that there is still potential for genetic conflict between some polyps in a given colony.

DISCUSSION

Genetic composition and levels of relatedness within *Ectopleura larynx* colonies

Our results demonstrate that colonies of *E. larynx* are genetically chimeric, containing multiple distinct genotypes that fall into potentially three distinct classes of genetic relationships, resulting in groups of polyps that likely represent clone-mates, offspring/siblings, as well as polyps with non-familial relationships. The different genetic relationships within a colony can be explained by the life history of *E. larynx*. Clones can arise early in development when a new polyp will undergo an initial round of determinate budding, resulting in four to six polyps (Pyefinch and Downing 1949; Petersen 1990). In addition, although the adult colony does not

bud asexually, *E. larynx* polyps have remarkable regenerative capabilities (Tardent 1963) and new polyps can form via regeneration in response to injury (P. Cartwright, pers. obs.). This initial asexual growth and regeneration response likely explains the persistence of clonal genotypes in a given colony, but it appears that *E. larynx* colonies achieve increased size from polyp fusion, including fusion of polyps not closely related to the main colony genotype. Colonies with genetically distinct polyps are thus produced through the fusion of juvenile polyps either brooded from the mother or from, unexpectedly, unrelated neighboring colonies.

We also determined whether or not polyps within colonies have a greater probability of identity by descent (IBD) than polyps in different colonies. We found that levels of relatedness were on average higher within colonies than between, but that there were many examples of polyps present in the same colonies that did not share either allele at the few polymorphic sites recovered and therefore may differ at relevant “conflict” loci.

The finding that polyps are generally more related within a colony than between colonies, but that there are some less-related polyps in a colony, suggests that *E. larynx* may possess a mechanism of conflict mediation that homogenizes chimeric polyps, such as somatic-cell takeover (Buss 1982, Michod 1982). However, the life history of *E. larynx*, particularly the budding and wound repair, can also explain this level of genetic relatedness as they ensure that a subset of polyps in a colony will be clonal. Additionally, juvenile polyps are limited in dispersal and thus frequently, but not always, settle on parental colonies, resulting in a colony consisting of polyps that are closely related. This is similar to work on the seaweed *Chondrus crispus* where limited dispersal of gametes results in levels of relatedness between males siring offspring with the same female higher than background relatedness (Krueger-Hadfield et al. 2015).

Our result that the presence of multiple genotypes are commonly found in colonies of *E. larynx* stands in contrast with previously reported examples of genetic heterogeneity amongst colonial organisms. Reported genetic mosaicism in some anthozoan and hydrozoan corals are represented by one or a few genetic changes likely generated by somatic mutations (Puill-Stephan et al. 2009; Schweinsberg et al. 2015; Schweinsberg et al. 2016; Schweinsberg et al. 2017), and not multiple distinct genotypes as reported here (Appendix 4). The level of chimerism discovered in *E. larynx* colonies approaches that of some red seaweeds, that are known to derive ecological benefits from chimerism and also appear able to limit the level of chimerism in certain tissues, perhaps reducing the burden of genetic conflict (González and Santelices 2017, Santelices et al. 2016).

Low genetic diversity and self/non-self recognition

Genetic chimerism is predicted to be much less common than mosaicism (intraorganismal diversity due to somatic mutation) due to the wider genetic distances involved and the potential involvement of the immune system in preventing wholesale fusion of organisms (Santelices 2004). The prevalence of genetic chimerism amongst *E. larynx* colonies, however, raises the question of why self/non-self recognition mechanisms appear not to be operating. One notable result of our work is the discovery that *E. larynx* has low population-level diversity (Table 3.1), even when compared to other cnidarians and invertebrates sampled at similar spatial scales (see Results).

Given this finding, it is possible that polyps of *E. larynx* in a local mating population are genetically similar enough that potential germline conflict is mitigated. Many experimental fusion studies of colonial animals such as the tunicate *B. schlosseri* and the hydrozoan

Hydractinia symbiolongicarpus (Cadavid et al. 2004; Lakkis et al. 2008; Rosengarten and Nicotra 2011; Taketa and De Tomaso 2015) and of certain reef-building corals (Puill-Stephan et al. 2009) show that level of relatedness is directly correlated with the capacity with which two genotypes will fuse with one another to form a chimera. In the hydrozoan *H. symbiolongicarpus*, shared allorecognition alleles largely explain whether colonies are able to form persistent chimeras (Cadavid et al. 2004). If all polyps of *E. larynx* in an area are genetically similar to one another, selection for mechanisms to keep colonies homogenous may be greatly reduced compared to selection for large colony size. Thus it is possible that *E. larynx* colonies cannot differentiate between self and non-self due to low genetic diversity.

Our measures of genome-wide relatedness (see “Genetic Composition” section above), can serve as a proxy for the potential for genetic conflict at such loci, in the absence of knowledge in *E. larynx* about specific “conflict” loci such as those involved in allorecognition. This result suggests the presence of a subset of polyps which in theory could differ at “conflict” loci within the same colony. Further work to identify and characterize actual allorecognition genes in the genome of *E. larynx* will shed light on whether or not more distantly related polyps in a given colony actually vary at sites important for self/non-self recognition.

One possible explanation for the low genetic diversity in populations of *E. larynx* is a climatically-mediated genetic bottleneck, potentially at either at geological time scales due to glaciation (Maggs et al. 2008) and seasonal scales, due to harsh winter conditions causing mortality (Drolet et al. 2013). *E. larynx* is most abundant in late summer (Guenther et al. 2009) although it is unclear if this is entirely due to new colonies or recovery from a winter dormancy (Calder 1990).

The potential for genetic conflict within colonies of *Ectopleura larynx*

Given that *E. larynx* colonies appear to be made up of a set of relationships including the fusion of distinct genotypes, there is the potential for germline conflict. Within a colony, the multiple distinct genetic lineages may all be competing for the opportunity to be represented in the gametes. Given that hydrozoans continually produce new germline cells from a population of multi/totipotent stem cells (Müller et al. 2004), it is possible that any of the fused polyps may have access to gamete production. Barfield et al. (2016) found that colony-specific somatic mutations in the coral *Orbicella faveolata* were not transferable to gametes, whereas Schweinsberg et al. (2014) demonstrated that more than one genotype from a colony of the coral *Acropora hyacinthus* was able to reproduce. Polyps in an *E. larynx* colony are almost always of the same sex, thus suggesting that a single germline is functioning. However, the relative contributions of environmental and genetic factors driving sex determination in hydrozoans is unclear (reviewed in Siebert and Juliano 2017). Future studies characterizing parental and offspring genotypes in *E. larynx* colonies are needed to definitively determine if germline chimerism or germline parasitism exists within an *E. larynx* colony.

Conclusions

Past studies of the presence of polymorphism in cnidarian colonies have largely used a selection of mitochondrial and microsatellite markers, making this study among the first to capture genome-scale information about intra-colonial divergence and diversity. Our work reveals that colonies of *E. larynx* are genetically chimeric and contain multiple types of within-colony genetic relationships, namely clones and familial relationships and, surprisingly, fusion of unrelated polyps from the local population with a low degree of allele-sharing with the rest of the

colony. This is consistent with a colony formation mechanism that relies on fusion of sexually-produced offspring from the surrounding population, and not just the fusion of recently released brooded juvenile polyps from the parent colony. Due to the observation of frequent within-colony polymorphism and chimerism in *E. larynx*, and the observation that there are multiple levels of genetic relatedness in a colony, it appears that there is potential for germline-soma conflict, but that this might be mitigated by low genetic diversity in populations of *E. larynx* and by processes that cause polyps in colonies to be more closely related on average than those from different colonies. Taken together, all of these results are consistent with the interpretation that there are multiple biological processes, including initial asexual budding, regeneration, fusion, and possibly a potential genetic conflict resolution mechanism, all contributing to the re-evolution of *E. larynx*'s large colony size from a solitary ancestor.

TABLES

Table 3.1 Per-collecting-location genetic diversity statistics for *E. larynx*

<i>Location</i>	<i># Polyps</i>	<i>Colonies</i>	<i># SNPs</i>	<i>% Polymorphic sites</i>	<i>Π</i>	<i>F_{IS}</i>
<i>ME.2</i>	39	7	164	0.7373	0.002	-0.0005
<i>IRE.2</i>	14	4	32	0.2794	0.0008	0
<i>IRE.1</i>	14	3	100	0.3050	0.001	0.0001
<i>ME.1</i>	55	12	58	0.7764	0.0012	-0.0001

FIGURES

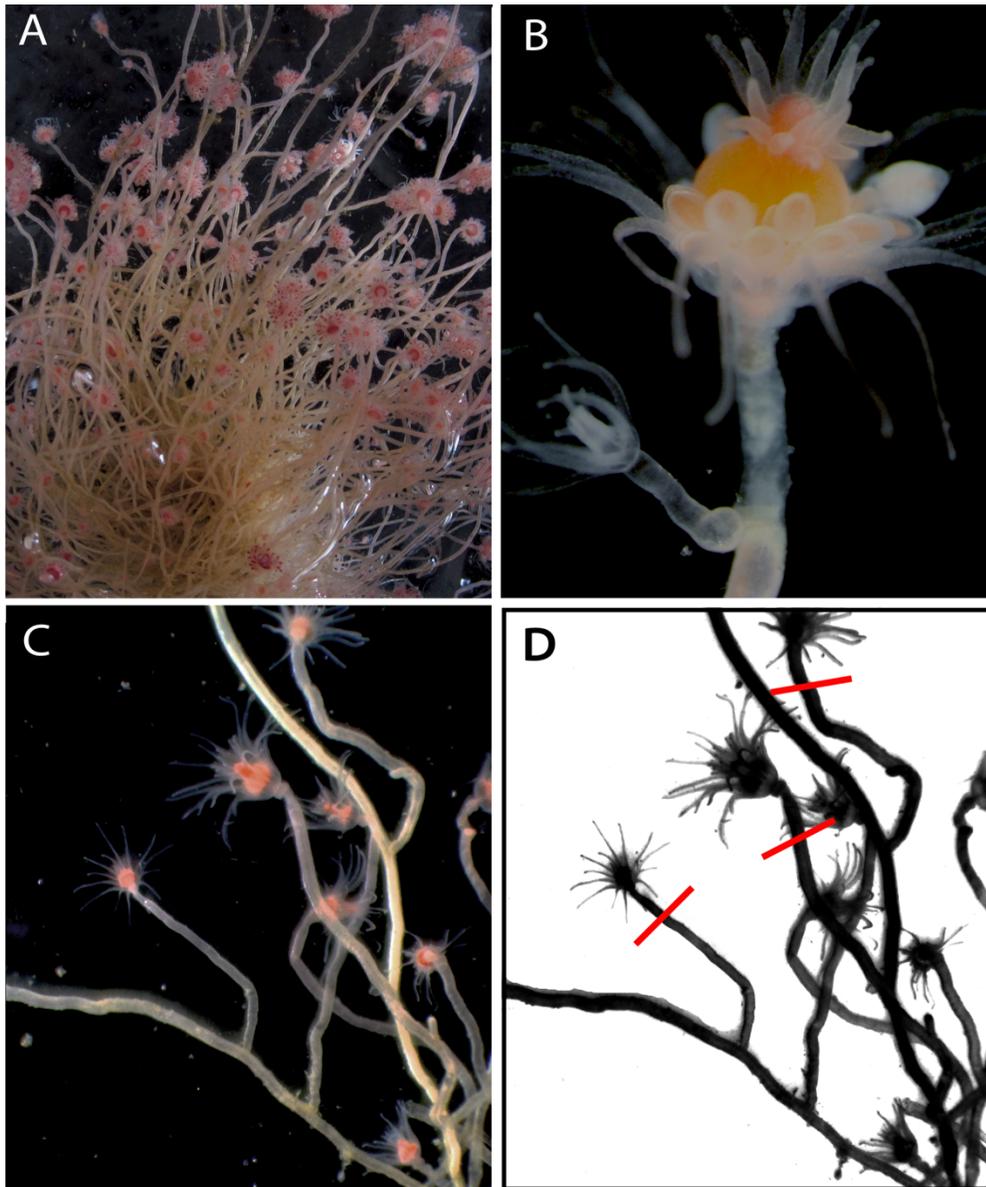


Figure 3.1: *Ectopleura larynx*

A) *E. larynx* colony. B) Juvenile polyp after recent settlement on adult, female colony. C) Juvenile polyps sharing continuous tissue with rest of adult colony after complete fusion. D) Schematic version of Panel C, with red lines indicating approximate cut sites for polyp harvesting.

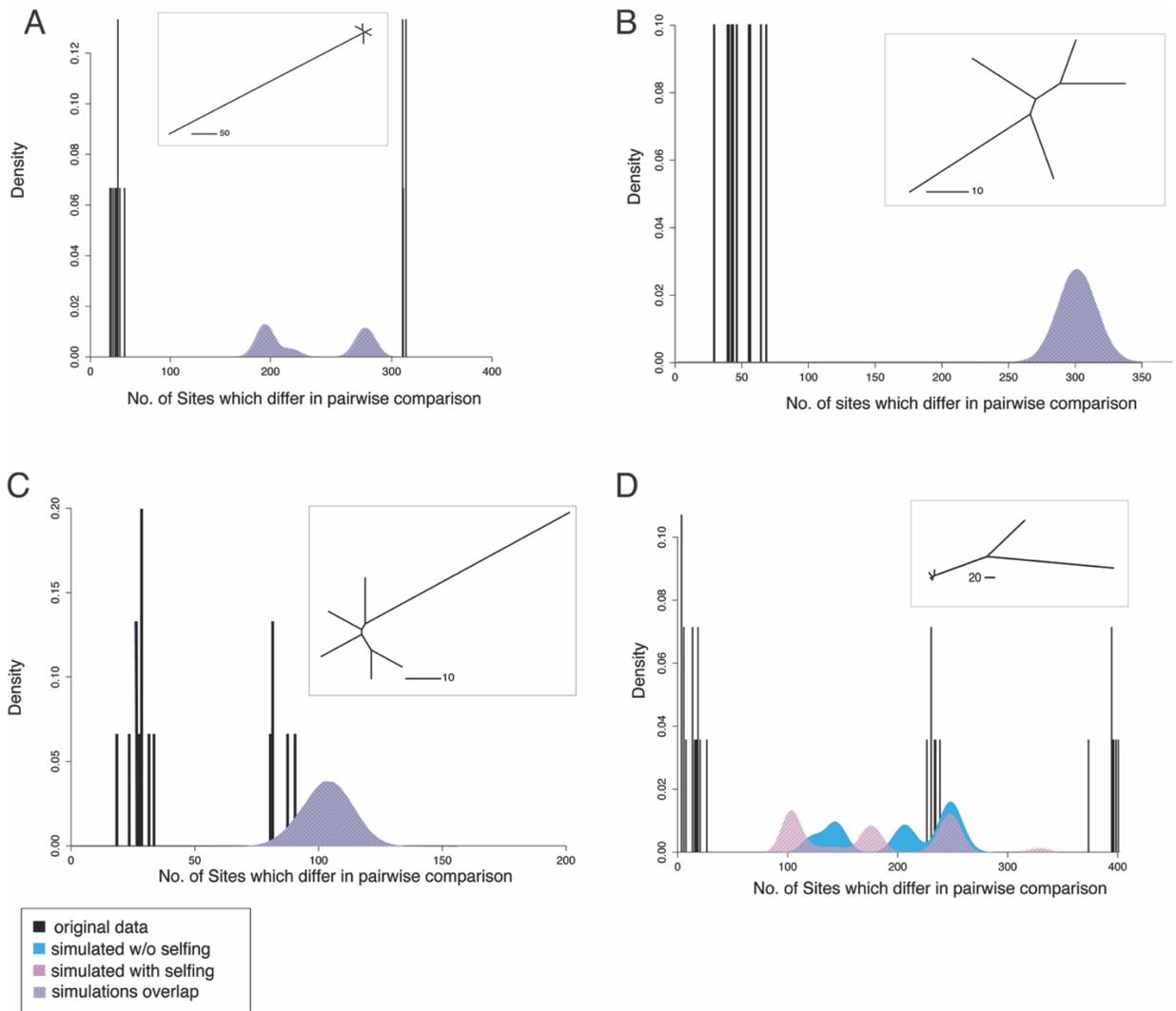


Figure 3.2: Genetic relationships within selected colonies of *E. larynx*.

Actual genetic distances (black bars, calculated as the number of loci which differ in each pairwise comparison) between polyps of *E. larynx* in every possible pairwise comparison in a colony, and those predicted under simulations of sex with selfing (pink), without (blue) and where those simulated distributions overlap (violet), for colony ME2.2 (A), ME1.3 (B), ME1.7 (C) and ME2.3 (D). Insets are neighbor-joining trees with each branch representing a polyp within the colony, branch lengths representing number of differing polymorphic sites.

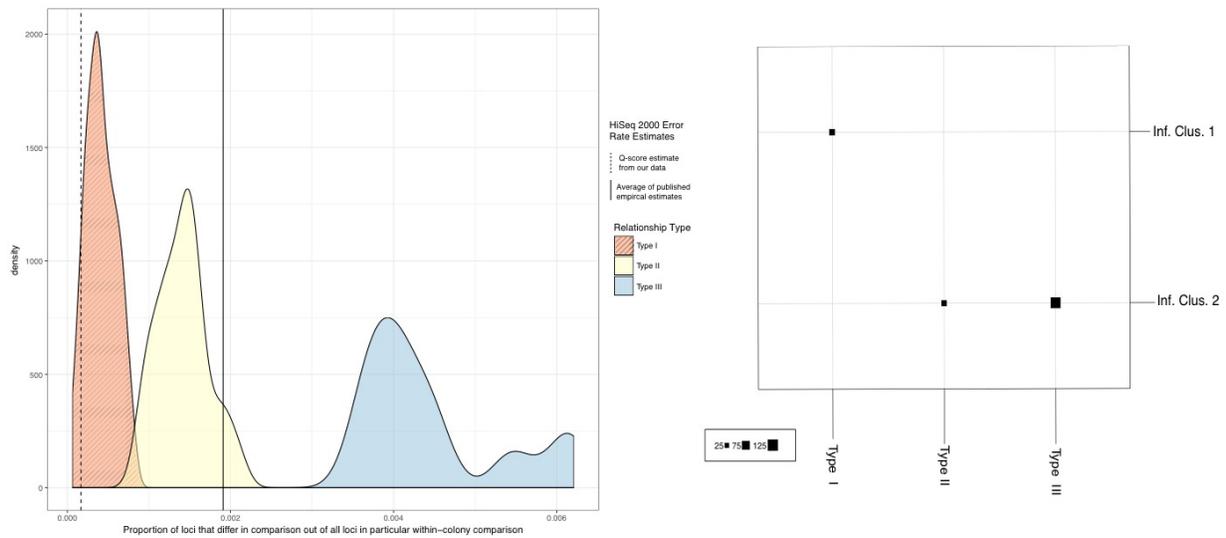


Figure 3.3: Density distributions of the combined pairwise genetic distances and results of k-means clustering for the three major relationship types

Panel A: Q-score estimate is the probability that a given base call is incorrect calculated from the average post-trimming Phred score for entire data set (37.79). The average published estimate is the average (.00191, SD=.00027) of three independent empirical studies quantifying substitution error rates on the HiSeq 2000/2500 line of sequencers (Minoche et al. 2011; Wall et al. 2014; Schirmer et al. 2016). Panel B depicts membership in two inferred clusters (rows), as compared with membership in the three types of relationships inferred via the simulation approach (columns). Type III relationships (those that are larger than expected due to within-colony sexual reproduction), fall in a distinct, non-overlapping cluster from Types I and II.

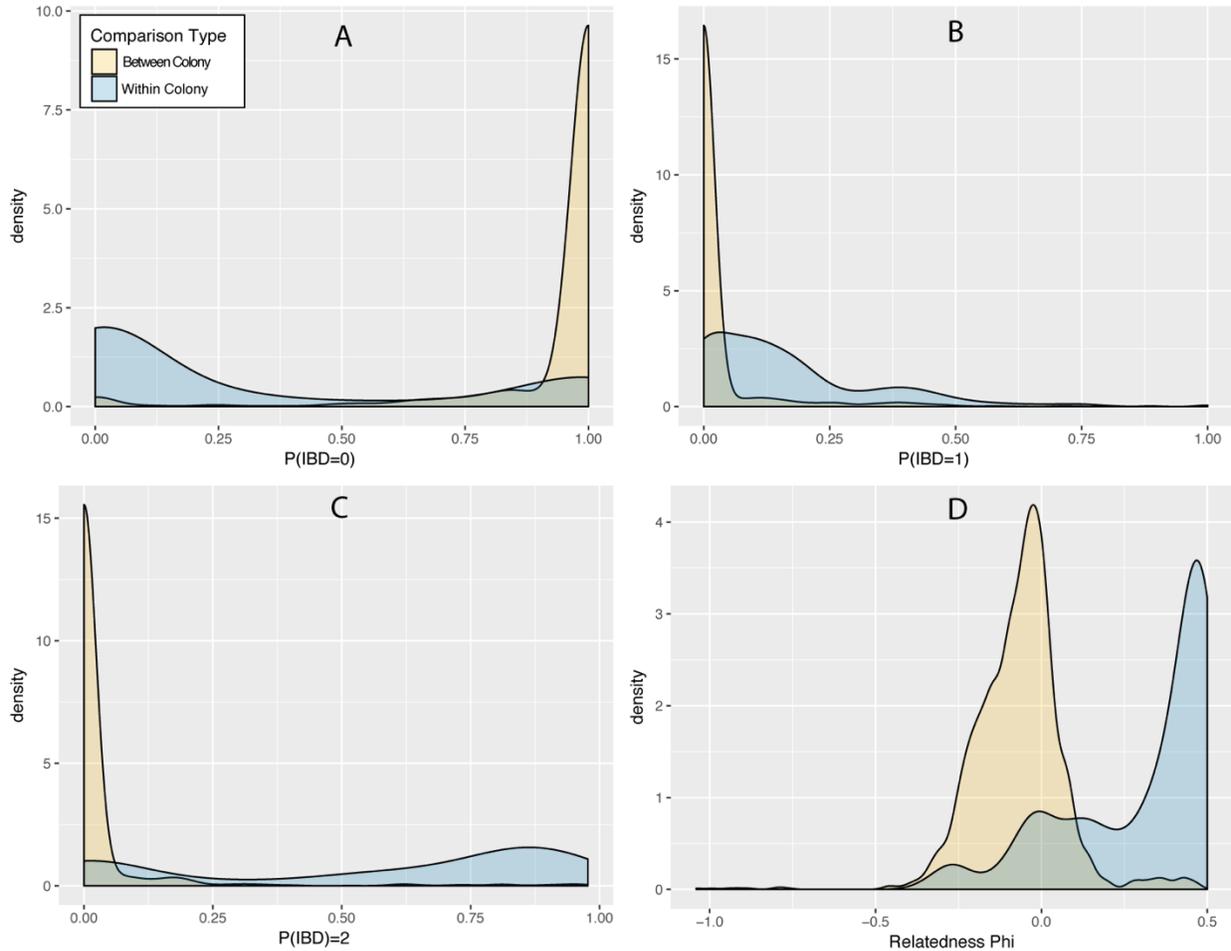


Figure 3.4: Distributions of values for several measures of relatedness for within- and between-colony comparisons between individuals.

Panel A displays values of the relatedness coefficient (equivalent to one half of Hamilton's r) as calculated using the software package KING. Higher values for the coefficient indicate a higher degree of allele sharing between individuals in a comparison, ranging from 0.5 for identical clones and 0 (or negative values treated as 0) for unrelated polyps. Panels B-D depict probabilities of different levels of identity by descent, as calculated using the software package PLINK. $\Pr(Z = 0)$ is the probability that individuals in a given comparison will be identical at a randomly selected SNP at no alleles, $\Pr(Z = 1)$ is the probability that individuals will be identical at one allele, and $\Pr(Z = 2)$ is the probability that a randomly selected SNP will be identical at both alleles.

LITERATURE CITED

- Andolfatto, P., D. Davison, D. Erezyilmaz, T. T. Hu, J. Mast, T. Sunayama-Morita, et al. 2011. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Res.* 21:610-617.
- Bailleul, D., S. Stoeckel, and S. Arnaud-Haond. 2016. RClone: a package to identify MultiLocus Clonal Lineages and handle clonal data sets in R. *Methods Ecol Evol* 7:966-970.
- Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, et al. 2008. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLOS ONE* 3:e3376.
- Barfield, S., G. V. Aglyamova, and M. V. Matz. 2016. Evolutionary origins of germline segregation in Metazoa: evidence for a germ stem cell lineage in the coral *Orbicella faveolata* (Cnidaria, Anthozoa). *Proc Biol Sci* 283:20152128.
- Bellis, E. S., D. K. Howe, and D. R. Denver. 2016. Genome-wide polymorphism and signatures of selection in the symbiotic sea anemone *Aiptasia*. *BMC Genomics* 17:160.
- Burian, A., P. Barbier de Reuille, and C. Kuhlemeier. 2016. Patterns of Stem Cell Divisions Contribute to Plant Longevity. *Current Biology* 26:1385-1394.
- Buss, L. W. 1982. Somatic cell parasitism and the evolution of somatic tissue compatibility. *Proc Natl Acad Sci U S A* 79:5337-5341.
- Buss, L. W. 1987. *The Evolution of Individuality*. Princeton University Press, Princeton, NJ.
- Cadavid, L. F., A. E. Powell, M. L. Nicotra, M. Moreno, and L. W. Buss. 2004. An invertebrate histocompatibility complex. *Genetics* 167:357-365.
- Calder, D. R. 1990. Seasonal cycles of activity and inactivity in some hydroids from Virginia and South Carolina, U.S.A. *Can J Zool* 68:442-450.
- Carpenter, M. A., J. H. Powell, K. J. Ishizuka, K. J. Palmeri, S. Rendulic, and A. W. De Tomaso. 2011. Growth and Long-Term Somatic and Germline Chimerism Following Fusion of Juvenile *Botryllus schlosseri*. *Biol Bull* 220:57-70.
- Cartwright, P. and A. M. Nawrocki. 2010. Character Evolution in Hydrozoa (phylum Cnidaria). *Integr Comp Biol* 50:456-472.
- Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko. 2013. Stacks: an analysis tool set for population genomics. *Mol Ecol* 22:3124-3140.
- Charlesworth, B. and D. Charlesworth. 2010. *Elements of Evolutionary Genetics*. Roberts and Company.
- Charrad, M., Ghazzali, N., Boiteau, V., and A. Niknafs. 2014. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, 61(6):1-36. <http://www.jstatsoft.org/v61/i06/>.
- Coates, A. G. J., J.B.C. 1985. Morphological themes in the evolution of clonal and aclonal marine invertebrates. Pp. 67-106 in L. W. B. R. E. C. J.B.C. Jackson, ed. *Population Biology and Evolution of Clonal Organisms*. Yale University Press, New Haven.
- Collins, A. G., S. Winkelmann, H. Hadrys, and B. Schierwater. 2005. Phylogeny of Capitata and Corynidae (Cnidaria, Hydrozoa) in light of mitochondrial 16S rDNA data. *Zool Scr* 34:91-99.
- Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27:2156-2158.
- Davey, J. W., T. Cezard, P. Fuentes-Utrilla, C. Eland, K. Gharbi, and M. L. Blaxter. 2013. Special features of RAD Sequencing data: implications for genotyping. *Mol Ecol*

- 22:3151-3164.
- Drolet, D., K. Kennedy, and M. A. Barbeau. 2013. Winter population dynamics and survival strategies of the intertidal mudflat amphipod *Corophium volutator* (Pallas). *J Exp Mar Bio Ecol* 441:126-137.
- Drury, C., K. E. Dale, J. M. Panlilio, S. V. Miller, D. Lirman, E. A. Larson, et al. 2016. Genomic variation among populations of threatened coral: *Acropora cervicornis*. *BMC Genomics* 17:286.
- Drury, C., S. Schopmeyer, E. Goergen, E. Bartels, K. Nedimyer, M. Johnson, et al. 2017. Genomic patterns in *Acropora cervicornis* show extensive population structure and variable genetic diversity. *Ecol Evol* 7:6188-6200.
- Ennis, H. L., D. N. Dao, S. U. Pukatzki, and R. H. Kessin. 2000. Dictyostelium amoebae lacking an F-box protein form spores rather than stalk in chimeras with wild type. *Proc Natl Acad Sci U S A* 97:3292-3297.
- Gautier, M., K. Gharbi, T. Cezard, J. Foucaud, C. Kerdelhué, P. Pudlo, et al. 2012. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Mol Ecol* 22:3165-3178.
- Gleason, L. U. and R. S. Burton. 2016. Genomic evidence for ecological divergence against a background of population homogeneity in the marine snail *Chlorostoma funebris*. *Mol Ecol* 25:3557-3573.
- González, A. V. and B. Santelices. 2017. Frequency of chimerism in populations of the kelp *Lessonia spicata* in central Chile. *PLOS ONE* 12:e0169182.
- Guenther, J., C. Carl, and L. M. Sunde. 2009. The effects of colour and copper on the settlement of the hydroid *Ectopleura larynx* on aquaculture nets in Norway. *Aquaculture* 292:252-255.
- Jackson, J. B. C. 1977. Competition on Marine Hard Substrata: The Adaptive Significance of Solitary and Colonial Strategies. *Am Nat* 111:743-767.
- Jombart, T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24:1403-1405.
- Jombart, T. and I. Ahmed. 2011. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 27:3070-3071.
- Kassambara, A. and F. Mundt (2017). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.5. <https://CRAN.R-project.org/package=factoextra>
- Knaus, B. J. and N. J. Grünwald. 2017. vcfr: a package to manipulate and visualize variant call format data in R. *Mol Ecol Res* 17:44-53.
- Krueger-Hadfield, S. A., D. Roze, J. A. Correa, C. Destombe, and M. Valero. 2014. O father where art thou? Paternity analyses in a natural population of the haploid-diploid seaweed *Chondrus crispus*. *Heredity* 114:185.
- Lakkis, F. G., S. L. Dellaporta, and L. W. Buss. 2008. Allorecognition and chimerism in an invertebrate model organism. *Organogenesis* 4:236-240.
- Maggs, C. A., R. Castilho, D. Foltz, C. Henzler, M. T. Jolly, J. Kelly, et al. 2008. Evaluating signatures of glacial refugia for north atlantic benthic marine taxa. *Ecology* 89:S108-S122.
- Manichaikul, A., J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W.-M. Chen. 2010. Robust relationship inference in genome-wide association studies. *Bioinformatics* 26:2867-2873.

- Michod, R. E. 1982. The Theory of Kin Selection. *Annual Review of Ecology and Systematics* 13:23-55.
- Miller, M. R., J. P. Dunham, A. Amores, W. A. Cresko, and E. A. Johnson. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res.* 17:240-248.
- Minoche, A. E., J. C. Dohm, and H. Himmelbauer. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol.* 12:R112.
- Monnahan, P. J., J. Colicchio, and J. K. Kelly. 2015. A genomic selection component analysis characterizes migration-selection balance. *Evolution* 69:1713-1727.
- Müller, W. A., R. Teo, and U. Frank. 2004. Totipotent migratory stem cells in a hydroid. *J Dev Biol* 275:215-224.
- Nawrocki, Annalise M. and P. Cartwright. 2012. A novel mode of colony formation in a hydrozoan through fusion of sexually generated individuals. *Current Biology* 22:825-829.
- Nawrocki, A. M., A. G. Collins, Y. M. Hirano, P. Schuchert, and P. Cartwright. 2013. Phylogenetic placement of Hydra and relationships within Aplanulata (Cnidaria: Hydrozoa). *Mol Phylogenet Evol* 67:60-71.
- Noce, T. and I. Takeuchi. 1985. Prestalk/prespore differentiation tendency of Dictyostelium discoideum cells as detected by a stalk-specific monoclonal antibody. *J Dev Biol* 109:157-164.
- Orive, M. E. 2001. Somatic Mutations in Organisms with Complex Life Histories. *Theor Popul Biol* 59:235-249.
- Otto, S. P. and I. M. Hastings. 1998. Mutation and selection within the individual. *Genetica* 102:507.
- Otto, S. P. and M. E. Orive. 1995. Evolutionary consequences of mutation and selection within an individual. *Genetics* 141:1173-1187.
- Pancer, Z., H. Gershon, and B. Rinkevich. 1995. Coexistence and Possible Parasitism of Somatic and Germ Cell Lines in Chimeras of the Colonial Urochordate Botryllus schlosseri. *Biol Bull* 189:106-112.
- Paradis, E., J. Claude, and K. Strimmer. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20:289-290.
- Petersen, K. W. 1990. Evolution and taxonomy in capitate hydroids and medusae (Cnidaria, Hydrozoa). *Zool J Linn Soc* 100:101-231.
- Postaire, B., P. Gélín, J. H. Bruggemann, M. Pratlong, and H. Magalon. 2017. Population differentiation or species formation across the Indian and the Pacific Oceans? An example from the brooding marine hydrozoan *Macrorhynchia phoenicea*. *Ecol Evol* 7:8170-8186.
- Puill-Stephan, E., B. L. Willis, L. van Herwerden, and M. J. H. van Oppen. 2009. Chimerism in Wild Adult Populations of the Broadcast Spawning Coral *Acropora millepora* on the Great Barrier Reef. *PLOS ONE* 4:e7751.
- Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, Manuel A R. Ferreira, D. Bender, et al. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics* 81:559-575.
- Pyefinch, K. A. and F. S. Downing. 1949. Notes on the general biology of *Tubularia Larynx* Ellis & Solander. *J Mar Biol Assoc U.K.* 28:21-43.

- Reitzel, A. M., S. Herrera, M. J. Layden, M. Q. Martindale, and T. M. Shank. 2013. Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. *Mol Ecol* 22:2953-2970.
- Rosengarten, R. D. and M. L. Nicotra. 2011. Model Systems of Invertebrate Allorecognition. *Current Biology* 21:R82-R92.
- Santelices, B. 2004. Mosaicism and chimerism as components of intraorganismal genetic heterogeneity. *J Evol Biol* 17:1187-1188.
- Santelices, B., V. González Alejandra, J. Beltrán, V. Flores, and C. Amsler. 2016. Coalescing red algae exhibit noninvasive, reversible chimerism. *Journal of Phycology* 53:59-69.
- Schirmer, M., R. D'Amore, U. Z. Ijaz, N. Hall, and C. Quince. 2016. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics* 17:125.
- Schuchert, P. 2006. The European athecate hydroids and their medusae (Hydrozoa, Cnidaria): Capitata part 1. *Rev Suisse Zool* 113:325-410.
- Schweinsberg, M., R. A. González Pech, R. Tollrian, and K. P. Lampert. 2014. Transfer of intracolony genetic variability through gametes in *Acropora hyacinthus* corals. *Coral Reefs* 33:77-87.
- Schweinsberg, M., R. Tollrian, and K. P. Lampert. 2017. Inter- and intra-colony genotypic diversity in hermatypic hydrozoans of the family Milleporidae. *Mar Ecol*, 38: e12388.
- Schweinsberg, M., L. C. Weiss, S. Striewski, R. Tollrian, and K. P. Lampert. 2015. More than one genotype: how common is intracolony genetic variability in scleractinian corals? *Mol Ecol* 24:2673-2685.
- Siebert, S. and C. E. Juliano. 2017. Sex, polyps, and medusae: Determination and maintenance of sex in cnidarians. *Mol Reprod Dev* 84:105-119.
- Speed, D. and D. J. Balding. 2014. Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics* 16:33.
- Stoner, D. S., B. Rinkevich, and I. L. Weissman. 1999. Heritable germ and somatic cell lineage competitions in chimeric colonial protochordates. *Proc Natl Acad Sci U S A* 96:9148-9153.
- Stoner, D. S. and I. L. Weissman. 1996. Somatic and germ cell parasitism in a colonial ascidian: Possible role for a highly polymorphic allorecognition system. *Proc Natl Acad Sci U S A* 93:15254-15259.
- Taketa, D. A. and A. W. De Tomaso. 2015. *Botryllus schlosseri* allorecognition: Tackling the Enigma. *Dev. Comp. Immunol.* 48:254-265.
- Tardent, P. 1963. Regeneration in the Hydrozoa. *Biol. Rev.* 38:293-333.
- Tay, Y. C., M. W. P. Chng, W. W. G. Sew, F. E. Rheindt, K. P. P. Tun, and R. Meier. 2016. Beyond the Coral Triangle: high genetic diversity and near panmixia in Singapore's populations of the broadcast spawning sea star *Protoreaster nodosus*. *R Soc Open Sci* 3:160253.
- Team, R. C. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- Wall, J. D., L. F. Tang, B. Zerbe, M. N. Kvale, P.-Y. Kwok, C. Schaefer, et al. 2014. Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Res.*
- Wickham, H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.
- Xu, T., J. Sun, J. Lv, H. Kayama Watanabe, T. Li, W. Zou, et al. 2017. Genome-wide discovery of single nucleotide polymorphisms (SNPs) and single nucleotide variants (SNVs) in

deep-sea mussels: Potential use in population genomics and cross-species application.
Deep Sea Res Part 2 Top Stud Oceanogr 137:318-326

Chapter 4

Species delimitation and the evolution of freshwater tolerance in the invasive hydrozoan *Cordylophora* using phylogenetic, population genomic and environmental evidence

ABSTRACT

The study of invasive species can provide us with new insights into ecological and evolutionary processes by presenting “natural experiments” as species adapt to new environments and native species respond to the invaders. For cnidarians, the evolutionary transition from a marine to freshwater habitat enables them to expand their potential range. An excellent system for investigating this evolutionary transition is the invasive hydrozoan *Cordylophora caspia* as it inhabits both freshwater and brackish habitats. Previous studies suggest it may be a species complex comprising lineages of exclusively brackish, freshwater and euryhaline colonies. However, much uncertainty still exists regarding taxonomy and species delimitation in the group, although currently all lineages are considered a single species due to lack of distinguishing morphological characters. In this work we provide a more detailed RAD-seq phylogenomic study of the complex in order to examine possible patterns of species divergence and examine the relationship between salinity and population structure across the complex. We include new samples from along estuarine salinity gradients where we may expect the different lineages to come into contact, and use an integrated phylogenetic and population genomic approach to examine species boundaries. We conclude that there are at least two major lineages of *Cordylophora* separated by both salinity regime and geography, and resurrect a previously synonymized species to represent one of the lineages.

INTRODUCTION

Tolerance to a wide array of environmental conditions has been shown to be a key trait for the success of an invasive species (Lenz et al. 2011; Bates et al. 2013) as it allows for organisms to exploit novel ecological niches. A key evolutionary innovation that has been crucial for the colonization of new environments in aquatic organisms is the transition from a marine ancestor to a freshwater-tolerant descendent. This transition is thought to pose many physiological challenges, particularly for marine invertebrates, which were relatively slow to invade freshwater environments after their initial appearance in the fossil record (Miller 2002). Marine invertebrates, which are usually at equal ion concentrations to the surrounding seawater, face the challenge of acquiring ions from a dilute environment against a steep ion gradient once in fresh water (Lee et al. 2011; Lee 2015). The colonization of fresh water is often accompanied by the evolution of complex changes to osmoregulatory and homeostatic systems (Lee and Bell 1999; Tsai and Lin 2007). Given the physiological challenges accompanying the marine to freshwater transition, relatively few major lineages have managed to make this transition (Lee et al. 2016). Despite these challenges, some invertebrates have expanded their habitats into a full range of salinities, such as the zebra mussel *Dreissena polymorpha*, which was originally native to the Ponto-Caspian region but is now invasive in freshwater and brackish habitats throughout the world (Molnar et al. 2008).

The evolution of freshwater tolerance can lead to genetic differentiation and isolation between populations with different salinity tolerances or preferences. Differences in salinity regimes have been shown to be associated with deep divergence between lineages in other aquatic animals such as killifish, cyclopoid copepods, gammarid amphipods and nudibranchs (Fuller et al. 2007; Chen and Hare 2008; Ueda et al. 2011; Korshunova et al. 2018). For example,

Ueda et al. (2011) demonstrated the presence of two distinct ecomorphotypes of the same species of a cyclopoid copepod in saline, lower portions of estuaries and upper, freshwater portions of estuaries along the east coast the United States. In the killifish, Fuller et al. (2007) discovered that populations of killifish with distinct distributions along the salinity gradient of estuaries are in fact genetically isolated to the point of being separate species. These studies point to the evolution of salinity tolerance differences as a driver of diversification and that salinity can serve as a potential barrier between recently diverged lineages.

Cnidarians (jellyfish, sea anemones, corals and relatives), which are typically thought of as a quintessentially marine phylum, have successfully made this transition to fresh water at least four independent times in their evolution (Jankowski et al. 2008; Chang et al. 2015). These cnidarian invasions of fresh water include members of the genus *Hydra*, the freshwater hydromedusae *Craspedecusta*, the parasites of freshwater fishes, *Polypodium* and Myxozoa, and the freshwater colonial hydroids in the genus *Cordylophora*. Freshwater cnidarians are unusual amongst other freshwater-adapted organisms in that they have no organ systems, and instead must adapt to osmotic stress at the tissue and cellular levels (Folino-Rorem and Renken 2018).

Amongst the freshwater cnidarians, only the invasive hydrozoans in the genus *Cordylophora* are able to tolerate a wide range of salinities, from nearly freshwater to almost entirely marine. Species delimitation within the genus *Cordylophora* continues to prove difficult due to the lack of distinguishing morphological synapomorphies (Schuchert 2004) and their appearance in both brackish and freshwater environments (Folino-Rorem et al. 2009).

Additionally, *Cordylophora* is phenotypically plastic, with different hydranth length and tentacle number at different salinities, making these characters unsuitable for species diagnoses and further confounding taxonomic efforts (Folino-Rorem 2000; Schuchert 2004; Folino-Rorem et

al. 2009). Past taxonomic classifications have included up to eight different species in the genus *Cordylophora* (Schuchert 2018). There are now just currently three accepted species of *Cordylophora* due to the synonymization of the other species with *C. caspia* (Schuchert 2018): *C. caspia*, *C. japonica* (Itô 1951), limited to Japan, and *C. solangiae* (Redier 1967), collected in the Fangataufa Atoll in the South Pacific. Notably, there is confusion as to the proper taxonomic status of *C. caspia* and *C. lacustris* which are currently considered synonymous (Cairns et al. 2002), meaning that *C. caspia* now refers to both brackish and freshwater *Cordylophora* (Cohen 1998; Smith 2001). Other authors, however, have suggested that *C. caspia* occurs only in brackish habitats while *C. lacustris* inhabits fresh water (Folino-Rorem 2000; Smith 2001). The most recent work on the physiology of *Cordylophora* (Folino-Rorem and Renken 2018) refers to the genus and species level taxonomy of *Cordylophora* as “tentative” and refers to all members simply by the genus designation *Cordylophora*.

The remarkable tolerance of *Cordylophora* to a wide range of salinities likely contributes to its notable ability to invade new aquatic habitats (Lee and Bell 1999; Reid and Orlova 2002), and this invasiveness is predicted to be exacerbated by climate change (Meek et al. 2012). *Cordylophora* likely spread globally from the Ponto-Caspian region via ship ballast water or ship fouling and is a relatively recent, potent invader of North America. It has been found throughout the United States including the Great Lakes and estuarine systems on the east and west coasts ((Mills et al. 1993; Cohen 1998; Ruiz et al. 2000; Pienimäki and Leppäkoski 2004; Streftaris N 2005; Wonham and Carlton 2005). *Cordylophora* is a biofouler of human-made structures such as power plants (Rajagopal et al. 2002; Folino-Rorem and Indelicato 2005; Folino-Rorem 2015), and is a predator which may disrupt the community structure of the benthic invertebrate communities which it invades (Folino-Rorem 2015). In addition to sexual reproduction,

Cordylophora can spread through asexual budding, producing dense colonies, as well as dispersing via dormant, drought-resistant fragments called menonts (Roos 1979). *Cordylophora* therefore represents an excellent system for studying the evolutionary transition towards fresh water tolerance.

Given the difficulty in delineating species within *Cordylophora* using morphology alone, genetics has been a useful tool for examining cryptic diversity and species boundaries in this genus. Previous work suggests that *Cordylophora* may in fact be made up of several genetic lineages which have distinct salinity regimes but broad geographical ranges (Folino-Rorem 2009). Specifically, using evidence from two mitochondrial loci (16S and CO1) and one nuclear marker (28S), Folino-Rorem et al. (2009) recovered three major monophyletic lineages: Clade 1A, containing freshwater genotypes collected from inland lakes in North America and Europe; Clade 1B, containing populations from a broad range of locations and salinity levels, and Clade 2, largely containing samples from the west coast of the U.S. collected at brackish salinities. Estimates of molecular divergence between these three lineages were found to be comparable to the interspecific distance range for a closely related genus of hydrozoans (Folino-Rorem 2009 et al.). These results suggest that different salinity tolerances evolved within the *C. caspia* species complex, rather than comprising a single species whose members are all euryhaline.

Species-level divergences are supported by experimental evidence, where it has been demonstrated that genotypes from Clade 1A (inland fresh water) and Clade 2 (brackish) possess optimal growth rates in distinct salinities and demonstrated observable phenotypic responses to changes in salinity (Folino-Rorem and Renken 2018). This study suggests that there is a genetic basis for differences in salinity tolerance between lineages of *Cordylophora*. The existence of

lineage-specific salinity tolerance differences between lineages suggests that salinity may serve as a barrier to gene flow in *Cordylophora*.

In this work, we perform species delimitation analyses for *C. caspia*, (*sensu lato*) and investigate the relationship between ecology, geography and species boundaries. In order to determine whether salinity-related differentiation rises to the level of speciation in *Cordylophora*, we use a multiplexed shotgun genotyping RAD-seq approach (Baird et al. 2008, Andalfatto et al. 2011) to reconstruct genome-wide patterns of differentiation, permitting determination as to whether distinct lineages have evolved in conjunction with distinct salinity tolerances. RAD-sequencing has been successfully used to provide species delimitation estimates in many groups of understudied or phylogenetically recalcitrant taxa and often reveals unknown diversity in the group of interest (Escudero et al. 2014; Pante et al. 2014; Herrera and Shank 2016; Chan et al. 2017).

We sampled extensively in a diversity of salinity habitats. In particular, we focused collecting efforts on the under-sampled but ecologically complex environments of the salinity gradient of North American estuaries, which appear to have been recently invaded (Wonham and Carlton 2005, Streftaris et al. 2005) and are potential zones of secondary contact for the diverged lineages of *Cordylophora*. To aid in species delimitation we test whether *Cordylophora* sampled from estuaries with a wide range of salinities are inhabited by one euryhaline lineage or multiple lineages with distinct salinity tolerances, and whether salinity serves as a barrier to gene flow. Using this phylogeny constructed from RAD-seq data, we reconstructed ancestral character states for salinity tolerance ranges in this group and used the phylogeny as the basis for hypotheses of species delimitation using modern Bayesian techniques (Leaché et al. 2014). Further, we leveraged the RAD-data to examine the fine scale population structure along several

of the sampled North American estuaries to confirm hypotheses of species delimitation. We clarify the species-level relationships within *C. caspia sensu lato*, identify cryptic diversity, patterns of salinity tolerances, and predictions of gene flow between lineages, thus providing insight into the patterns and processes involved in the evolutionary transition to fresh water from a marine ancestor.

MATERIALS AND METHODS

Sample Collection Methods

Previous researchers have found that a small sampling of populations is unlikely to be representative of all populations within an invasive species complex (Lee 2015), so it was important to sample across both the complex salinity environment of the estuaries and to expand sampling of the *Cordylophora* complex in general. Thus, we collected colonies of *Cordylophora* from multiple locations along several estuary systems in the United States, in addition to broadly enhancing sampling throughout the invaded North American range. In particular, we focused on collection along the Columbia River, San Francisco Bay, and James River estuaries (see Fig. 4.1 for an example collection scheme for estuaries) and locations along the Great Bay in New Hampshire and the New York Finger Lakes, as well as harbors in Lake Michigan (see Table 4.1). Where applicable, proper permissions were obtained to collect samples from private marinas and permits were acquired for collection within Virginia State Parks. Notably, collection success at relatively high salinity locations (i.e. >12 psu) was limited due to what appeared to be out-competition of *Cordylophora* by other encrusting organisms (i.e. sponges, bryozoans, marine hydroids) (E.S.C, personal observation).

The exact collection method we utilized depended upon the environment in which the colonies of *Cordylophora* were found. This collection often entailed scraping large pieces of

substrate and manually examining the pieces for *Cordylophora* tissue, or extracting hydranths from the substrate directly when colonies were more visible. Samples of *Cordylophora* were immediately placed in 100% ethanol for future DNA extraction. Where possible, in locations where *Cordylophora* was very abundant, samples were taken at least a meter apart to avoid re-sampling the same colony, in line with data from (Darling and Folino-Rorem 2009) demonstrating that the scale of clonality for *Cordylophora* is less than a few feet. At each location, current salinity was measured using either a Deepwater Aquatics ATC Refractometer or YSI meter.

DNA Extraction and Sanger-Marker PCR

We pooled three to ten hydranths per colony for DNA extraction and for colonies with only stolonial material collected, we manually macerated this material with a plastic pestle prior to the application of the initial Proteinase K step. DNA was extracted with the QIAGEN DNeasy Blood and Tissue kit with several modifications to increase DNA purity and yield: Samples were left to digest in Proteinase K for an extended overnight period so that skeletal material could be digested, 4ul of RNase A was added to each sample post-digestion, and the final elution step was repeated, finishing with an elution into 30ul of EB Buffer.

Generation and analysis of Sanger-sequenced markers

PCR amplification. We carried out PCR amplification of 16S, 28S and CO1 sequences from extracted DNA using the same primers and thermocycler programming as described in Folino-Rorem et al. (2009). All PCR products were run out on an agarose gel and products of the expected molecular weight were sent to GeneWiz (Plainfield, NJ) to be sequenced directly in

both forward and reverse directions using the amplification primers.

Phylogenetic analyses. Newly sequenced 16S, 28S and CO1 sequences were trimmed and assembled in Geneious 8.1. Sequences from Folino-Rorem et al. 2009 were downloaded from NCBI and were aligned with the newly generated sequences for each marker using default settings in Geneious 8.1. These alignments were further edited by eye in the Geneious 8.1 alignment viewer. We also created concatenated alignments for each possible combination of the three markers for every colony that was successfully sequenced for each pair of markers resulting in alignments containing multiple Sanger markers.

For each alignment, maximum likelihood phylogenetic reconstruction was carried out using the IQTree algorithm (Nguyen et al. 2015) on the IQ-Tree Web Server (Trifinopoulos et al. 2016) with automated choice of sequence evolution model via ModelFinder (Kalyaanamoorthy et al. 2017). For concatenated alignments, partitioned models of sequence evolution were used (Chernomor et al. 2016). Support for each tree was assessed with 10,000 ultrafast bootstrap replicates (Hoang et al. 2018). Tree visualization was carried out using the GGtree package for R (Yu et al. 2016; Team 2017). Information about each Sanger-marker data set and the models chosen by ModelFinder can be found in Table 2. Trees were rooted using one or more of closely related marine hydrozoans *Turritopsis rubra*, *Leuckartia octona* and *Clava multicornis* depending on sequence availability, as in Folino-Rorem et al. (2009). Each of the Sanger-marker phylogenies recovered the same three major clades that were also congruent with those recovered in Folino-Rorem et al. (2009). For each of the major clades recovered in the single-marker analyses, we calculated measures of phylogenetic support and diagnosability of species using the Species Delimitation Geneious plugin (Masters et al. 2010).

RAD-sequencing and phylogenetic analyses

Library preparation. To obtain genome-wide sequencing representing loci across the entire genome, we took a restriction site-associated-digest (RAD-Seq) approach (Miller et al. 2007; Baird et al. 2008) using a modified multiplex-shotgun-genotyping (MSG) method (Andolfatto et al. 2011) as implemented in Monnahan et al. (2015) and Chang et al. (2018). Because initial rounds of MSG library preparations yielded very low amounts of DNA, we opted use the frequently-cutting restriction enzyme AseI (NEB Biolabs, Ipswich, MA, USA) for DNA digestion to increase the overall amount of input DNA for sequencing. In total, 192 samples were prepared in two 96-sample batches for sequencing, with the assistance of the University of Kansas Genome Sequencing Core (Lawrence, KS) (GSC). To allow for demultiplexing of sequences from different samples, each sample of our first 96 samples was ligated to one of 48 unique barcode adaptors, and each such set of 48 samples was pooled independently. Each of these sublibraries was then combined with different Illumina indices during the PCR step of the library preparation (Monnahan et al. 2015), and a total of 14 rounds of PCR was performed on the pooled products. The second library, containing a second set of 96 samples, was ultimately subdivided into six sublibraries, which were pooled and had between 6 and 14 rounds of PCR performed. For both libraries, we added 10% PhiX spike-in to increase library complexity for sequencing. Illumina sequencing of the two libraries was performed by the GSC in one lane each of high-output paired-end 100bp reads on an Illumina HiSeq 2500 System, although only forward reads were used in further analyses.

Processing reads. The raw Illumina sequence data were quality-filtered and demultiplexed into sample-specific FASTQ files using the `process_radtags` program of Stacks

v.1.44 (Catchen et al. 2013) on default settings. At this stage, sequences for several samples that had fewer than <10,000 retained reads after the process_radtags step were removed from further analysis.

Reconstruction of homologous loci. Sample-specific FASTQ files trimmed by Stacks were used as input into the ipyrad v. 0.7.24 (Eaton 2014) pipeline for *de novo* assembly of homologous RAD-loci beginning at Step 3, which is the within-sample clustering of homologous loci based on sequence similarity. Subsequently, ipyrad does joint estimation of heterozygosity and error rate, consensus base calling and filtering, aligning and clustering reads among samples, and the final filtering and export of loci to a variety of output formats (Eaton et al. 2014). To get an initial sense of how successful our sequencing efforts had been, we ran the entire ipyrad pipeline first on default settings. We examined our results including visualization of patterns of missing data in the data set using the MatrixCondenser Web app (https://bmedeiros.shinyapps.io/matrix_condenser/).

During this analysis, we detected a pattern of non-overlapping markers that correlated with the clades recovered from the Sanger sequenced markers (Fig. 4.4). In order to increase the overall number of loci spanning across the entire group and reducing the amount of missing data, as well as to test the effect of assembly parameters in downstream analyses, we tested several important assembly parameters over several more runs of the pipeline. One such parameter was the “mindepth_majrule”, which we lowered to four (from a default of six), which lowers the minimum depth at which majority rule based calls will be made by ipyrad during the estimation and consensus base calling and filtering steps. A lower number should result in more loci with base calls but potentially an underestimation of heterozygosity rates. Additionally, we tried several values of the clustering threshold (lowered to .80 from a default of .85) which is the level

of similarity at which sequences are considered homologous and clustered together, affecting both the within- and between-sample clustering steps. Lowering this parameter allows for more divergent loci to be identified as homologous, and therefore can increase the sample coverage of a given locus, but may also cause truly non-homologous RAD-loci to be treated as homologs. Finally, we also tested a larger number of indels allowed per locus (10 allowed rather than 8), higher values of which may help retain loci assembled between more distantly related individuals.

A full summary of the parameter sets and the number of loci assembled in each run can be found in Table 4. We found that a final assembly with a `mindepth_majrule` of four, a clustering threshold of the default 0.85 and an increase of the number indels allowed per loci to 10 increased the total number of loci retained in future analysis and the amount of data overlap between clades (Fig. 4.4). Therefore, further analyses were done using this assembly (referred to as `Depth4_clust85_Indel10_Min8`) unless otherwise noted. At this step we discarded additional individuals with fewer than 100 assembled loci.

Phylogenetic analyses: One major goal of our study was to facilitate species delimitation and reconstruction of trait evolution by reconstructing a well-supported tree of the *Cordylophora* species complex using genome-scale data. To do so, from the `Depth4_c85_indel10` ipyrad assembly described above, we exported data sets containing different subsets of individuals and loci in order to examine the effects of taxon sampling and proportion of missing data on tree topology and branch support. To investigate the role of taxon sampling, we created data sets that consisted of just the three colonies with the highest number of sequenced loci from each collecting site, and the set of all colonies that had passed previous filtering steps. For each set of colonies, we tested three different minimum numbers of individuals a locus must be present in to

be exported; minimum four individuals, minimum eight individuals, and minimum 50% of individuals. Four individuals represents the fewest number of individuals needed for a loci to be potentially phylogenetically informative (i.e. quartet-informative, Eaton et al. 2017). For each of these analyses, we examined patterns of missing data (using MatrixCondenser as above) and generated a phylogeny as described below (see Table 4.2 for the number of loci and individuals in each data set).

Each data set was exported from ipyrad as a phylip-formatted file containing one SNP per RAD-loci (-u output format option), as including multiple SNPs from the same assembled locus may capture signatures of linkage. The non-variant sites in these files were removed on the IQ-Tree Web Server to generate alignment files containing only SNPs. These files were then used as input for maximum likelihood phylogenetic reconstruction using the IQTree algorithm (Nguyen et al. 2015) on the IQ-Tree Web Server (Trifinopoulos et al. 2016) with automated choice of sequence evolution model via ModelFinder (Kalyaanamoorthy et al. 2017), allowing for free-rate heterogeneity (+R) and including the ascertainment bias correction (+ASC) to condition the likelihood on the use of only variable sites in our SNP alignments. Support for each tree was assessed with 10,000 ultrafast bootstrap replicates (Hoang et al. 2018). To assess the value of combining SNP and Sanger markers for phylogenetic analysis we also created alignments by doing a partitioned analysis with our combined SNP and Sanger data for each individual that was present in both our RAD-sequencing data set and for that particular marker. These were also run on the IQ-tree web server with separate data partitions for the two types of data in each alignment. Descriptions of these data sets and the models selected by ModelFinder can be found in Table 4.2.

Tree visualization was carried out using the GGtree package for R (Yu et al. 2016). Because taxa outside of the *Cordylophora* species complex were not sequenced in the RAD-sequencing data set, phylogenies produced with our RAD-data were rooted manually at the node recovered in our Sanger-marker phylogenies. For further analyses based on the RAD-sequencing-only tree, we used the tree based on a minimum of eight individuals per loci. Tree topology using this dataset was consistent with all other RAD-data trees produced with different levels of this parameter, but possessed the highest average bootstrap support across the phylogeny. As for our 16S tree, we calculated measures of clade differentiation using the Species Delimitation plugin for Geneious (Masters et al. 2010).

Trait evolution and ancestral state reconstruction

For the RAD-sequence tree presented in Figure 4.5, we tested whether the major clades we recovered had significantly different ancestral salinity range, as suggested in Folino-Rem (2009). We used salinity data recorded at the collecting site for each colony in the phylogeny as the response variable and clade as an explanatory variable to perform the non-parametric Kruskal-Wallis test for comparison between all subclades, and a Welch's t-test to compare mean salinity between major clades 1 and 2. We created boxplots using ggplot2 (Wickham 2009) with the `stat_compare_means()` function from the ggpubr package (<http://www.sthda.com/english/rpkgs/ggpubr/>) in order to test and visualize significance between every different pairwise comparison between the clades. For the tree in Figure 4.5, we also performed Bayesian ancestral character state reconstruction in R using the `anc.bayes()` function from the phytools package (Revell 2011), which uses Bayesian MCMC to sample from the posterior distribution for states at the internal nodes of a given tree. We ran the MCMC sampler

for 10000 generations per internal node, sampled every 1000 generations. Average inferred ancestral salinity levels were represented for selected nodes using Adobe Illustrator CC 2018 to produce Figure 4.7.

Tests for gene flow within and between major clades of *Cordylophora*

Given that our phylogenetic analyses recovered several groupings within Clade 1 (Fig. 4.5), we wanted to assess whether or not these lineages were undergoing gene flow between them. Specifically, we wanted to test whether or not individuals in different subclades of Clade 1 (Fig. 4.5) are actually distinct biological species from one another despite some of them co-occurring in the same estuary systems. Structure 2.3.4 (Hubisz et al. 2009) was used to determine the number of genetic clusters amongst all individuals that fell in Clade 1 according to Figure 4.5. To minimize the amount of missing data, we required that a locus be recovered in at least 50% of all Clade 1 individuals, resulting in a data set of 3221 unlinked SNPs and 60 individuals. An initial short run with $K=1$ was carried out in order to estimate Lambda (actual independence between markers), and settings for further runs were run at defaults (with ADMIX=TRUE to allow for the possibility admixture), besides adjusting parameter lambda to match the estimated value for our data set. K values (number of clusters) between one and ten were evaluated, with ten replicates for each K and 10,000 burn-in followed by 100,000 Markov chain steps for each replicate. The most likely number of clusters was determined using the method of Evanno et al. 2005 as implemented using the structureHarvester online server (Earl and vonHoldt 2012). Results for each value of K were summarized and visualized using the CLUMPAK online server (Kopelman et al. 2015).

Because the most likely number of clusters was two for each analysis (Appendix 26 and 27), this gave us the opportunity to assess which individuals, if any, had shared ancestry between the two major recovered clusters by calculating their hybrid indices (Buerkle 2005). We chose individuals with a probability (Q-score from Structure) of 1.0 of ancestry from one or the other subclades as parental populations and any individuals with less than 1.0 as potential admixed individuals for the hybrid index analysis. Hybrid indices, which measure the genetic contribution of each parental population to putatively admixed individuals, were calculated using the Introgress R Package (Gompert and Buerkle 2010). The same data set was used for this analysis as in the Structure analysis, except that any locus that had missing data for all members of any group (P1, P2 or admixed) had to be removed to calculate the hybrid indices resulting in a final data set of 1962 loci.

Species Delimitation Analyses

To supplement the measurements of population structure described above, and to directly compare different hypotheses of which clades of *Cordylophora* represent different species, we performed the Bayes Factor Delimitation (BFD) method (Leaché et al. 2014). In BFD, individuals in a data set are given *a priori* species assignments, and the likelihood of this configuration of individuals into species is calculated given the input SNP data set. Species can then be reassigned to different models of species delimitation and the likelihoods recalculated for different scenarios of species delimitation. The different proposed models of species delimitation within Clade 1 that we compared are listed in Table 4.8. Since these marginal likelihood calculations are computationally intensive, we pruned the number of individuals in Clade 1 to 37

individuals spanning all of the major subclades, and considered only loci recovered in 50% or more of these individuals, resulting in a data set of 1239 SNPs.

We implemented BFD using the SNAPP (SNP and AFLP Package for Phylogenetic Analysis) (Bryant et al. 2012) package for the BEAST2 software program (Bouckaert et al. 2014). The marginal likelihood of each model was estimated via path sampling using 24 steps, an alpha of 0.3, and a MCMC chain length of 1,000,000 with a pre-burnin of 10,000. These analyses were run with the following relatively un-informed priors because we had little knowledge of true values for these parameters in our system, following Chan et al. (2017): mutation rates (u and v) and the shape parameter for the gamma distribution prior on population sizes (α) were set at 1.0; the beta scale parameter was set at 350 and the speciation rate prior (λ) was sampled from a broad gamma distribution of $\alpha = 2$ and $\beta = 250$. We considered the posterior distribution to be adequately sampled when effective sample size values for parameters were >200 , as suggested by Bouckaert et al. (2014). Once marginal likelihoods were calculated for each delimitation scenario, natural log Bayes factors (BF) were used to compare the log marginal likelihoods (MLE) of competing models using the equation $BF = 2[MLE(\text{model1}) - MLE(\text{model2})]$, with Model1 always referring to the more complex model with more hypothesized species (Leaché et al. 2014).

RESULTS

Collection of *Cordylophora* from across North American Range

New samples of *Cordylophora* were obtained from a wide variety of locations throughout its invaded range in North America, including thirteen new collecting sites representing several different estuary systems on the East and West coasts of the United States, as well as greatly enhancing sampling in previously sampled locations such as Lake Michigan

and the Finger Lakes (Table 4.1). Unfortunately, not all samples present in the Sanger phylogenies of Folino-Rorem et al. (2009) were represented by contemporary biological material so were therefore not included in this newly generated RAD-sequencing library, nor was every single colony from a given location Sanger-sequenced for each marker (see data sets in Table 4.2).

Results of phylogenetic analyses using Sanger-sequenced markers

Phylogenies inferred from individual and concatenated Sanger-sequenced markers 16S, 28S and CO1, including newly collected North American samples, are largely congruent with one another and with those produced in previous molecular phylogenies of the *Cordylophora* species complex (Fig. 4.2 and Appendices 17-18). All phylogenies recover two major clades, designated 1 and 2 after Folino-Rorem (2009). Given the overall congruency between the trees produced with different Sanger-sequence markers, and the fact that our 16S sequence possessed the best taxon sampling, we focus on results from this phylogeny (Fig. 4.2). Further, the 16S region of the mitochondrial genome is commonly used as a species-level barcoding sequence for hydrozoans (Zheng et al. 2014; Lindsay 2015; Miglietta et al. 2015).

Within the two major clades, several subclades are consistently recovered. In Clade 1, Clade 1A (bootstrap support of 98) consists of individuals collected from freshwater, including the Great Lakes, Finger Lakes and inland European Lakes, and also includes newly sequenced samples from inland, low-salinity portions of the James River. Clade 1B consists of samples from a wide variety of salinities and geographic regions, including representatives from all of newly sampled estuaries: Great Bay, James River, the Columbia River and the San Francisco Bay. In all phylogenies, we consistently recover a second clade (Clade 2), which consists of

individuals solely from the two estuaries on the West Coast (2A), and 2B, which includes estuarine samples from the Great Bay as well as samples from Europe and South America (2B). In contrast to Clade 1 which contains mostly colonies sampled from freshwater, all individuals in Clade 2 were collected at higher salinities (4.0 psu and higher).

Statistical measures of clade separation and species diagnosability used in DNA barcoding are presented in Table 4.3. One measure we include is the ratio of within-clade to between-nearest-clade distances, which has been demonstrated to be an effective measure of species diagnosability (Ross et al. 2008). A within-to-between-clade distance ratio of 0.25 or smaller is considered to be evidence that the clades could be distinct species (López-López et al. 2012; Churchill et al. 2014), which is observed in each comparison in Table 4.3. Also calculated was the probability of correctly identifying an unknown member of the putative species given their 16S sequence, i.e. the P ID(Strict), which is calculated from a third-order polynomial regression describing the probability of successful identification into pre-defined species assignments as a function of the within-to-between-clade distance ratio above (Masters et al. 2010). Although there is no strict species-level cutoff for this value, it should be noted that these values are very close to 1.0, except in the case of Clade 2B (Table 4.3), which supports the separation of Clades 1A, 1B and 2 as separate species. Clade 1A and 1B are particularly distinct from one another, with very low ratios of within- to between-clade differences. Although the major clades and subclades are well supported (with the exception of the node support for sample Dp13 and the rest of Clade 1) bootstrap support is relatively low within each of the subclades. Finally, we did not find that concatenating Sanger-markers changed major topology or markedly increased branch support, and in fact had the negative effect of decreasing taxon

sampling or increasing missing data in each concatenated tree as compared to our 16S phylogeny (Appendices 19-20).

We also conducted tests to see whether or not our recovered clades had significantly different average salinities from one another (Fig. 4.3). We tested every pairwise comparison between different subclades (values on brackets in Fig. 4.4) and also calculated an overall significance value for a test of no difference between clades (Kruskall-Wallis test, $p=7.9e-14$). When comparing the salinities of environments in which individuals were collected, Clade 1A is limited to freshwater or nearly freshwater habitats and has a significantly different average salinity when compared to any other subclade. Subclades of Clade 2 are not significantly different from one another, although Clade 2A has a much wider range of salinity values than 2B. Overall, the average salinity value for Clade 2 is significantly higher than that of Clade 1 (3.65 psu in Clade 1 vs. 12.31 psu in Clade 2, $t = -5.5465$, $df = 60.905$, $p\text{-value} = 6.68e-07$ for Welch's two-sample t-test). The patterns in this 16S phylogeny thus suggest that Clade 1 is a primarily fresh water clade (with a few outliers) and Clade 2 is a euryhaline clade, and does not include any truly freshwater representatives.

RAD-sequencing Data Set

After several rounds of filtering and exclusion of individuals with particularly low numbers of loci, the final data set included 164 individuals from localities listed in Table 1. One feature of this data set, particularly when assembled with default parameters, is a striking pattern of missing data wherein there was a noticeable non-overlap between the RAD-loci recovered in individuals from Clades 1 and 2 (Figure 4.4A) and no RAD-locus was present in all 164 of the retained individuals (Table 4.4, last column). This effect was somewhat alleviated by adjusting

filtering parameters to be less stringent, such as the minimum sequencing depth for a locus and the number of insertions/deletions allowed (Figure 4.4B). This took the data set from 87% missing data in to 76% missing data. Overall, adjusting assembly parameters had the expected effects on the assembly, with decreasing the minimum read depth increasing the number of loci present, and increasing the number of individuals a locus must be present reducing the number of loci (see parameter sets in Table 4.4). We noted a negative relationship between matrices of relatedness (calculated via 16S) between pairs of individuals and the number of shared loci between them (Mantel test z-score: 1721217, p-value <.001) which is a well known pattern in RAD-sequencing data due to mutations affecting homologous cut sites (Herrera et al. 2015, Eaton et al. 2017). However, despite the pattern of missing data, it has been shown that for large data sets, well-supported phylogenies can be produced with large amounts of missing data (Eaton et al. 2017; Tripp et al. 2017).

Reconstruction of RAD-sequencing trees

We found that adjusting the assembly parameters did not affect our overall topology except for those that drastically reduced the number of loci (i.e. only including loci present in a minimum of 50% of individuals, compare Appendices 22 and 23). Reducing the taxon set also did not affect the major clades recovered (see phylogeny including just three samples per site, Appendix 22, in comparison with Fig. 4.5). Given this, the focus for the rest of this work is on the phylogeny inferred from our final assembly (Depth4_clust85_Indel10_Min8) that includes the largest possible sampling of individuals (Fig. 4.5).

In the phylogeny reconstructed from our RAD-seq assembly, we broadly recover the same major monophyletic groups as in our 16S phylogeny, keeping the clade naming scheme

consistent with the 16S phylogeny. With the increased sampling, we also recovered several major ecological and biogeographical patterns (discussed below), as well as increased resolution between individuals in the same clade and higher bootstrap support overall (Fig. 4.5).

This phylogeny recovered separation between East Coast (Clade 1), and West Coast (Clade 2) invasions. All of the West Coast locations (with the exception of some freshwater members) fall into Clade 2, and vice versa. More specifically, all samples from the Chesapeake Bay estuary, the Great Bay estuary and the Great Lakes and Finger Lakes are in Clade 1. Clade 1A has the same basic membership as in Fig. 4.2 with the addition of samples from the nearly freshwater portions of the James River and Great Bay, as well as freshwater locations in Irvine (Iv) and Brannan Isle (Br) CA. We also recover a small subclade, designated 1C (not present in 16S phylogeny), made up solely of individuals from a brackish location on the James River (Kingsmill Landing, Kml, 4.4 psu), and a larger subclade (1B, also recovered in the 16S phylogeny) made up of samples from the Great Bay and several samples from the United Kingdom. Newly collected colonies from inland of their respective west-coast estuaries are recovered in Clade 2, which increase the lower salinity limit of Clade 2 from that in the 16S tree in Figure 4.3. Clade 2 contains monophyletic groups within 2A that are either from Columbia River or the San Francisco Bay, respectively.

As for the 16S phylogeny, we calculated statistical measures of clade separation and species diagnosability (Table 4.5). Unlike what we recovered for our 16S phylogeny, we find that most intra-clade to inter-clade ratios of genetic distances are above the 0.25 cutoff for species-level divergence, with the exception of Clade 1B and 1C. However, more similar to the case of our 16S data set, the PID (Strict) values (the mean probability of correctly identifying an unknown member of the putative species using criterion that it must fall within clade), are all

close to 1.0 and above 0.5. These values suggest in theory one could use RAD-sequencing data as a means of identifying members of the different clades and subclades reliably. The use of an entire RAD-seq data set gives us a less clear signal than a single Sanger-marker, as the single marker represents the signal from a single gene tree. On the other hand, the many loci in the RAD-seq data sets presents us with an enhanced ability to detect incongruence between loci due to factors such as incomplete lineage sorting and the different evolutionary histories of different parts of the genome.

Comparing the salinity ranges between clades in this phylogeny (Fig. 4.6), Clade1A is lower in salinity than all other Clade 2 subclades besides 2C, largely consistent with what we found in Figure 3. We also detect some differences in average salinity in the subclades of both major clades (Fig. 4.6) and calculated an overall significance value for a test of no difference between clades, which was significant (Kruskall-Wallis test, $p=6.9e-16$). Overall, the average salinity value for Clade 2 is significantly higher than that of Clade 1 (2.0 psu in Clade 1 vs. 6.94 psu in Clade 2, $t = -5.7121$, $df = 113.21$, $p\text{-value} = 9.145e-08$ for Welch's two-sample t-test). This is approximately the same mean difference in salinities between Clade 1 and Clade 2 (approximately 5.0 psu) as recovered in our analyses using the 16S phylogeny, above.

Overall, there is evidence that each estuary is made up of individuals from at least two subclades with different salinity regimes (see tip label colors in Figure 4.6), suggesting that colonies of *Cordylophora* inhabiting estuaries are at least partially partitioned by salinity. This is particularly striking in that inland members of three different estuaries were recovered in the exclusively freshwater Clade1A.

Finally, as in the concatenated Sanger-marker trees, we did not find that concatenating Sanger-markers to our RAD-sequencing data set changed major topology or markedly increased

branch support, and in fact had the negative effect of decreasing taxon sampling included in each concatenated tree as compared to the main RAD-data phylogeny in Figure 6 (not all combination of markers tested due to computation constraints, Appendix 21).

Ancestral Character State Reconstruction using RAD-sequencing trees

Using the phylogeny described above, we reconstructed the evolutionary history of the salinity trait in the *Cordylophora* species complex using a Bayesian approach. Full results of ancestral character state reconstruction for each internal node can be found in Appendix 25 (key to node numbers in Appendix 24) and the average predicted values for selected nodes are visualized in Figure 4.7. Most ancestral nodes, including the ones for each of the major Clades 1 and 2, as well as for the whole species complex, are predicted to have nearly freshwater salinities (<2.0 psu). Subclade 1A and more shallow nodes within 1A are predicted to be even closer to being freshwater. Subclade 1B (containing largely relatively high salinity individuals from New Hampshire collecting sites) has a relatively high salinity ancestor. The clade containing individuals from Oregon collected at nearly marine salinities (marked with 5.9, collecting site Gar), also has a brackish inferred ancestral salinity. None of the three most ancestral nodes have high predicted ancestral salinities (<2.5 psu), suggesting that freshwater tolerance may have evolved outside of the *Cordylophora* species complex.

Population Genetics Analyses of Clade 1

We examined population-level boundaries and potential for gene flow within Clade 1A and 1C, both excluding and including Clade 1B as a potential source of genetic material. Figure 4.8 summarizes the results from our population structure analyses, with each individual

represented by a vertical segment, with colored segments representing the probability that this individual is a member of each of the two inferred populations. For the analysis of population structure in subclades 1A and 1C, the Evanno method (Evanno et al. 2005) for selecting the optimal number of clusters of individuals (Appendix 26) clearly suggested two clusters ($K=2$), after evaluation of K values from 1 to 10. Most individuals from Clade 1A had a probability (Q-score) of 1.0 for being descended from one putative population (solid blue bars from Figure 1A), and all individuals from Clade 1C were shown to be from the other predicted cluster (solid orange bars from Figure 4.8A). Non-admixed Clade 1C individuals came from Kingsmill Landing (4.4 psu) and Chippokes Plantation State Park (Chp, 1.1 psu) (depicted in solid orange). The entirety of samples collected in Great Lakes Region (BH, LME, DP) fell within the other cluster (depicted in solid blue). Some estuarine samples in Clade 1A from James River (Jcc, Chp) and California (Br, Is, Iv) appear to have roughly equal probabilities of assignment to the two putative clusters (combined blue and orange bars in Fig. 4.8A). For further work, we considered any individual with Q-score equal to 1.0 for either of the two inferred clusters to be members of those clades (i.e. represented by only one color in Figure 4.8A and 4.8B) and any individuals with less than 1.0 to be potentially admixed individuals. In total, out of sixty individuals, there were 25 admixed individuals and 27 non-admixed individuals of Clade 1A (blue cluster, Fig. 4.8A) and seven non-admixed individuals from clade 1C (orange cluster, Fig. 4.8A).

Population structure analyses of all of Clade 1 (including Clade 1B) also yielded two putative clusters (Appendix 27). Most individuals from Clade 1A were from one predicted cluster, similar to that recovered in Figure 1A, but in this analysis, Clade 1C individuals cluster with 1B rather than 1A. One cluster (solid orange bars in Figure 4.8B) includes members of NH

estuary (Ex, Clade 1B), Kingsmill Landing (4.4 psu) (Clade 1C), some members of relatively freshwater Newmarket, NH, and some members of Guk and Chp (Clade 1B). Members of Great Lakes locations (Bh and Mb) were entirely within the other (blue) clade. Once again in this analysis, some estuarine members are potential hybrid individuals (those with both colors in their bars, representing individuals from Jcc, Is, Iv, Br). For further work, we considered any individual with Q-score equal to 1.0 for either of the two inferred clusters to be members of those clades (i.e. represented by only one color in Figure 4.8A and 4.8B) and any individuals with less than 1.0 to be potentially admixed individuals. In total, out of sixty individuals, there were 31 non-admixed individuals from Clade 1A (blue cluster), 27 potentially admixed individuals and 20 non-admixed individuals from Clade 1C+1B (orange cluster).

Hybrid Index analysis of Clade 1

Because we obtained two clusters in both of our structure analyses, we were able to calculate a hybrid index (HI) for each putatively admixed individual. The hybrid index uses allele frequencies to calculate a measure that represents the amount of genetic contribution from either of two parental populations (Buerkle 2005). This index can range from 0 to 1, 0 meaning 100% genetic contribution from one population and 1 meaning 100% contribution from the other parental population, and we treat anything between .25 and .75 as being a potential hybrid individual (Buerkle 2005; Gompert and Buerkle 2010). Hybrid indices follow nearly the exact same pattern as the Q-scores (probability of being descended from one or the other population) calculated using Structure (Tables 4.6 and 4.7). Calculation of the hybrid index confirms that some individuals are hybrids of Clade 1A and 1C when excluding Clade 1B (Table 4.6), and that when including Clade 1B, the same individuals now appear to be hybrids of Clade 1A and Clade

1B (Table 4.7). Clade 1C in particular appears to be B-like ($HI > 0.75$), rather than A-like (Fig. 4.8).

Bayes Factor Delimitation analysis

The results from using the Bayes Factor Delimitation method comparing different models of species delimitation can be found in Table 4.8. Wholly positive Bayes Factor values indicate that the model where each subclade of Clade 1 (Model 1A+1B+1C+2) are each a separate species is the most likely model. Smaller Bayes Factor values mean that there is a smaller difference in likelihood between this model and model in question, suggesting that the model in which Clades 1B and 1C are grouped together (Model 1A+1BC+2) is the second most likely model of species boundaries. Further, the model where only major Clades 1 and 2 are separate (no speciation within Clade 1) is actually a worse model than one where designations are purely geographical (i.e. separated by body of water).

DISCUSSION

RAD-seq phylogeny of the *Cordylophora* species complex

Data from RAD-sequencing produced a well-supported maximum-likelihood phylogeny of the species complex, providing new detail into evolutionary patterns within *Cordylophora*. Despite a large amount of missing data and substantial non-overlap between loci sequenced in the two major clades of *Cordylophora* (Fig. 4.4), our resultant tree (Fig. 4.5) displayed better bootstrap support, particularly at deep nodes, and more resolution within each clade, as compared with the trees recovered from Sanger-sequenced markers (Fig. 4.3, Appendices 17-20). Our ability to resolve this phylogeny despite large amounts of missing data (77% in our final

data matrix) places it among a number of studies that suggest the relative importance of the large number of informative characters produced by RAD-seq over the proportion of missing data for increasing phylogenetic resolution. For example, (Tripp et al. 2017) were able to create a fully resolved phylogeny of a radiation of the desert-dwelling *Petalidium* plants using a matrix with over 90% missing data, and (Pante et al. 2014) were able to resolve relationships within a genus of deep sea corals using matrices with up to 83% missing data. In general, RAD-sequencing appears robust to varying degrees of relatedness between taxa, having been used successfully to investigate systems ranging from cichlid species diverged <15,000 years ago (Wagner et al. 2013) to deep sea corals diverged from one another ~80mya (Herrera and Shank 2016). Our approach involved relaxing certain assembly parameters in order to include more SNPs in the final matrix, which is an approach that has been shown to increase phylogenetic resolution both empirically and in simulation (Rubin et al. 2012; Wagner et al. 2013; Wessinger et al. 2016; Eaton et al. 2017; Tripp et al. 2017).

Evidence for multiple lineages of *Cordylophora* in estuaries

By incorporating detailed sampling of *Cordylophora* from North American estuary systems (San Francisco Bay, James River, Columbia River, Great Bay), we were able to determine that the salinity gradient in estuaries is inhabited by multiple lineages with distinct salinity regimes, rather than a single, euryhaline lineage. It might be expected that estuaries would contain multiple genotypes, given the ecological complexity of the habitat as well as their contact with vectors of introduction of invasive species, leading to repeated invasions (Williams and Grosholz 2008). Partitioning of an estuary with different populations or species has been uncovered in many groups of marine invertebrates (some reviewed by Bilton et al. 2002),

sometimes rising to the level of potential species-level lineages (Chen and Hare 2008) or even the evolution of distinct ecomorphotypes in different parts of an estuary (Ueda et al. 2011). The work of (Folino-Rorem et al. 2009) recovered multiple lineages present at particular estuarine collecting sites, but no systematic analysis along gradients of multiple estuaries had been previously undertaken for *Cordylophora*.

Overall, we found some evidence of partitioning-by-salinity in that most of the genotypes collected in the lowest salinity regions of three of the four estuaries sampled, the James River, the San Francisco Bay and the Great Bay, fell within the freshwater Clade 1A, which also contains samples from the inland Great Lakes and Finger Lakes (Fig. 4.5). Most intriguingly, specimens collected in California appear to span both major Clades 1 and 2, with some freshwater individuals recovered as members of Clade 1A. Additionally, we note that each estuary contains members of at least two subclades. However, those are not necessarily partitioned by salinity, suggesting that estuaries may contain cryptic diversity that may not be associated with salinity regime. This is in contrast to what we find for genotypes from the Great Lakes and Finger Lakes region, which all fall into Clade 1A, despite increased sampling of colonies the region as compared with Folino-Rorem et al. (2009). Previous work (Darling and Folino-Rorem 2009) found cryptic diversity of genotypes just within Lake Michigan, but signatures of that very localized population structure may be swamped out in the larger genus-level phylogenomic analyses we conduct here.

Geographic patterning in the *Cordylophora* species complex

In contrast to Folino-Rorem et al. (2009), who found little correspondence between phylogenetic affiliation and geography in the *Cordylophora* species complex, we find strong

geographical patterning in our RAD-sequence phylogeny. As compared with our 16S phylogeny (Fig. 4.3), increased resolution in our RAD-seq phylogeny recovered geographically-associated subclades within the major clades 1 and 2, where each member of the subclade is from the same estuary or even the same collecting site within an estuary (i.e. Clade 1B and most subclades within Clade 2). In their review, Bilton and Bishop (2002), find that this pattern of genetic differentiation along or between estuaries even on the same coast (i.e. San Francisco Bay vs. Columbia River in our system) to be common among marine animals but suggest that this is highly dependent on the dispersal ability of the organism in question. Invertebrates like *Cordylophora*, which have limited larval dispersal, may be more prone to localized geographic structuring than very mobile organisms, such as the estuarine riverbream fish, which was shown to have complete panmixia between estuary systems in South Africa (Oosthuizen et al. 2016).

Further work on the evolution of this species complex, specifically with collecting in the Ponto-Caspian region will help us to understand whether or not these smaller-scale differences between estuaries come from standing diversity in the native range or are due to divergence since the invasion of North America. On a larger scale, Clade 1 and 2 clearly have very different geographic compositions, with Clade 1 containing inland estuary portions and East Coast collecting sites, and Clade 2 containing solely samples from West Coast estuaries. This is in line with other invasive species (reviewed in Lee 2016) which often appear to have distinct invasions of the east and west coasts, most likely due to their position in different global shipping routes and different potential source populations. Notably, however, Clade 1A appears to contain freshwater members of multiple estuaries, regardless of geography, suggesting that the same freshwater lineage is present in multiple water systems throughout North America.

Patterns of Salinity Tolerance Evolution in *Cordylophora*

Consistent with the work of Folino-Rorem et al. (2009), we find that Clade 1A (freshwater members of most estuaries, Great Lakes and Finger Lakes) has a very low, limited salinity range. Specifically, Clade 1 has a lower average salinity than Clade 2 by about 5.0 psu (Figs. 4.3 and 4.6). Recently, experimental work on two genotypes, one from Clade 1A and one from 2B has demonstrated that they have distinct optimal salinities for growth (Folino-Rorem and Renken 2018). Specifically, the growth rate of a genotype from Clade 1A was highest at the low salinity of 0.5 psu and decreased substantially at higher salinity levels. In contrast, a genotype from Clade 2 displayed an optimal grown rate at 10 psu and showed declining growth rates when subjected to increasingly higher or lower salinities than this optimum. This suggests that at least some of the difference in salinity preference may be genetically determined and associated with phylogenetic affinity. These results are consistent with a potential scenario wherein members of Clade 1A have actually lost the ability to tolerate higher salinities and are an incipient obligate freshwater population.

We also used our well-resolved RAD-sequencing phylogeny and our salinity data for each collecting site to estimate the ancestral salinity level at the common ancestors of the major subclades of *Cordylophora* (Fig. 4.7). Our results indicate that the common ancestor to the *Cordylophora* species complex has a relatively freshwater predicted salinity, given that each major clade contains low-salinity members, in contrast to what might be expected based on a hypothesis of *Cordylophora* evolving from a marine ancestor. However, given the limitations in sampling, especially that we were unable to obtain samples from the native Ponto-Caspian region of *Cordylophora*, there is uncertainty in these reconstructions regarding the origin of lower salinity tolerances within this group.

Integrative efforts for species delimitation in *Cordylophora*

Guided by our well-resolved RAD-sequencing phylogeny of *Cordylophora*, we attempted to resolve taxonomic uncertainty in the genus by delineating species boundaries, focusing on whether or not the freshwater lineage 1A is a separate species from the rest of Clade 1. Evidence from our RAD-sequencing phylogeny suggests separation between all subclades of Clade 1, with 1A and 1C forming a monophyletic group to the exclusion of 1B (1AC+B+2 model). However, analyses of population structure and hybridization (Fig. 4.7 and Tables 4.6 and 4.7), support the grouping of 1B and 1C (1A+BC+2 model) and support the genetic contribution from both 1A and 1B to members of 1C. In a Bayesian Factor Delimitation framework (Leaché et al. 2014), a model with each subclade as its own species (1A+B+C+2) had the highest marginal likelihood, just above the same model favored by the structure analyses (Table 4.8).

Given the different predictions between different methods about the separation between subclades 1A+1B+1C, it is likely that some mechanism of gene tree incongruence exists in this system, such as incomplete lineage sorting or ongoing gene flow between the subclades in question. Such incongruence is common in RAD-seq species delimitation studies, such as in (Herrera and Shank 2016) and (Pante et al. 2014) who both encountered a similar pattern in deep sea octocorals. In *Cordylophora*, despite deep divergence between Clades 1 and 2, it is likely that the subclades of Clade 1 are still undergoing divergence post-invasion in their respective estuaries or have simply been invaded by different European genotypes. Additionally, since 1A and 1C co-occur in the James River and the Great Bay, ongoing hybridization is indeed an obvious explanation. Further work on this data set using methods that more directly estimate gene flow, such as FastSimCoal (Excoffier et al. 2013) and/or explicitly test for introgression vs.

incomplete lineage sorting such as the ABBA-BABA test (Durand et al. 2011) should shed light on this issue.

Further, increased sampling outside the North American range could provide further insight into subclade divergence. One well-known limitation of BFD analysis is that it is inherently a “validation” process, needing hypothesized models of species assignment to be established *a priori*. It is possible that our selected models of delimitation within Clade 1 may not adequately capture the true patterns of speciation in this group (see comparison of “validation” and “discovery” approaches in Satler et al. 2013). Secondly, a criticism of SNAPP and other species delimitation approaches that use a multispecies coalescent model is that they are delimiting simple population structure rather than independently evolving species, and thus may be choosing models that overestimate the number of species (Sukumaran and Knowles 2017).

Our approach to integrate multiple lines of evidence (phylogenomic, population genomic, ecological, physiological) provides a clearer picture of evolutionary patterns within the *Cordylophora* species complex. The integration of multiple lines of evidence is currently regarded as the best practice in species delimitation, especially in the face of conflicting recommendations about methodology and data type (Carstens et al. 2013; Rannala 2015) and these hybrid approaches incorporating genome-scale data are becoming very common in taxonomy (Chan et al. 2017; Bryson Jr. et al. 2018). Given the deep divergence between Clades 1 and 2, each with distinct average salinities, we propose that the species complex presented here, long regarded as a single species *C. caspia*, be delimited into two separate species as described below.

Taxonomic recommendations

Our phylogenomic analyses support highly distinct Clade 1 and Clade 2 lineages within *Cordylophora caspia* that have different average salinity levels. Further, Folino-Rorem and Renken (2018) has demonstrated that genotype representatives from 1A and 2B have significantly different optimal salinity range, and our analyses of salinity level in both our Sanger and RAD phylogenies recover a difference of about 5.0 psu between Clade1 and Clade 2 overall. Although our phylogenomic analyses recover 1B and 1C as distinct clades, which was also the highest-ranking species delimitation scenario, there is evidence of hybridization or incomplete lineage sorting between 1B and 1C and between 1A and 1C. Schuchert (2004) synonymized *C. lacustris* due to lack of morphological distinction, the reciprocal monophyly and distinct salinity profiles in Clade 1 and 2, but by contrast, our work supports the existence of at least two species. Given these lines of evidence, we resurrect *Cordylophora lacustris* to represent all of Clade 1. Despite the distinct patterns of freshwater tolerance in 1A, the potential for incomplete lineage sorting and/or ongoing hybridization between different subclades of Clade 1 suggests that designation of 1A as a distinct species from 1B and 1C is premature. Therefore, we at this time designate Clade 1 and 2 as separate species, with Clade 2 remaining as *C. caspia*. Our overall interpretation is that *Cordylophora* is a species complex comprised of lineages that are all at different parts along the speciation continuum, from having a low probability that they will become separate species, to those that already distinct species (Nosil et al. 2009), and that only Clade 1 and Clade 2 at this time rise to the level of independent species.

Conclusions

In this work, we provide a detailed phylogenomic study of the *Cordylophora* species complex, including intensive sampling along estuarine salinity gradients where we may expect the different lineages to come into contact. Our well-supported, high-resolution RAD-seq phylogeny of the *Cordylophora* species complex allows us to take an integrative approach to identifying species boundaries and clarifying patterns of salinity tolerance evolution in this group. Our integrative approach includes phylogenetic, population genomic and ecological lines of evidence. We find that differences in salinity regime may contribute to patterns of deep differentiation between lineages of *Cordylophora*, and that these lineages may also differ in their invaded range in North America. These data, combined with previous evidence that there is a physiological and potentially genetic salinity difference between clades, allows us to conclude that there are two distinct species, resurrecting *C. lacustris* as the freshwater/lower-salinity species. These results, including the genomic resources we developed for this study, are an important step for further developing *Cordylophora* into a model for understanding the evolutionary transition to freshwater from a marine ancestor.

TABLES

Table 4.1 Description of localities sampled for *Cordylophora*

<i>Code</i>	<i>Location</i>	<i>City</i>	<i>Region</i>	<i>Estuary</i>	<i>Salinity (psu)</i>	<i># RAD samples</i>
<i>SL</i>	Seneca Lake		NY, USA		0.3	
<i>WH</i>	Woods Hole		MA, USA		3	
<i>Lme</i>	Lake Michigan East		WI, USA		Fresh ⁺	2
<i>Mb</i>	Muskegon Bay		WI, USA		Fresh ⁺	6
<i>GW</i>	Grafham Waters		UK		0	4
<i>IV</i>	Irvine Water Treatment Plant	Irvine	CA, USA		1.3	7
<i>NR</i>	Napa River		CA, USA	San Francisco Bay	16	5
<i>An</i>	Antioch		CA, USA	San Francisco Bay	1.3	8
<i>PB</i>	Pittsburg		CA, USA	San Francisco Bay	2.9	
<i>IS</i>	Iselton		CA, USA	San Francisco Bay	0.1	2
<i>SC</i>	Suisun City		CA, USA	San Francisco Bay	6.7	10
<i>C</i>	Huinay		Chile		Brackish*	
<i>CR</i>	Columbia River		OR, USA	Columbia River	0	
<i>F</i>	Canet. St.	Nazaire Lake	France		Brackish*	
<i>LS</i>	LaSalle Lake	Marseilles	France		0	
<i>G</i>	Ryck River	Greifswald	Germany		5	
<i>LB</i>	Lake Balaton	Tihany	Hungary		0.45	
<i>Bh</i>	Burnham Harbor	Chicago	IL, USA		0.1	14
<i>DP</i>	DesPlaines River	Joliet	IL, USA		0.5	6
<i>FN</i>	59th Street Marina	Chicago	IL, USA		0.4	
<i>H</i>	Illinois River	Henry	IL, USA		1	
<i>I</i>	Shannon River	Limerick	Ireland		2	
<i>Chp</i>	Chippokes State Park		VA, USA	James River	1.1	3
<i>Jcc</i>	James City		VA, USA	James River	1.6	9
<i>Kml</i>	Kingsmill Landing		VA, USA	James River	4.4	5
<i>N</i>	Waal River	Nijmegen	Netherlands		0.3	
<i>Ex</i>	Squamscott	Exeter	NH, USA	Great Bay	10	11
<i>J</i>	Jackson Landing	Durham	NH, USA	Great Bay	25	
<i>Nm</i>	Lamprey River	Newmarket	NH, USA	Great Bay	1.2	10
<i>Ccs</i>	Cayuga Lake	Ithaca	NY, USA		0.2	6

<i>LO</i>	Lake Ontario	Rochester	NY, USA		0.3	
<i>Co</i>	Coos Bay		OR, USA	Columbia River	11	
<i>P39</i>	Pier 39	Astoria	OR, USA	Columbia River	0.2	9
<i>Hm</i>	Hammond		OR, USA	Columbia River	2	9
<i>Gar</i>	Garibaldi		OR, USA	Columbia River	22.3	10
<i>P</i>	Panama Canal	Gamboa	Panama		0	
<i>BR</i>	Brennan Isle		CA, USA	San Francisco Bay	1	6
<i>PR</i>	Petaluma River		CA, USA	San Francisco Bay	22	
<i>V</i>	James River	Jamestown	VA, USA		0.5	
<i>Cb</i>	Chinook Bay		WA, USA	Columbia River	1	10
<i>Ilw</i>	Ilwaco		WA, USA	Columbia River	3.5	12

+Collected from freshwater environment, exact salinity not known but presumably <2.0 ppt

*Exact salinity unknown but not freshwater

Table 4.2 Description of data sets used to estimate phylogenies of the *Cordylophora* species complex

Marker(s)	# Individ.	Seq. Length OR # of SNPs	Model Selected
28s	67	792	TIM3e+I
16s	131	531	TIM3+F+G4
Co1	64	560	TIM+F+G4
16s, CO1	57	1091	K3Pu+F+I+G4
16s, CO1, 28s	42	2883	Nuclear: TIM2e , Mitochondrial: K3Pu+F+G4
RAD Top3, 50% Coverage	63	1752	PMB+F+ASC+R4
RAD Top3, Min8	63	27063	PMB+F+ASC+R4
RAD Full, 50%	164	61	PMB+F+ASC+R4
RAD Full, Min8	164	23584	GTR+F+ASC+R3
RAD+Co1	34	17421	CO1: TIM2+F+I, RAD:TR+F+ASC
RAD+16s	46	7565	16s:TIM3+F+G4, RAD:TVM+F+ASC
RAD+28s	40	15914	28s:TIM3e+I, RAD:GTR+F+ASC

Table 4.3 Measures of phylogenetic support and diagnosability for major clades in 16S phylogeny

Putative Species	Closest Species	Intra_Dist	Inter Dist¹	Intra/Inter²	P ID(Strict)³
Clade 1	Clade 2	0.054	0.306	0.18	0.93 (0.88, 0.99)
Clade 2	Clade 1	0.078	0.306	0.25	0.91(0.86,0.96)
Clade 1A	Clade 1B	0.017	0.143	0.12	0.95 (0.90, 1.0)
Clade 1B	Clade 1A	0.002	0.143	0.02	0.98 (0.92, 1.0)
Clade 2	Clade 1	0.054	0.269	0.2	0.92 (0.87, 0.98)
Clade 2B	Clade 2A	0.036	0.152	0.24	0.78 (0.65, 0.90)
Clade 2A	Clade 2B	0.008	0.152	0.06	0.97 (0.92, 1.0)
Clade 1A	Clade 1B	0.002	0.143	0.02	0.98 (0.92, 1.0)
Clade 1B	Clade 1A	0.017	0.143	0.12	0.95 (0.90, 1.0)

1The average pairwise tree distance between the members of one putative species and the members of the closest second putative species. 2The ratio of Intra Dist to Inter Dis. 3The mean (95% CI) probability of correctly identifying an unknown member of the putative species using criterion that it must fall within clade. *Statistics for data set used in further phylogenetic analyses

Table 4.4 Descriptive statistics of selected RAD-sequencing assemblies

Depth¹	Clustering	Indel²	Min Indiv	# Var. Sites	# PIS¹	Min Loci²	Max Loci³	Max Overlap⁴
4	85	8	4	320698	162528	212	30468	121
4	85	8	8	137598	91639	201	21266	121
4	85	8	50%	393	303	1	59	121
4	80	8	4	358493	182799	103	384545	122
*4	85	10	8	139649	92848	178	25629	124
4	85	10	4	324626	164360	214	39827	124
6	85	8	4	209503	110601	4	23645	111
6	85	8	8	101579	67924	3	15247	111
6	85	8	50%	99	87	1	11	111
6	85	10	8	103052	68855	1	19504	111

¹Minimum read depth required for a locus to be retained in assembly. ²Maximum number of inferred indels for a locus to be retained in assembly. ³Number of parsimony informative sites generated for given assembly. ⁴Number of loci assembled in individual with the least number of loci. ⁵Number of loci assembled in individual with the greatest number of loci. ⁶Maximum number of samples covered by a single locus in given assembly.

Table 4.5 Measures of phylogenetic support and diagnosability for major clades recovered in RAD-sequence phylogeny

Species	Closest Species	Intra Dist	Inter Dist – Closest¹	Intra/Inter²	P ID(Strict)³
Clade 1	Clade 2	0.595	0.985	0.6	0.81 (0.76, 0.86)
Clade 2	Clade 1	0.457	0.985	0.46	0.85 (0.80, 0.91)
Clade 1A	Clade 1C	0.338	0.774	0.44	0.86 (0.81, 0.92)
Clade 1C	Clade 1A	0.09	0.357	0.25	0.83 (0.72, 0.93)
Clade 1B	Clade 1A+C	0.088	0.357	0.25	0.91 (0.86, 0.96)
Clade 2	Clade 1	0.457	0.556	0.82	0.71 (0.65, 0.76)

¹The average pairwise tree distance between the members of one putative species and the members of the closest second putative species. ²The ratio of Intra Dist to Inter Dis. ³The mean (95% CI) probability of correctly identifying an unknown member of the putative species using criterion that it must fall within clade.

Table 4.6 Per-individual results of hybrid index calculations using Clade 1A and Clade 1C as parental populations

Individual	Geography	Hybrid Index¹	Interpretation	Q-Score 1A²
Iv2	California	0.5089045	Hybrid	0.392
Iv5	California	0.4797266	Hybrid	0.517
Br12	California	0.3757671	Hybrid	0.539
Ccs1	Cayuga Lake	0.6151806	Hybrid/B-like	0.382
Ccs4	Cayuga Lake	0.0513312	A-like	0.96
Ccs5	Cayuga Lake	0.3853461	Hybrid	0.602
Ccs6	Cayuga Lake	0.457948	Hybrid	0.559
Chp3	James River	0.5091257	Hybrid	0.51
Dp10	Lake Mich.	0.003742	A-like	0.994
Dp11	Lake Mich.	0	A-like	0.999
Dp13	Lake Mich.	0.2154665	A-like	0.812
Ex3	Great Bay	0.0326126	A-like	0.98
Ex8	Great Bay	0.0339201	A-like	0.972
Iv1	California	0.4679238	Hybrid	0.439
Iv3	California	0.318912	Hybrid	0.577
Iv4	California	0.2315522	Hybrid	0.722
Jcc1	James River	0.0258834	A-like	0.964
Jcc11	James River	0.0791791	A-like	0.919
Jcc2	James River	0.4559557	Hybrid	0.538
Jcc3	James River	0.118102	A-like	0.88
Jcc4	James River	0.0061978	A-like	0.998
Jcc5	James River	0.418563	Hybrid	0.608
Jcc6	James River	0.4065774	Hybrid	0.601
Mb1	Lake Mich.	0.0071935	A-like	0.999
Mb2	Lake Mich.	0.028083	A-like	0.981

¹Hybrid indices range from 0-1, 0.0 representing only genetic contribution from Clade 1A, and 1.0 representing only contribution from Clade 1C. ²Probability from Structure analysis that individual belongs to Clade 1A (one minus probability that individual belongs to 1C).

Table 4.7 Per-individual results of hybrid index calculations using Clade 1A and Clade 1B as parental populations

Individual	Geography	Hybrid Index¹	Interpretation	Q-Score 1A²
Is2	California	0.4886842	Hybrid	0.707
Iv6	California	0.4542854	Hybrid	0.532
Br12	California	0.3512301	Hybrid	0.472
Ccs1	Cayuga Lake	0.5755387	Hybrid	0.594
Ccs4	Cayuga Lake	0.0465696	A-Like	0.051
Ccs5	Cayuga Lake	0.3078141	Hybrid	0.332
Ccs6	Cayuga Lake	0.4303164	Hybrid	0.451
Chp3	James River	0.4806837	Hybrid	0.494
Ex13	Great Bay	0.809446	B-Like	0.841
Nm3	Great Bay	0	A-Like	0.002
Ex4	Great Bay	0.8774368	B-Like	0.998
Nm8	Great Bay	0.0247336	A-Like	0.036
Iv1	California	0.4464049	Hybrid	0.492
Iv3	California	0.3023335	Hybrid	0.356
Iv4	California	0.1570509	A-Like	0.283
Jcc1	James River	0.0197143	A-Like	0.028
Jcc11	James River	0.0765752	A-Like	0.078
Jcc2	James River	0.3476668	Hybrid	0.426
Jcc3	James River	0.1155041	A-Like	0.126
Jcc5	James River	0.3943678	Hybrid	0.404
Jcc6	James River	0.3381854	Hybrid	0.47
Kml1	James River	0.979187	B-Like	0.997
Kml2	James River	0.9679671	B-Like	0.997
Kml3	James River	0.9753235	B-Like	1
Kml5	James River	0.9831674	B-Like	1
Kml9	James River	0.9654301	B-Like	0.999
Mb2	Lake Mich.	0.0143939	A-Like	0.014

¹Hybrid indices range from 0-1, 0 representing only genetic contribution from Clade 1A, and 1 representing only contribution from Clade 1B. ²Probability from Structure analysis that individual belongs to Clade 1A (one minus probability that individual belongs to 1B).

Table 4.8 Results of Bayes Factor Delimitation analysis

Model	# Species	MLE	BF	Rank
1A+1B+1C+2	4	-59175	-----	1
1A+1BC+2	3	-69788	21226	2
1AC+1B+2	3	-70495	22640	3
1ABC+2	2	-87307	56264	5
Geography	6	-86392	54434	4

Models were split or lumped according to plausible biogeographic scenarios and phylogenetic topologies. Competing models were compared and ranked using log marginal likelihood estimates (MLE) and Bayes factors (BF). A positive BF value indicates support for the most complex model (1A+1B+1C+2) over the given model.

FIGURES

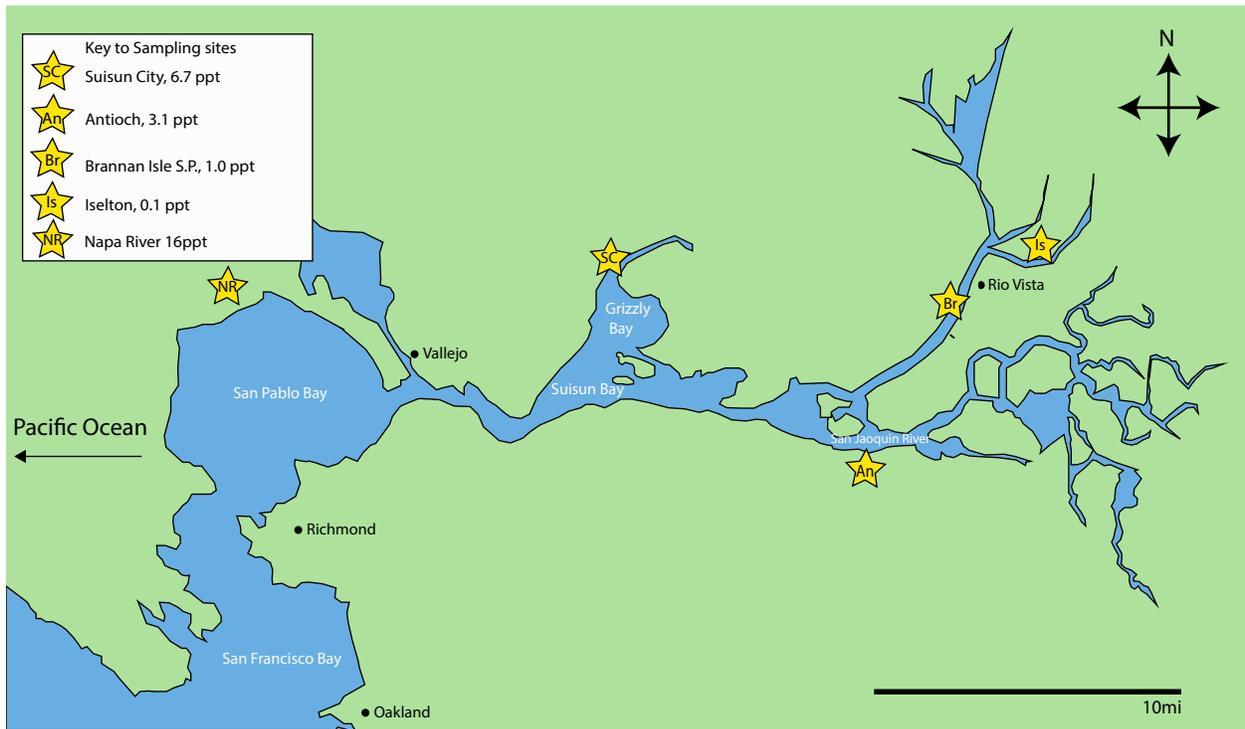


Figure 4.1 Schematic of collecting sites along the San Francisco Bay

Sites from which *Cordylophora* colonies were successfully collected for use in our data set are marked with yellow stars with abbreviations that match those in Table 4.1, with location name and salinity information presented in the key. This represents an example of our collection strategy, as similar efforts to sample systematically along the James River and Columbia River were undertaken.

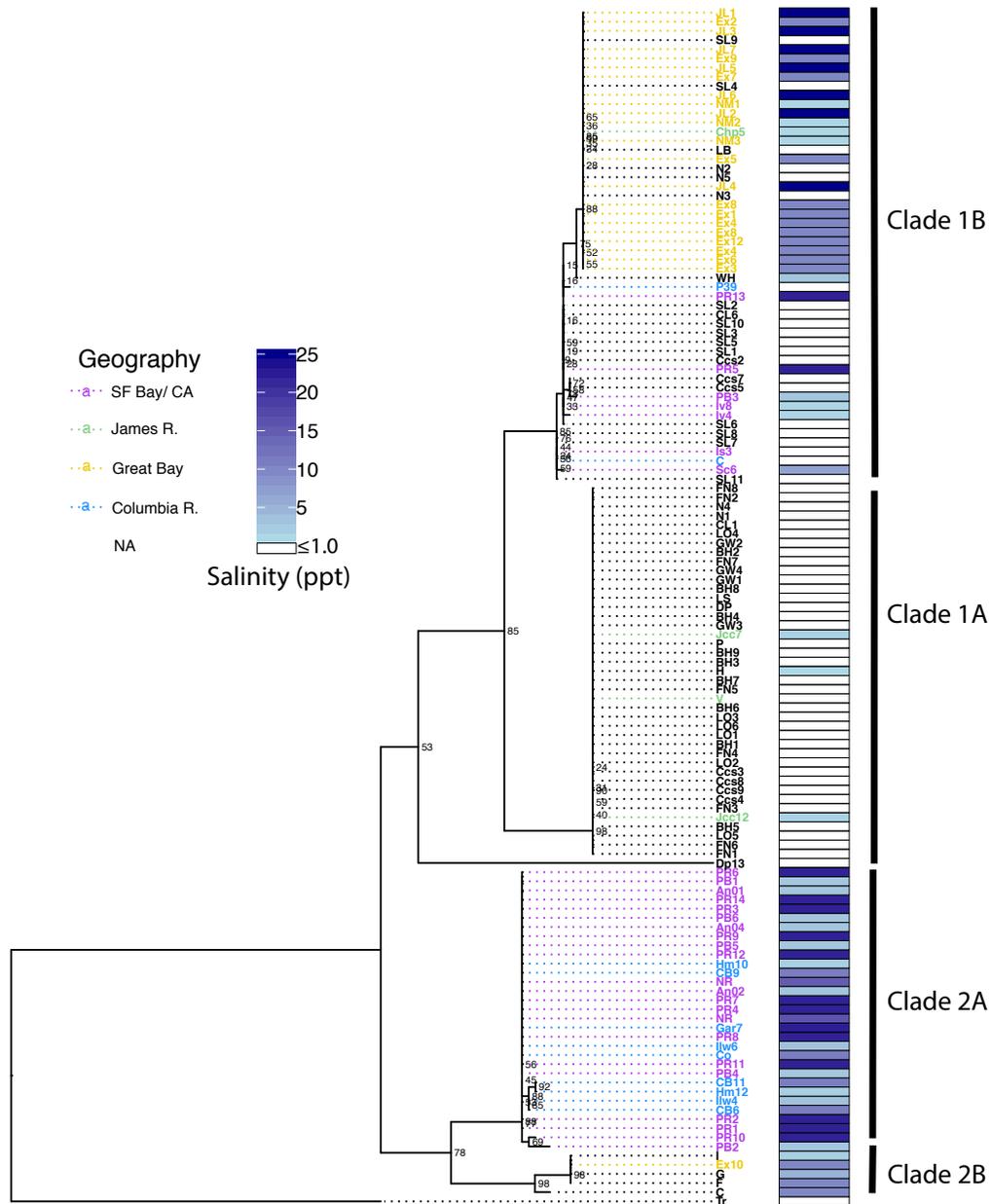


Figure 4.2 Phylogenetic tree of the *Cordylophora* species complex based on maximum-likelihood analysis of 16S sequences

Phylogenetic reconstruction was carried out in IQTree, with numbers at the nodes representing percent bootstrap support out of 10,000 ultrafast bootstrap replicates. Colored taxon names indicate newly sampled estuarine samples, and clades are named following precedent in Folino-Rorem et al. 2009. Taxa with black labels are not part of one of the estuary systems, and generally represent individuals from the Great Lakes, Finger Lakes and Europe (see Table 1).

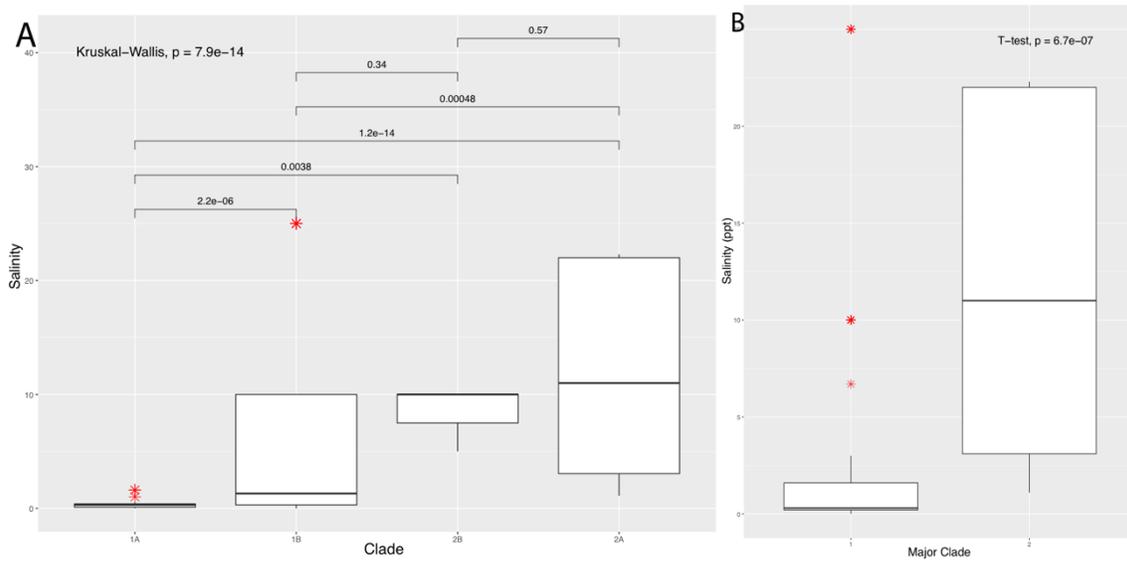


Figure 4.3 Boxplot of salinity vs. clade based on 16S phylogeny

A) The values on the horizontal bars represent p-values for Kruskal-Wallis tests on differences in salinity (in psu) between selected clades, along with overall significance level for the test for differences between groups. Most compared clades have significant differences in salinity (p-value ≤ 0.05). Red stars in each image represent outliers for a particular clade. B) A comparison of salinity levels between the two major clades recovered in Figure 2, Clade 1 (left) and Clade 2 (right).

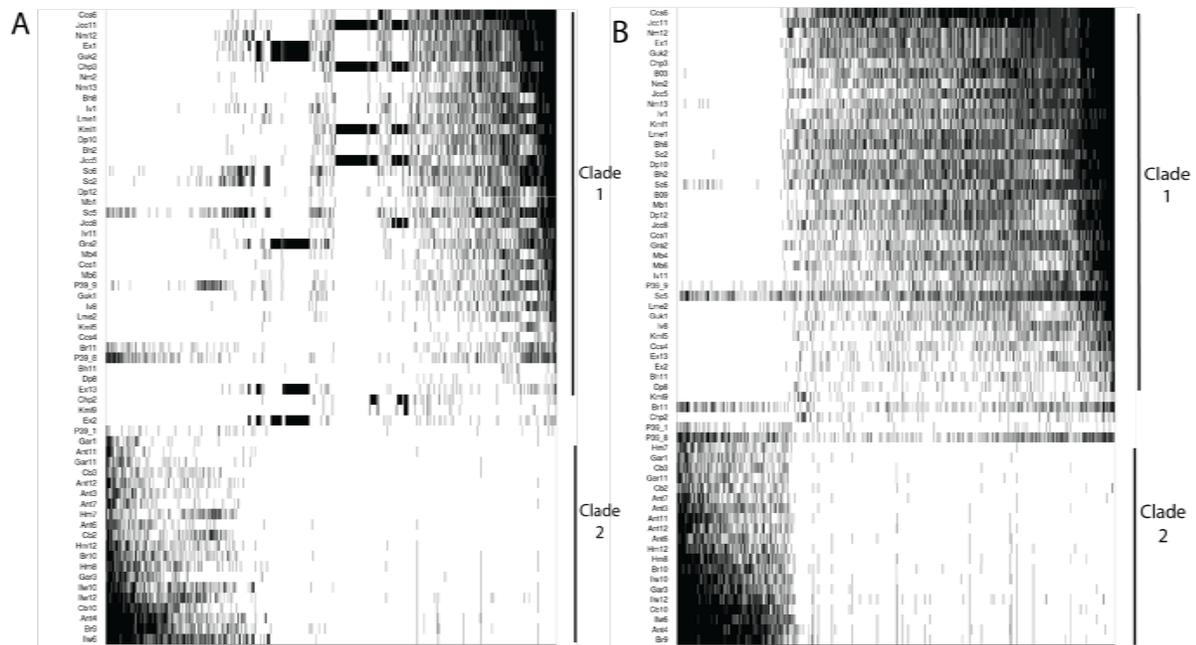


Figure 4.4 Presence-absence matrix of loci retained in default (A) and final (B) RAD-seq assemblies

Individual sample names are on vertical axis and loci are represented on the horizontal axis by a black bar if assembled for a given sample, based on the default (Depth6_Cov85_Indel8_Min8) or Depth4_Cov85_Indel10_Min8 assembly. Data has been sorted by similarity in coverage between samples, and only the three individuals per sampling location have been rendered due to computational constraints. Parts of matrix where individuals are solely from one clade or the other have been labeled.

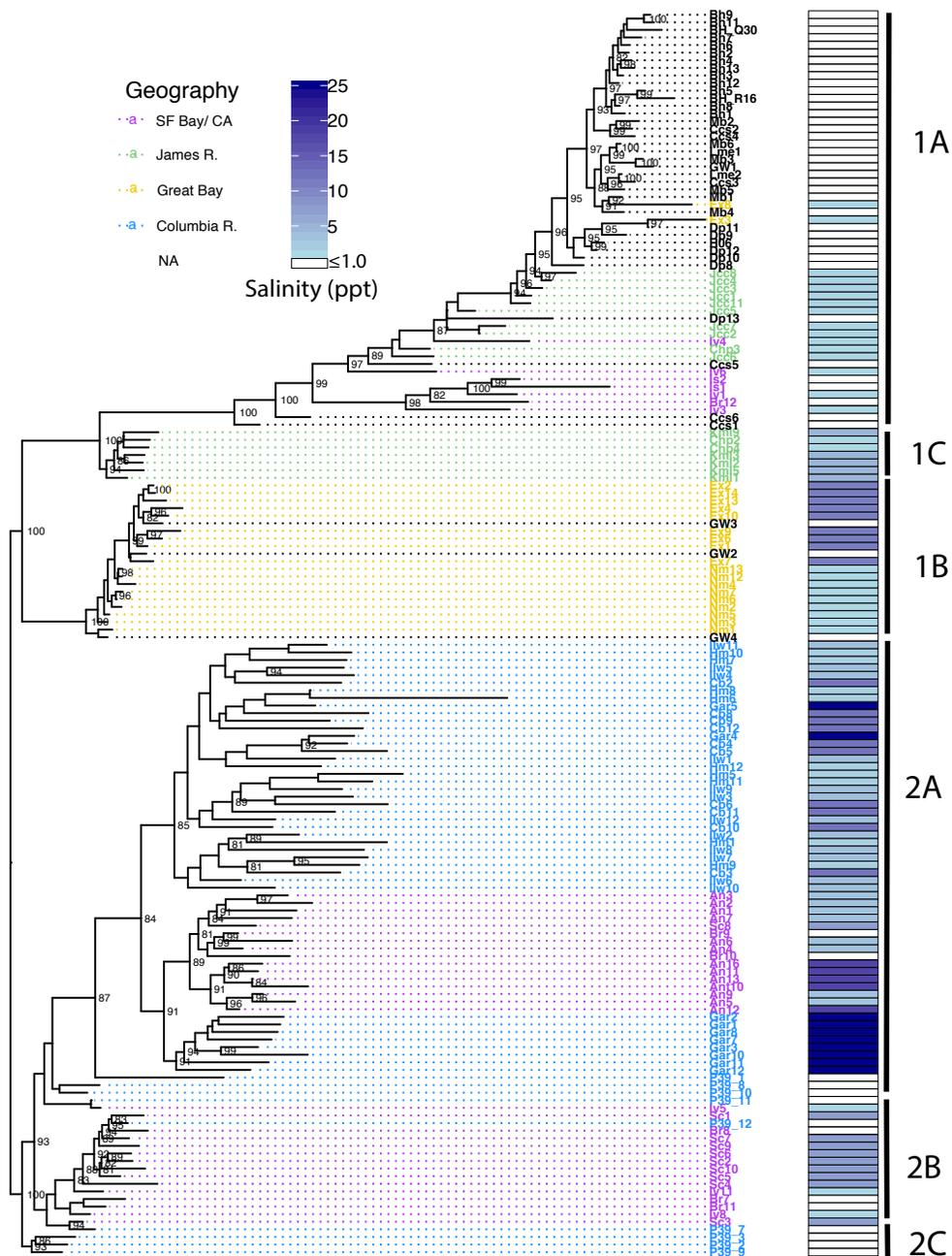


Figure 4.5 RAD-seq maximum-likelihood phylogeny of *Cordylophora* species complex based on alignment of 23852 SNPs

Estuaries and salinity levels are represented as in Figure 2. Phylogenetic reconstruction was carried out in IQTree, with numbers at the nodes representing percent bootstrap support out of 10,000 ultrafast bootstrap replicates and are included if bootstrap support is greater than 80%. All major monophyletic groups recovered have been labeled based on the scheme in Figure 2.

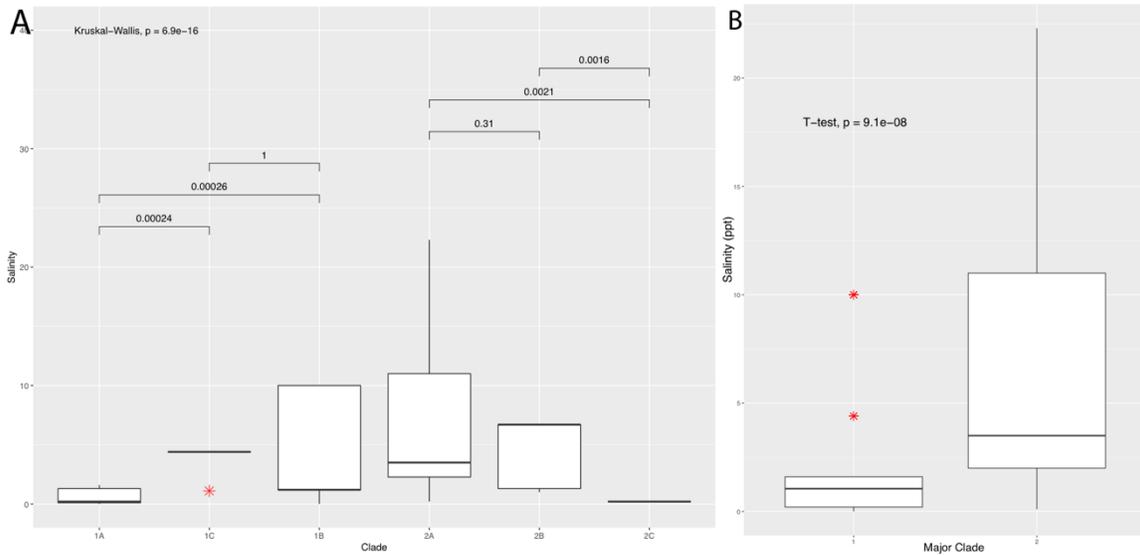


Figure 4.6 Boxplot of salinity vs. clade based on RAD-seq phylogeny

A) the values on the horizontal bars represent p-values for Kruskal-Wallis tests on differences in salinity (in psu) between selected clades, along with overall significance level for the test for differences between groups. Many compared clades have significant differences in salinity (p-value ≤ 0.05). Red stars in each image represent outliers for a particular clade. B) depicts a comparison of salinity levels between the two major clades Clade 1 and Clade 2.

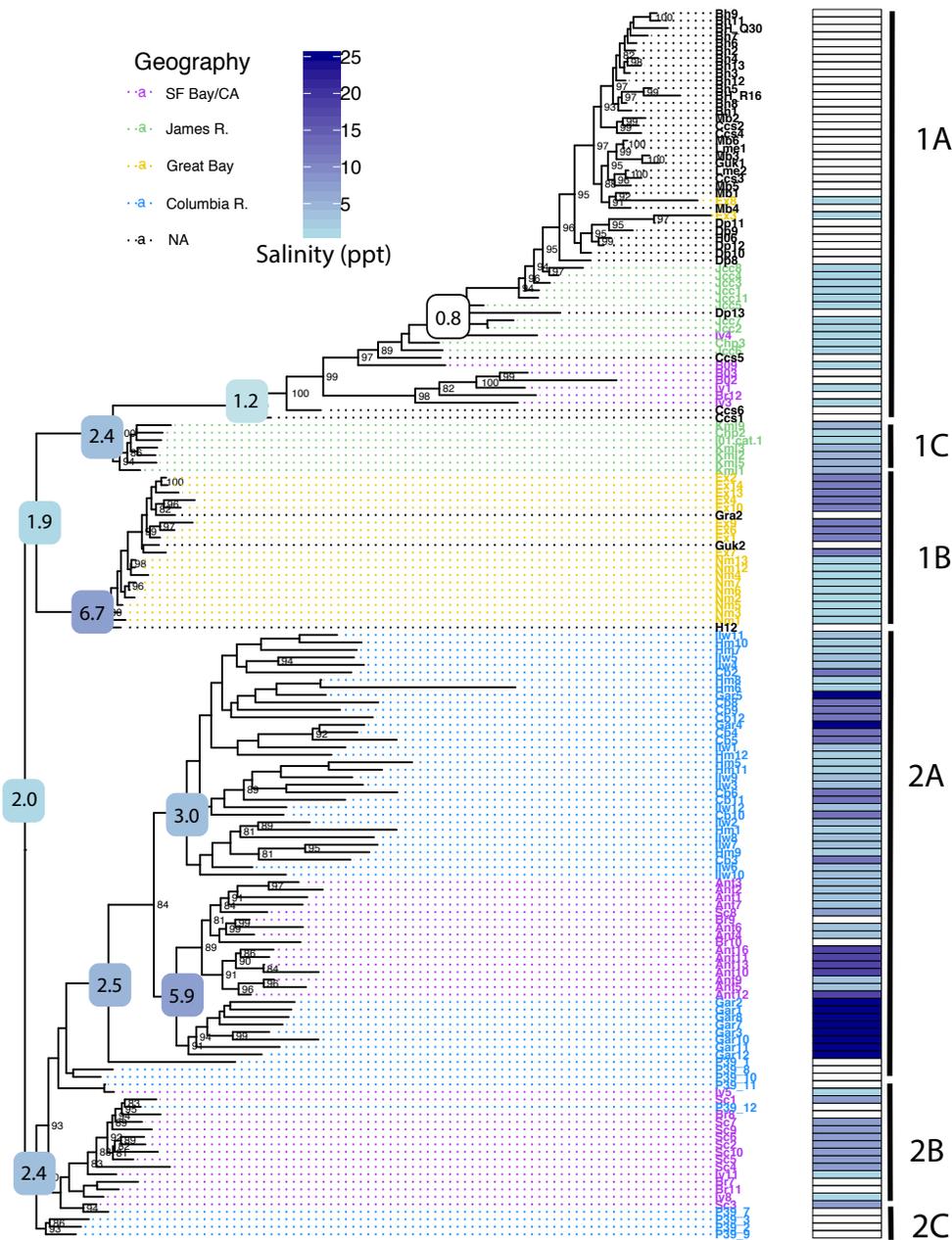
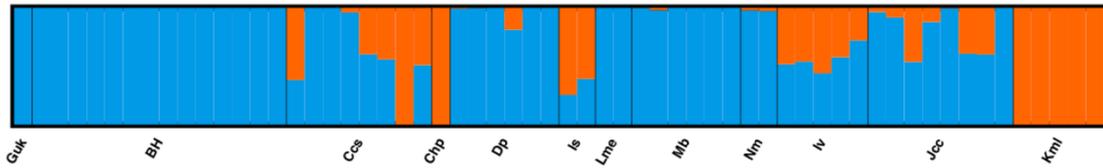


Figure 4.7 Bayesian ancestral character state reconstruction of native salinity based on RAD-sequencing phylogeny

Inferred salinity preference, averaged over 1000 MCMC replicates per node, for selected major nodes are indicated with colored squares, using the same color scale as for the actual salinity data. None of the three most ancestral nodes have high predicted ancestral salinities (<2.5 psu), suggesting that freshwater tolerance may have evolved outside of the *Cordylophora* species complex.

A.



B.

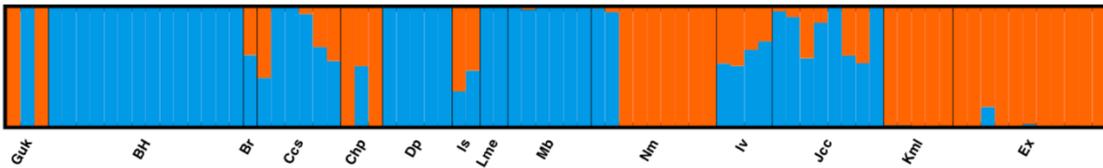


Figure 4.8 Inferred population structure for individuals making up Clade 1, excluding subclade 1B (A) and including subclade 1B (B)

Structure analysis of the RAD-sequencing data set predicted two as the most likely number of populations in each analysis. Each individual, represented by a vertical segment, is partitioned into colored segments that represent the probability that this individual is a member of each of the two inferred populations (orange and blue). See Table 1 for location abbreviations.

LITERATURE CITED

- Andolfatto, P., D. Davison, D. Erezyilmaz, T. T. Hu, J. Mast, T. Sunayama-Morita, et al. 2011. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research* 21:610-617.
- Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, et al. 2008. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLOS ONE* 3:e3376.
- Bates, A. E., C. M. McKelvie, C. J. B. Sorte, S. A. Morley, N. A. R. Jones, J. A. Mondon, et al. 2013. Geographical range, heat tolerance and invasion success in aquatic species. *Proceedings of the Royal Society B: Biological Sciences* 280:20131958.
- Bilton, D. T., J. Paula, and J. D. D. Bishop. 2002. Dispersal, Genetic Differentiation and Speciation in Estuarine Organisms. *Estuarine, Coastal and Shelf Science* 55:937-952.
- Bouckaert, R., J. Heled, D. Kühnert, T. Vaughan, C.-H. Wu, D. Xie, et al. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLOS Computational Biology* 10:e1003537.
- Bryant, D., R. Bouckaert, J. Felsenstein, N. A. Rosenberg, and A. RoyChoudhury. 2012. Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis. *Molecular biology and evolution* 29:1917-1932.
- Bryson Jr., R. W., D. A. Wood, M. R. Graham, M. E. Soleglad, and J. E. McCormack. 2018. Genome-wide SNP data and morphology support the distinction of two new species of *Kovarikia* Soleglad, Fet & Graham, 2014 endemic to California (Scorpiones, Vaejovidae). *ZooKeys*:79-106.
- Buerkle, C. A. 2005. Maximum-likelihood estimation of a hybrid index based on molecular markers. *Molecular Ecology Notes* 5:684-687.
- Cairns, S., D. Calder, A. Brinckmann-Voss, C. Castro, D. Fautin, P. Pugh, et al. 2002. Common and scientific names of aquatic invertebrates from the United States and Canada: Cnidaria and Ctenophora.
- Carstens, B. C., T. A. Pelletier, N. M. Reid, and J. D. Satler. 2013. How to fail at species delimitation. *Molecular Ecology* 22:4369-4383.
- Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko. 2013. Stacks: an analysis tool set for population genomics. *Molecular Ecology* 22:3124-3140.
- Chan, K. O., M. Alexander Alana, L. L. Grismer, Y. C. Su, L. Grismer Jesse, S. H. Quah Evan, et al. 2017. Species delimitation with gene flow: A methodological comparison and population genomics approach to elucidate cryptic species boundaries in Malaysian Torrent Frogs. *Molecular Ecology* 26:5435-5450.
- Chang, E. S., M. Neuhof, N. D. Rubinstein, A. Diamant, H. Philippe, D. Huchon, et al. 2015. Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. *Proceedings of the National Academy of Sciences* 112:14912.
- Chang, E. S., M. E. Orive, and P. Cartwright. 2018. Nonclonal coloniality: Genetically chimeric colonies through fusion of sexually produced polyps in the hydrozoan *Ectopleura larynx*. *Evolution Letters* 2:442-455.
- Chen, G. and M. P. Hare. 2008. Cryptic ecological diversification of a planktonic estuarine copepod, *Acartia tonsa*. *Molecular Ecology* 17:1451-1468.
- Chernomor, O., A. von Haeseler, and B. Q. Minh. 2016. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Systematic Biology* 65:997-1008.

- Churchill, C. K. C., A. Valdes, and D. O. Foighil. 2014. Molecular and morphological systematics of neustonic nudibranchs (Mollusca : Gastropoda : Glaucidae : Glaucus), with descriptions of three new cryptic species. *Invertebrate Systematics* 28:174-195.
- Cohen, A. N. K., T.; Li, K.; Kohn, A.; Cordell, J.; Bookheim, B.; Secord, D.; Wonham, M.; Mills, C.; Berry, H. 1998. A Rapid Assessment Survey of Nonindigenous Species in the Shallow Waters of Puget Sound. SFEI Contribution 223.
- Darling John, A. and C. Folino-Rorem Nadine. 2009. Genetic analysis across different spatial scales reveals multiple dispersal mechanisms for the invasive hydrozoan *Cordylophora* in the Great Lakes. *Molecular Ecology* 18:4827-4840.
- Durand, E. Y., N. Patterson, D. Reich, and M. Slatkin. 2011. Testing for ancient admixture between closely related populations. *Molecular biology and evolution* 28:2239-2252.
- Earl, D. A. and B. M. vonHoldt. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* 4:359-361.
- Eaton, D. A. R. 2014. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* 30:1844-1849.
- Eaton, D. A. R., E. L. Spriggs, B. Park, and M. J. Donoghue. 2017. Misconceptions on Missing Data in RAD-seq Phylogenetics with a Deep-scale Example from Flowering Plants. *Systematic Biology* 66:399-412.
- Escudero, M., D. A. R. Eaton, M. Hahn, and A. L. Hipp. 2014. Genotyping-by-sequencing as a tool to infer phylogeny and ancestral hybridization: A case study in *Carex* (Cyperaceae). *Molecular Phylogenetics and Evolution* 79:359-367.
- Evanno, G., S. Regnaut, and J. Goudet. 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology* 14:2611-2620.
- Excoffier, L., I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll. 2013. Robust Demographic Inference from Genomic and SNP Data. *PLOS Genetics* 9:e1003905.
- Folino-Rorem, N. C. 2000. The freshwater expansion and classification of the colonial hydroid *Cordylophora*. Pp. 139-144. *Marine bioinvasions*, Cambridge, MA
- Folino-Rorem, N. C. 2015. Phylum Cnidaria. Pp. 159-179 in J. R. Thorp, DC., ed. *Ecology and General Biology: Thorp and Covich's Freshwater Invertebrates*. Academic Press, Amsterdam.
- Folino-Rorem, N. C., J. A. Darling, and C. A. D'Ausilio. 2009. Genetic analysis reveals multiple cryptic invasive species of the hydrozoan genus *Cordylophora*. *Biological Invasions* 11:1869-1882.
- Folino-Rorem, N. C. and J. Indelicato. 2005. Controlling biofouling caused by the colonial hydroid *Cordylophora caspia*. *Water Research* 39:2731-2737.
- Folino-Rorem, N. C. and C. J. Renken. 2018. Effects of salinity on the growth and morphology of the invasive, euryhaline hydroid *Cordylophora* (Phylum Cnidaria, Class Hydrozoa). *Invertebrate Biology* 137:78-90.
- Fuller, R. C., K. E. McGhee, and M. Schrader. 2007. Speciation in killifish and the role of salt tolerance. *Journal of Evolutionary Biology* 20:1962-1975.
- Gompert, Z. and C. A. Buerkle. 2010. introgress: a software package for mapping components of isolation in hybrids. *Molecular Ecology Resources* 10:378-384.
- Herrera, S. and T. M. Shank. 2016. RAD sequencing enables unprecedented phylogenetic resolution and objective species delimitation in recalcitrant divergent taxa. *Molecular Phylogenetics and Evolution* 100:70-79.

- Hoang, D. T., O. Chernomor, A. von Haeseler, B. Q. Minh, and L. S. Vinh. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular biology and evolution* 35:518-522.
- Hubisz, M. J., D. Falush, M. Stephens, and J. K. Pritchard. 2009. Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources* 9:1322-1332.
- Itô, T. 1951. A new athecate hydroid, *Cordylophora japonica* n. sp., from Japan. . *Memoirs of the Ehime University, Section II* 1:81-86.
- Jankowski, T., A. G. Collins, and R. Campbell. 2008. Global diversity of inland water cnidarians. *Hydrobiologia* 595:35-40.
- Kalyaanamoorthy, S., B. Q. Minh, T. K. F. Wong, A. von Haeseler, and L. S. Jermin. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* 14:587.
- Kopelman, N. M., J. Mayzel, M. Jakobsson, N. A. Rosenberg, and I. Mayrose. 2015. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources* 15:1179-1191.
- Korshunova, T., K. Lundin, K. Malmberg, B. Picton, and A. Martynov. 2018. First true brackish-water nudibranch mollusc provides new insights for phylogeny and biogeography and reveals paedomorphosis-driven evolution. *PLOS ONE* 13:e0192177.
- Leaché, A. D., M. K. Fujita, V. N. Minin, and R. R. Bouckaert. 2014. Species Delimitation using Genome-Wide SNP Data. *Systematic Biology* 63:534-542.
- Lee, C. E. 2015. Evolutionary mechanisms of habitat invasions, using the copepod *Eurytemora affinis* as a model system. *Evolutionary Applications* 9:248-270.
- Lee, C. E. and M. A. Bell. 1999. Causes and consequences of recent freshwater invasions by saltwater animals. *Trends in ecology & evolution* 14:284-288.
- Lenz, M., B. A. da Gama, N. V. Gerner, J. Gobin, F. Groner, A. Harry, et al. 2011. Non-native marine invertebrates are more tolerant towards environmental stress than taxonomically related native species: results from a globally replicated study. *Environmental research* 111:943-952.
- Lindsay, D. J., Grossmann, Mary M., Nishikawa, Jun, Bentlage, Bastian, and Collins, Allen Gilbert. 2015. DNA Barcoding of Pelagic Cnidarians: Current Status and Future Prospects. *Bulletin- Plankton Society of Japan* 62:39-43.
- López-López, A., P. Hudson, and J. Galián. 2012. The blackburni/murchisona species complex in Australian *Pseudotetracha* (Coleoptera: Carabidae: Cicindelinae: Megacephalini): evaluating molecular and karyological evidence. *Journal of Zoological Systematics and Evolutionary Research* 50:177-183.
- Masters, B. C., V. Fan, and H. A. Ross. 2010. Species delimitation – a geneious plugin for the exploration of species boundaries. *Molecular Ecology Resources* 11:154-157.
- Meek, M. H., A. P. Wintzer, W. C. Wetzel, and B. May. 2012. Climate Change Likely to Facilitate the Invasion of the Non-Native Hydroid, *Cordylophora caspia*, in the San Francisco Estuary. *PLOS ONE* 7:e46373.
- Miglietta, M. P., D. Odegard, B. Faure, and A. Faucci. 2015. Barcoding Techniques Help Tracking the Evolutionary History of the Introduced Species *Pennaria disticha* (Hydrozoa, Cnidaria). *PLOS ONE* 10:e0144762.

- Miller, M. F., Labandeira, Conrad C. 2002. Slow crawl across the salinity divide: delayed colonization of freshwater ecosystems by invertebrates. *Geological Society of America Today* 12:4-10.
- Miller, M. R., J. P. Dunham, A. Amores, W. A. Cresko, and E. A. Johnson. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research* 17:240-248.
- Mills, E. L., J. H. Leach, J. T. Carlton, and C. L. Secor. 1993. Exotic Species in the Great Lakes: A History of Biotic Crises and Anthropogenic Introductions. *Journal of Great Lakes Research* 19:1-54.
- Molnar, J. L., R. L. Gamboa, C. Revenga, and M. D. Spalding. 2008. Assessing the global threat of invasive species to marine biodiversity. *Frontiers in Ecology and the Environment* 6:485-492.
- Monnahan, P. J., J. Colicchio, and J. K. Kelly. 2015. A genomic selection component analysis characterizes migration-selection balance. *Evolution* 69:1713-1727.
- Nguyen, L.-T., H. A. Schmidt, A. von Haeseler, and B. Q. Minh. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular biology and evolution* 32:268-274.
- Nosil, P., L. J. Harmon, and O. Seehausen. 2009. Ecological explanations for (incomplete) speciation. *Trends in ecology & evolution* 24:145-156.
- Oosthuizen, C. J., P. D. Cowley, S. R. Kyle, and P. Bloomer. 2016. High genetic connectivity among estuarine populations of the riverbream *Acanthopagrus vagus* along the southern African coast. *Estuarine, Coastal and Shelf Science* 183:82-94.
- Pante, E., J. Abdelkrim, A. Viricel, D. Gey, S. C. France, M. C. Boisselier, et al. 2014. Use of RAD sequencing for delimiting species. *Heredity* 114:450.
- Pienimäki, M. and E. Leppäkoski. 2004. Invasion Pressure on the Finnish Lake District: Invasion Corridors and Barriers. *Biological Invasions* 6:331-346.
- Rajagopal, S., G. Van der Velde, M. Van der Gaag, and H. A. Jenner. 2002. Laboratory Evaluation of the Toxicity of Chlorine to the Fouling Hydroid *Cordylophora caspia*. *Biofouling* 18:57-64.
- Rannala, B. 2015. The art and science of species delimitation. *Curr. Zool.* 61:846-853.
- Redier, L. 1967. Un nouvel hydraire, *Cordylophora solangiae* n. sp. (Atoll de Fangataufa-Tuamoty). *Cah. pacif.* 11:117-128.
- Reid, D. F. and M.I. Orlova. 2002. Evolution of Physiological Tolerance and Performance during Freshwater Invasions in the Baltic Sea and North American Great Lakes. *Canadian Journal of Fisheries and Aquatic Sciences* 59:1144-1158.
- Revell, L. J. 2011. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 3:217-223.
- Roos, P. J. 1979. Two-stage life cycle of a *Cordylophora* population in the Netherlands. *Hydrobiologia* 62:231-239.
- Ross, H. A., S. Murugan, and W. L. Sibon Li. 2008. Testing the Reliability of Genetic Methods of Species Identification via Simulation. *Systematic Biology* 57:216-230.
- Rubin, B. E. R., R. H. Ree, and C. S. Moreau. 2012. Inferring Phylogenies from RAD Sequence Data. *Plos One* 7:12.
- Ruiz, G. M., P. W. Fofonoff, J. T. Carlton, M. J. Wonham, and A. H. Hines. 2000. Invasion of Coastal Marine Communities in North America: Apparent Patterns, Processes, and Biases. *Annual Review of Ecology and Systematics* 31:481-531.

- Satler, J. D., B. C. Carstens, and M. Hedin. 2013. Multilocus Species Delimitation in a Complex of Morphologically Conserved Trapdoor Spiders (Mygalomorphae, Antrodiaetidae, Aliatypus). *Systematic Biology* 62:805-823.
- Schuchert, P. 2004. Revision of the European athecate hydroids and their medusae (Hydrozoa, Cnidaria): Families Oceanidae and Pachycordylidae. *Revue Suisse De Zoologie* 111:315-369.
- Schuchert, P. 2018. World Hydrozoa Database. *Cordylophora caspia* (Pallas, 1771), World Register of Marine Species
- Smith, D. 2001. Pennak's freshwater invertebrates of the United States: Porifera to Crustacea.
- Streftaris N, Z. A., Papathanassiou E. 2005. Globalisation in marine ecosystems: the story of non-indigenous marine species across European seas. *Oceanography and Marine Biology Annual Review* 43:419-453.
- Sukumaran, J. and L. L. Knowles. 2017. Multispecies coalescent delimits structure, not species. *Proceedings of the National Academy of Sciences of the United States of America* 114:1607-1612.
- Team, R. C. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- Trifinopoulos, J., L.-T. Nguyen, A. von Haeseler, and B. Q. Minh. 2016. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Research* 44:W232-W235.
- Tripp, E. A., Y. H. E. Tsai, Y. Zhuang, and K. G. Dexter. 2017. RADseq dataset with 90% missing data fully resolves recent radiation of *Petalidium* (Acanthaceae) in the ultra-arid deserts of Namibia. *Ecology and Evolution* 7:7920-7936.
- Tsai, J.-R. and H.-C. Lin. 2007. V-type H⁺-ATPase and Na⁺,K⁺-ATPase in the gills of 13 euryhaline crabs during salinity acclimation. *Journal of Experimental Biology* 210:620.
- Ueda, H., A. Yamaguchi, S.-i. Saitoh, S. O. Sakaguchi, and K. Tachihara. 2011. Speciation of two salinity-associated size forms of *Oithona dissimilis* (Copepoda: Cyclopoida) in estuaries. *Journal of Natural History* 45:2069-2079.
- Wagner, C. E., I. Keller, S. Wittwer, O. M. Selz, S. Mwaiko, L. Greuter, et al. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Mol Ecol* 22:787-798.
- Wessinger, C. A., C. C. Freeman, M. E. Mort, M. D. Rausher, and L. C. Hileman. 2016. Multiplexed shotgun genotyping resolves species relationships within the North American genus *Penstemon*. *American Journal of Botany* 103:912-922.
- Wickham, H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.
- Williams, S. L. and E. D. Grosholz. 2008. The Invasive Species Challenge in Estuarine and Coastal Environments: Marrying Management and Science. *Estuaries and Coasts* 31:3-20.
- Wonham, M. J. and J. T. Carlton. 2005. Trends in marine biological invasions at local and regional scales: the Northeast Pacific Ocean as a model system. *Biological Invasions* 7:369-392.
- Yu, G., D. K. Smith, H. Zhu, Y. Guan, and T. T.-Y. Lam. 2016. *ggtree*: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* 8:28-36.

Zheng, L. M., J. R. He, Y. S. Lin, W. Q. Cao, and W. J. Zhang. 2014. 16S rRNA is a better choice than COI for DNA barcoding hydrozoans in the coastal waters of China. *Acta Oceanol. Sin.* 33:55-76.

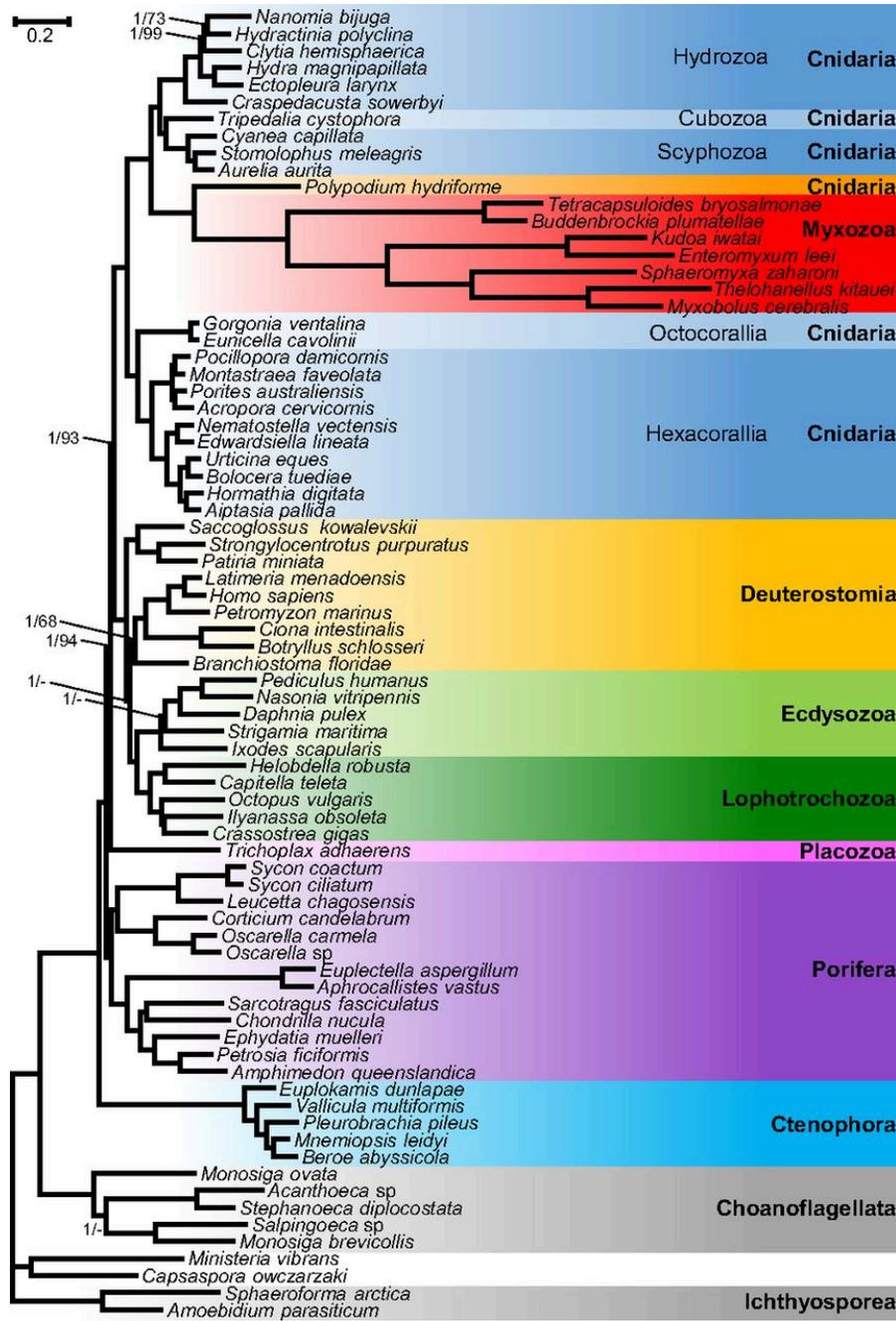
Chapter 5 Appendix

Appendix 1: Sample information and accession numbers

Species	DNA / RNA	Platform	reads	BioProject	BioSample	SRA Experiment	Genome/transcriptome shotgun assembly	Locality
<i>Kudoa iwatai</i>	DNA	G2x	76bp PE	PRJNA261422	SAMN0306868 1	SRX704259	JRUY00000000	Red Sea, Eilat
<i>Kudoa iwatai</i>	DNA	HiSeq 2000	100bp PE	PRJNA261052	SAMN0306868 1	SRX702459	JRUX00000000	Red Sea, Eilat
<i>Kudoa iwatai</i>	RNA	HiSeq 2000	100bp PE	PRJNA248713	SAMN0280092 5	SRX554567	GBGI00000000	Red Sea, Eilat
<i>Enteromyxum leei</i>	DNA	HiSeq 2000	100bp PE	PRJNA284325	SAMN0370140 5	SRX103492 8	LDNA00000000 0	Red Sea, Eilat
<i>Sphaeromyxa zaharoni</i>	DNA	HiSeq 2000	100bp PE	PRJNA284326	SAMN0370140 0	SRX103491 4	LDMZ00000000 0	Red Sea, Eilat
<i>Myxobolus cerebralis</i>	RNA	HiSeq 2000	100bp PE	PRJNA258474	SAMN0299809 6	SRX685586	GBKL00000000	California
<i>Polypodium hydriforme</i>	RNA	HiSeq 2000	100bp PE	PRJNA251648	SAMN0283786 0	SRX570527	GBGH00000000 0	Grand Lake State Park, OK
<i>Polypodium hydriforme</i>	DNA	HiSeq 2000	100bp PE	PRJNA259515	SAMN0299811 2	SRX687102		Grand Lake State Park, OK

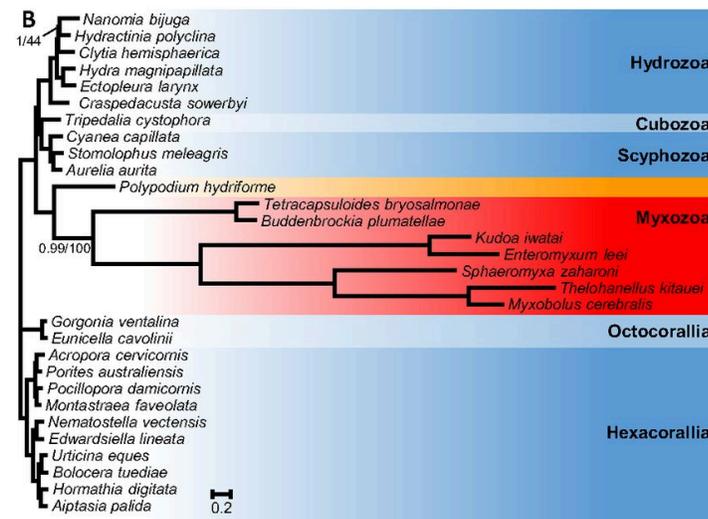
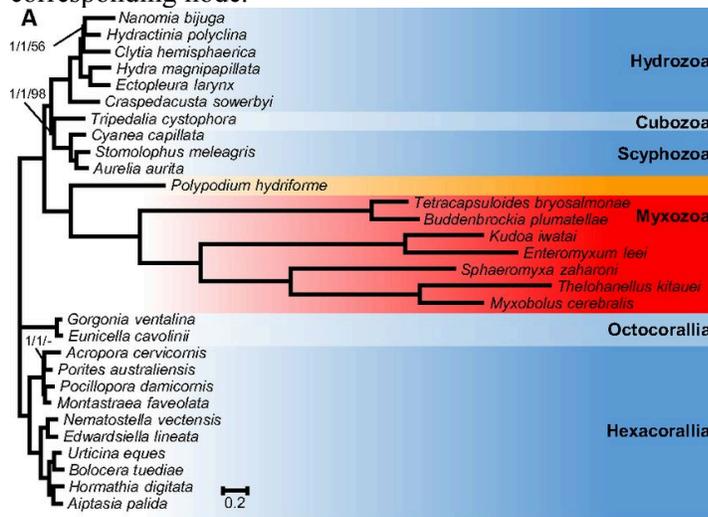
Appendix 2: Phylogenetic tree generated from a matrix of 41,237 amino acid positions, which excludes ribosomal genes, and 77 taxa using Bayesian inference under the CAT model.

Support values are indicated only for nodes that did not receive maximal support. Bayesian posterior probabilities/ML bootstrap supports under the PROTGAMMAGTR are given near the corresponding node. A minus sign (“-”) indicates that the corresponding node is absent from the ML bootstrap consensus tree.

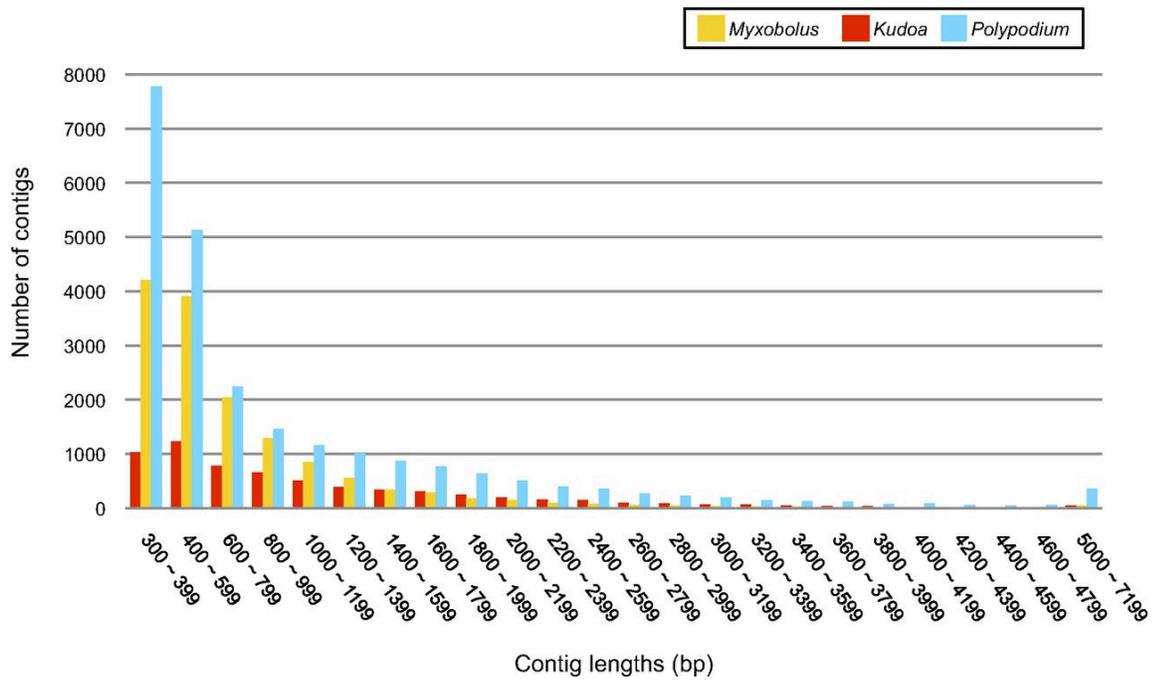


Appendix 3: Phylogenetic reconstructions with only cnidarian taxa

A) Phylogenetic tree generated from a matrix of 51,940 amino acid sequences and 30 cnidarian taxa using Bayesian inference under the CAT model. Support values are indicated only for nodes that did not receive maximal support. Bayesian posterior probabilities under the CAT model/Bayesian posterior probabilities under the CAT-GTR model/ML bootstrap supports under the PROTGAMMAGTR are given near the corresponding node. A minus sign (“-”) indicates that the corresponding node is absent from the ML bootstrap consensus tree. (B) Phylogenetic tree generated from a matrix that excludes ribosomal genes, comprising 41,237 amino acid sequences and 30 cnidarian taxa using Bayesian inference under the CAT model. Support values are indicated only for nodes that did not receive maximal support. Bayesian posterior probabilities/ML bootstrap supports under the PROTGAMMAGTR are given near the corresponding node.

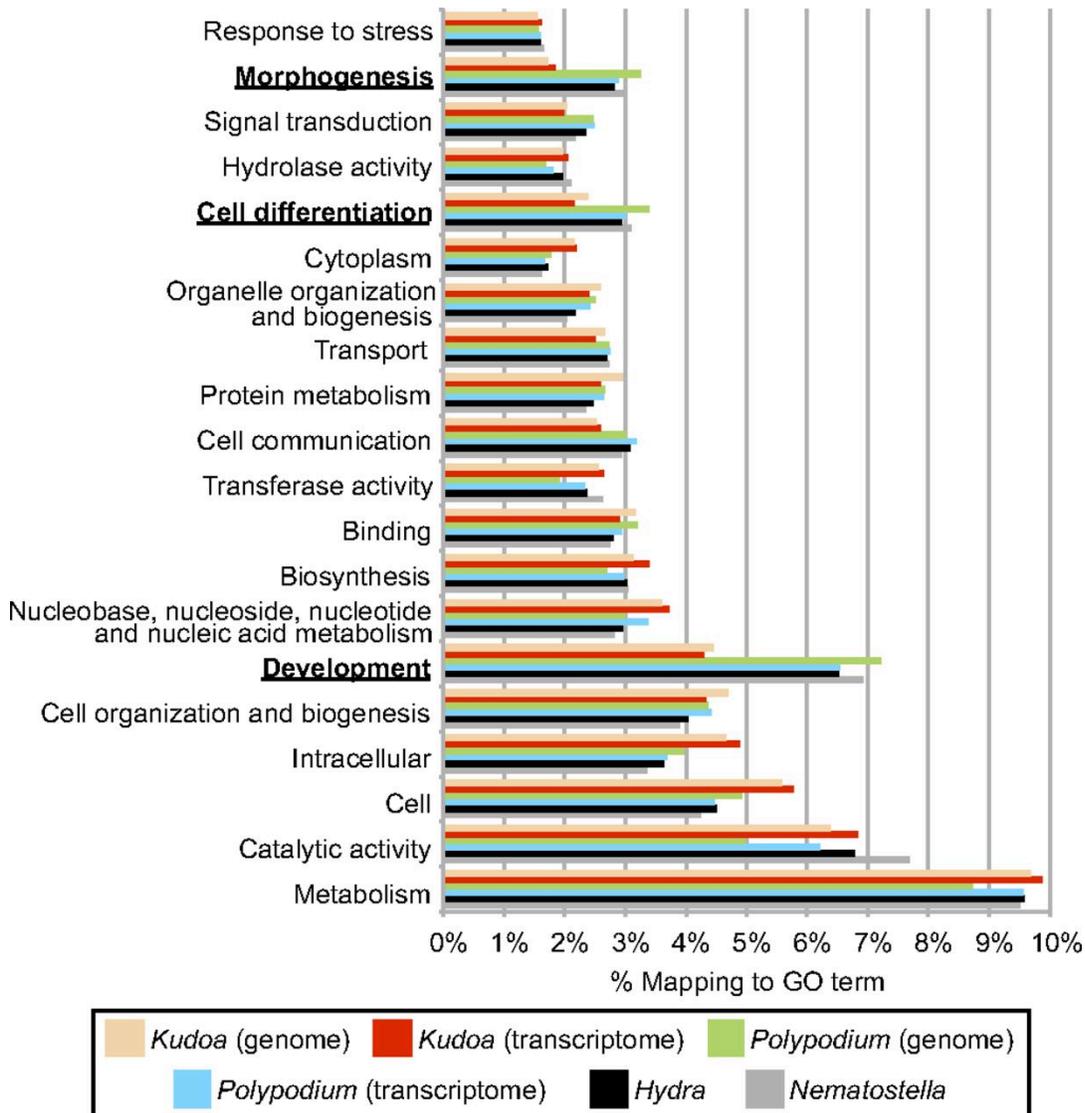


Appendix 4: Sequence size distribution of the assembled transcriptome sequences



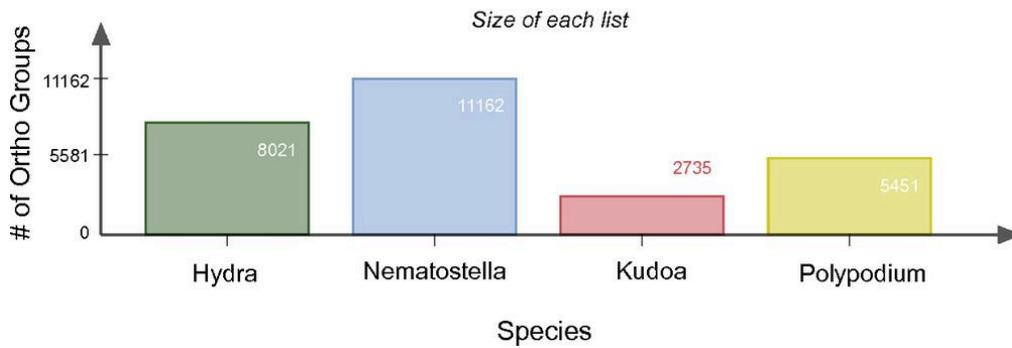
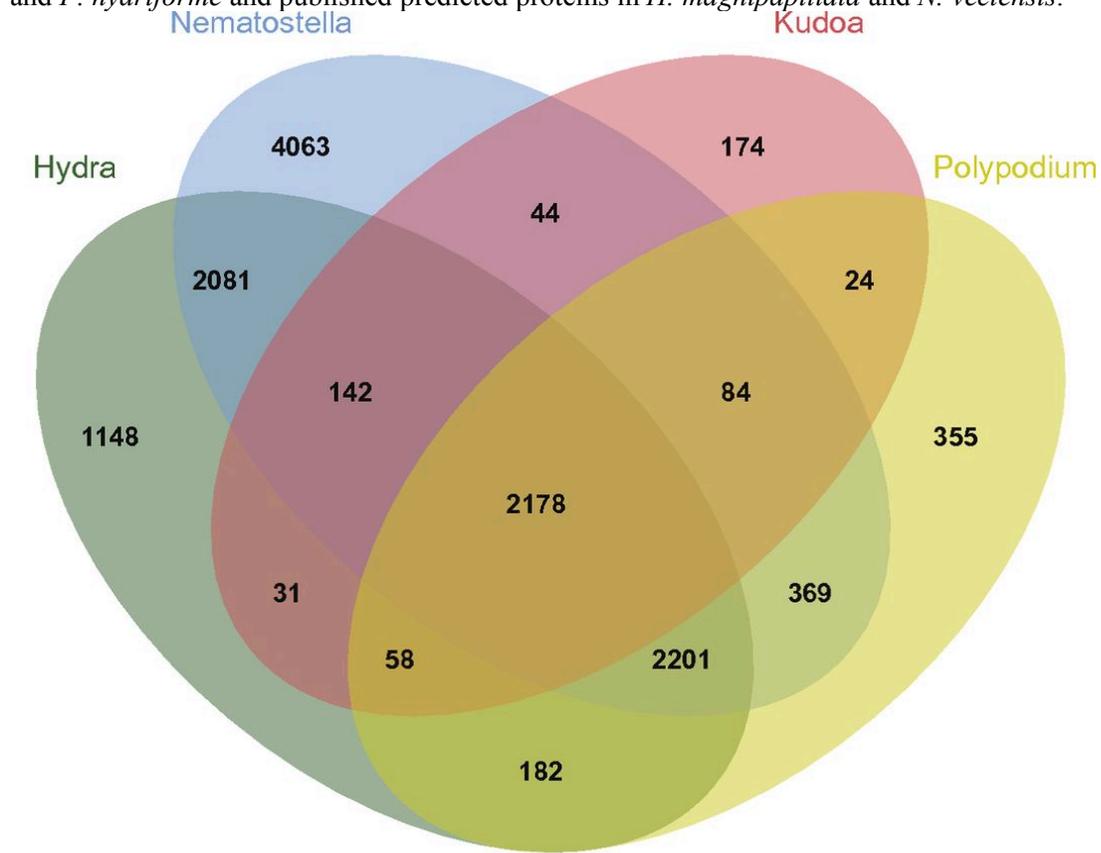
Appendix 5: GO annotation of unigenes in genomes and transcriptomes

The top 20 GO categories are shown as a percentage of total GO terms from the assemblies of *K. iwatai*, *P. hydriforme*, and the published protein sequences of *H. magnipapillata* and *N. vectensis*. Categories for which *K. iwatai* presents significantly fewer GO terms than other cnidarians are indicated in boldface type.



Appendix 6: Comparison of OGs in myxozoan and other cnidarian transcriptomes

VENN diagram comparing OGs for the OrthoMCL database from transcriptome assemblies of *K. iwatai* and *P. hydriforme* and published predicted proteins in *H. magnipapillata* and *N. vectensis*.



Appendix 7: Collection sites of *Ectopleura larynx* colonies from the coast of Maine and Ireland

Abbrev.	Site Name	Town	Region	Lat.	Long.
ME1	Darling Marine Center	Walpole	Maine, USA	43.94277778	-69.56888889
ME2	Downeast Institute	Beal's Island	Maine, USA	44.47888889	-67.59666667
IRE1	University Marine Lab	Pontaferry	Northern Ireland	54.380705	-5.550311
IRE2	Sketrick Pontoon	Sketrick Island	Northern Ireland	54.48916	-5.646843

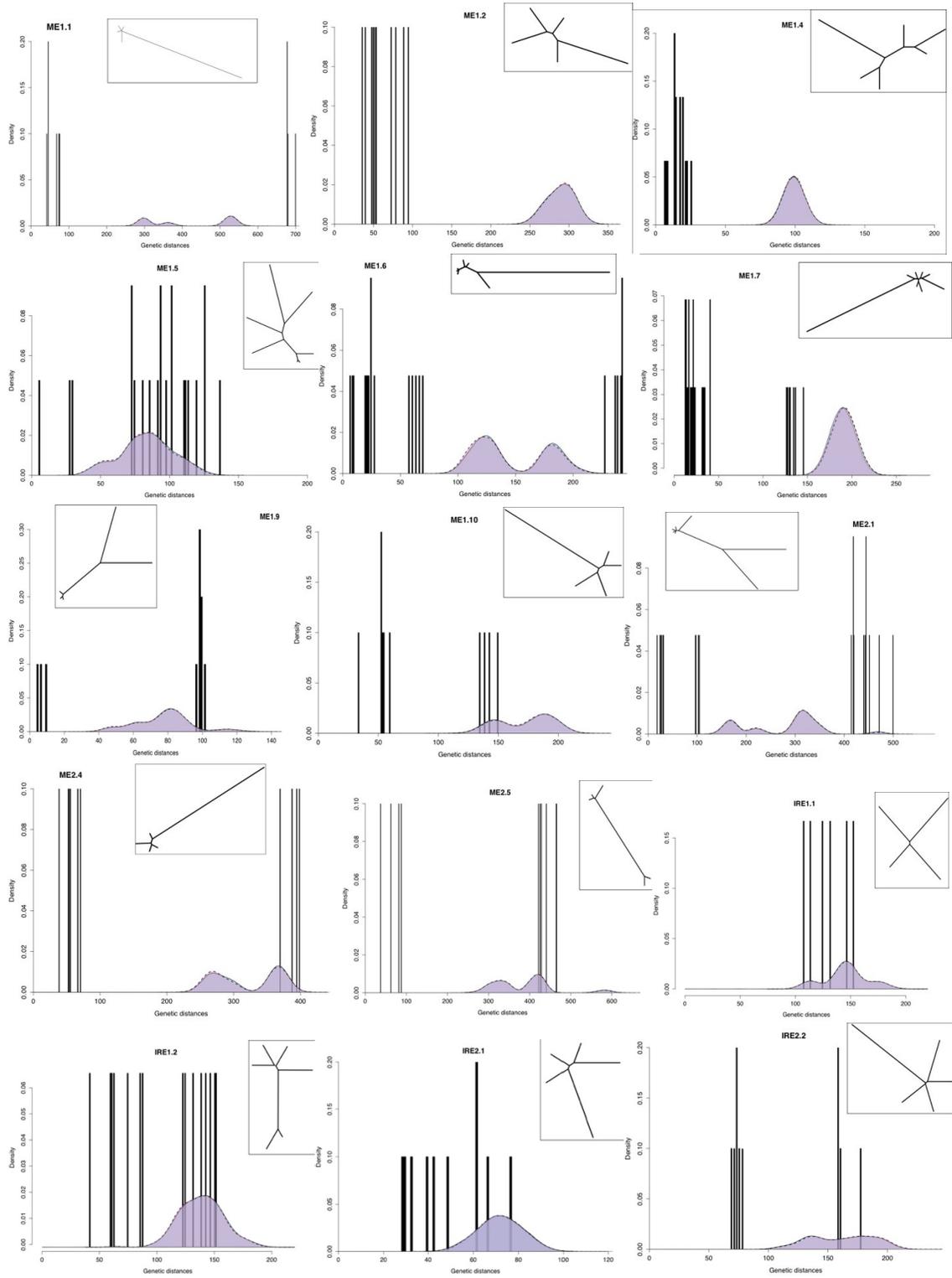
Appendix 8: . Summary of loci for each colony-level data set

Collecting Site	Colony	Sex	Polyps	Loci	RD [†]
Maine 1	ME1.1	M	5	772	276.3
	ME1.2	M	5	439	267.0
	ME1.3	M	5	430	279.5
	ME1.4	M	6	143	240.5
	ME1.5	F	7	287	480.5
	ME1.6	F	7	183	292.4
	ME1.7	F	8	336	291.3
	ME1.8	F	6	167	291.4
	ME1.9	F	5	152	207.7
	ME1.10	M	5	300	89.3
N. Ireland 1	IRE1.1	F	4	270	88.5
	IRE1.2	F	6	244	112.7
Maine 2	ME2.1	M	7	665	305.5
	ME2.2	M	6	607	214.3
	ME2.3	F	8	435	277.0
	ME2.4	F	5	553	190.7
	ME2.5	H	5	639	234.9
N. Ireland 1	IRE2.2	F	5	132	266.3
	IRE2.3	M	5	326	94.4

Appendix 9: Description of data sets used in each analysis in Chapter 3

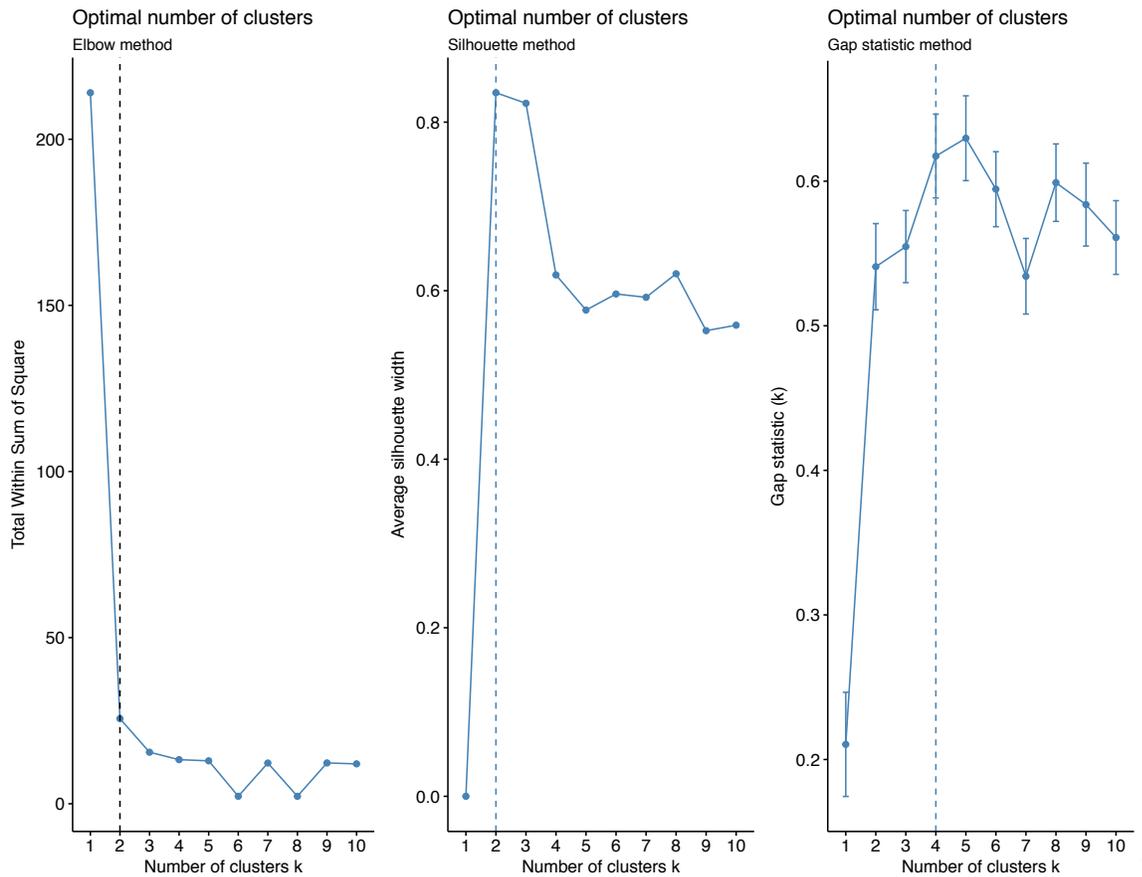
Analysis	Stacks Populations module filtering for loci	Number of Individuals	Number of Loci
Colony-level analyses (Rclone, per-locus study)	Prepared per colony, all polyps in a given colony, m=8	Varies by colony, see Table 1	Varies by colony, see Table 1
Collecting Site Diversity Statistics (Table 2 and Figure 4)	Prepared per collecting site., r=1.0, m=6	See Table S2	See Table 2
Collecting-site level pairwise F_{ST}	Prepared per collecting site., r=1.0, m=6	See Table S2	Varies by pairwise comparison
PLINK and KING analyses of genetic relatedness	Prepared per collecting site, r=.75, m=4, p=.8, --write_single_snp	See Table S2	ME1=954, ME2=1025, IRE1=328, IRE2=364

Appendix 10: Genetic relationships within selected colonies of *E. larynx* not already included in Figure 3.2



Appendix 11: Results of different methods for choosing a best k (number of clusters)

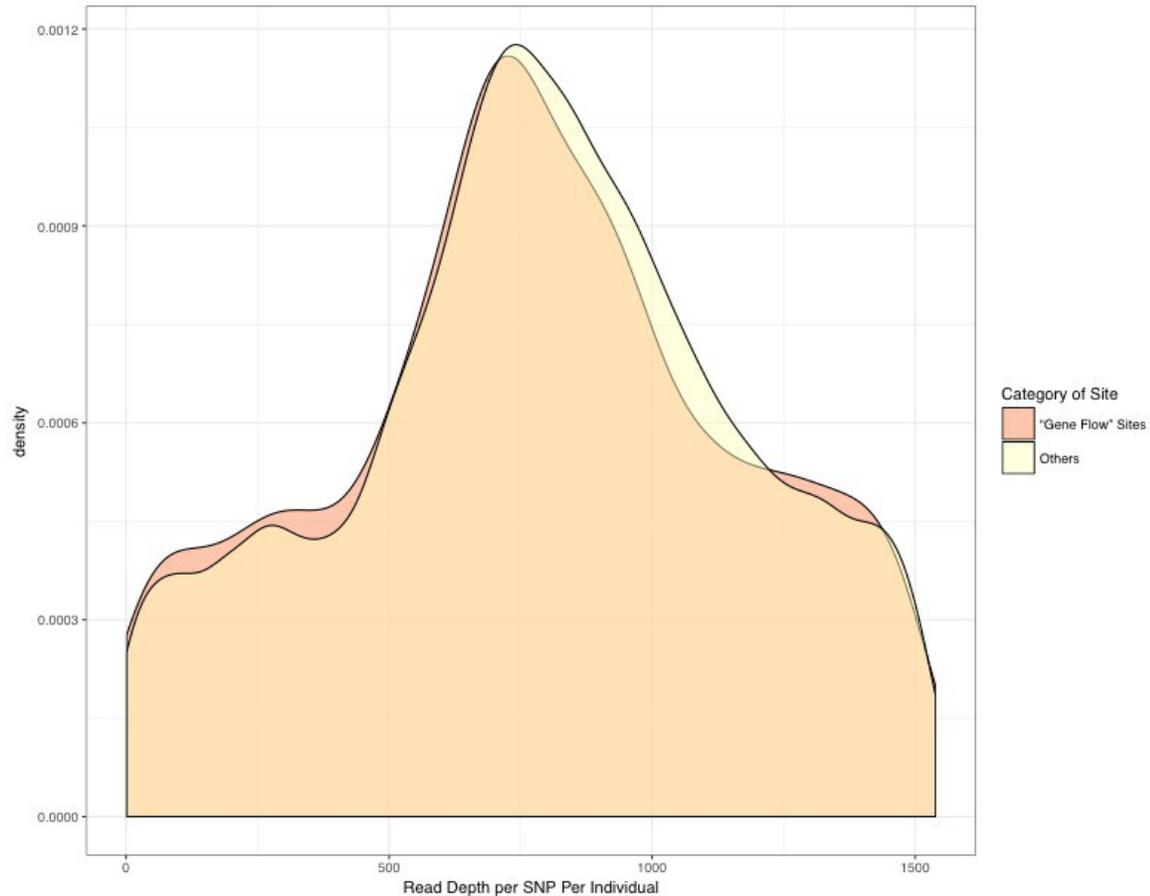
Data set used is the same as for colony level analysis (see Appendix 9). Panels depict graphical representations of several methods of selecting the minimal number of clusters needed to explain most of the variation in our data set.



Appendix 12: Number of clusters (K) for the for the proportion of sites which differ in a within-colony comparison reported from different algorithms

method	K inferred
kl	8
ch	0
hartigan	4
mcclain	2
gamma	2
gplus	2
tau	4
dunn	2
sdindex	3
sdbw	3
cindex	9
silhouette	2
ball	3
ptbserial	2
gap	2
frey	2

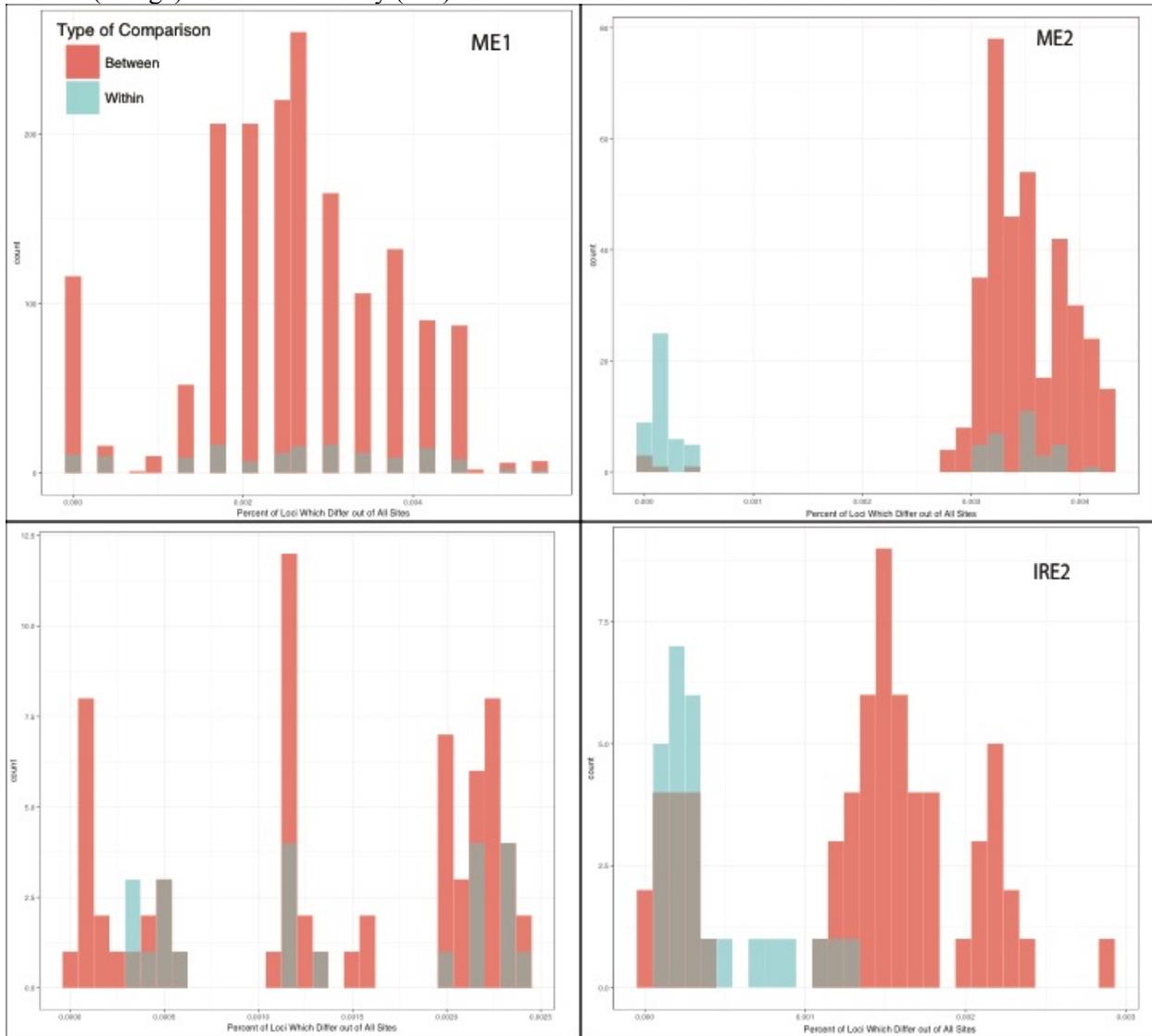
Appendix 13: Comparison of read-depth distributions between SNPs where polyps were alternative homozygotes vs. all other SNPs



SNPs were taken to be evidence of polyp fusion into a particular colony if at least one polyp in the colony differed from the clonal genotype at this SNP by more than one mutational/error step (that is, polyps shared no alleles at this SNP). SNPs were added to the data set for this figure in a per-colony basis. That is, the only read depths considered for the potential sites were those depths associated with that particular colony in which the pattern appeared at that site, and the sites in that colony that were not the alternate homozygote were included as the “others”.

Appendix 14: Histogram of the number of sites at which polyps differ in comparisons between colonies at a given collecting site

Colors: (orange) or within a colony (teal). Data sets are the same as for main Table 3.2.



Appendix 15: Between-colony pairwise Fst values for collecting sites

ME1

	ME1.9	ME1.10	ME1.2	ME1.3	ME1.4	ME1.5	ME1.6	ME1.7	ME1.8
ME1.1	0.095	0.050	0.184	0.171	0.157	0.052	0.091	0.190	0.157
ME1.9		0.094	0.196	0.197	0.153	0.085	0.153	0.165	0.191
ME1.10			0.165	0.153	0.142	0.067	0.116	0.171	0.143
ME1.2				0.278	0.302979	0.119	0.214	0.289	0.286
ME1.3					0.281	0.091	0.195	0.255	0.234
ME1.4						0.090	0.226	0.249	0.266
ME1.5							0.080	0.090	0.078
ME1.6								0.262	0.200
ME1.7									0.209

ME2

	ME2.1	ME2.2	ME2.3	ME2.4	ME2.5
ME2.5		0.142421	0.0870675	0.146802	0.115565
ME2.6			0.148295	0.176256	0.18661
ME2.7				0.120654	0.129049
ME2.8					0.123994

IRE1

	IRE1.1	IRE1.2
IRE1.1		0.256271

IRE2

	IRE2.1	IRE2.2
IRE2.1		0.273354

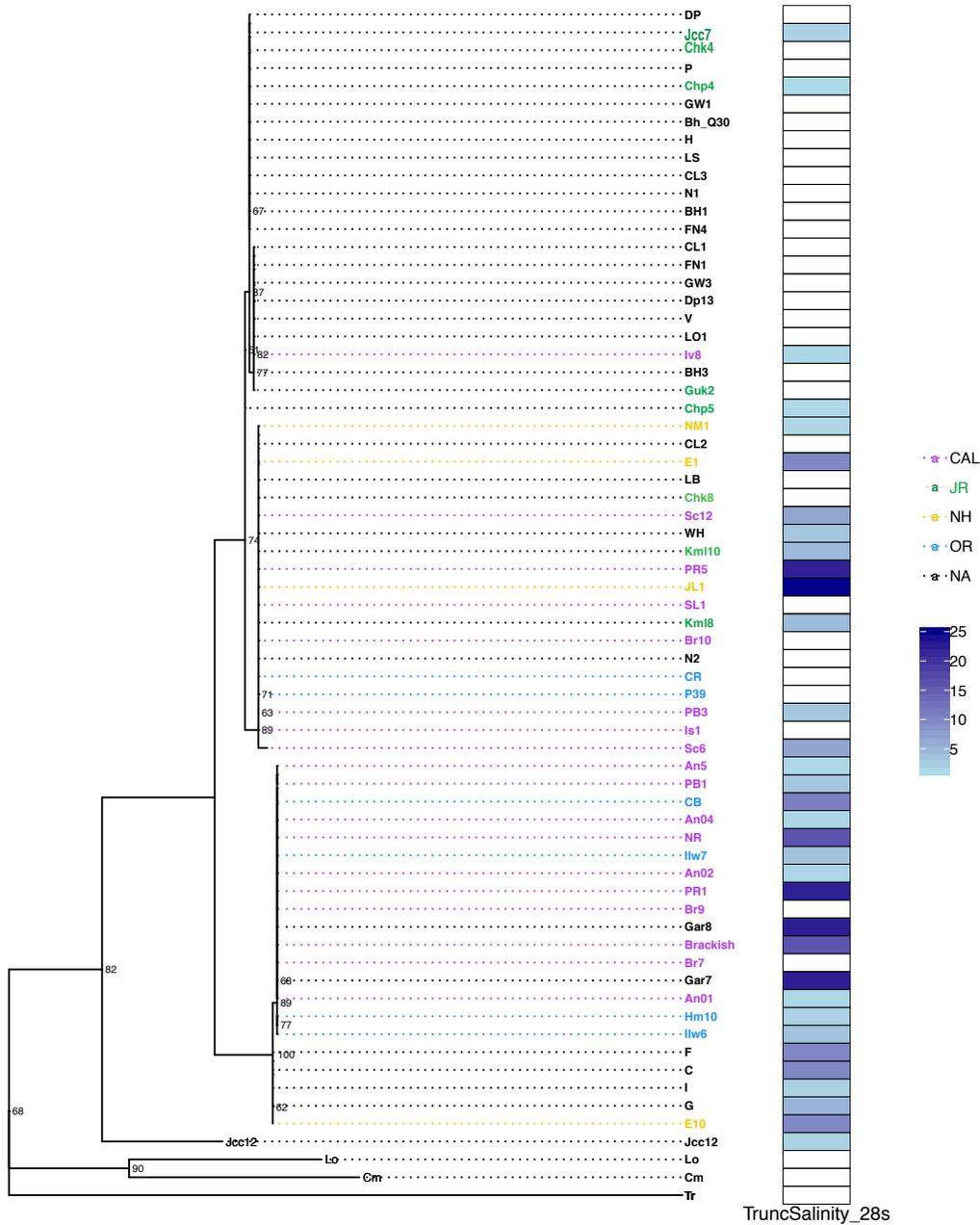
Appendix 16: Summary of major per-colony results

ID	Relationship Types	Gene flow?	Ratio of genotypes to polyyps
ME1.1	I,III	Yes	0.4
ME1.2	I	No	0.2
ME1.3	I	No	0.2

ME1.4	I	No	0.17
ME1.5	I,II,III	Yes	0.717
ME1.6	I,II,III	Yes	.43
ME1.7	I,II	Yes	0.25
ME1.8	I,II	Yes	0.34
ME1.9	I,II	Yes	0.4
ME1.10	I,II	Yes	0.4
ME2.1	I,II,III	Yes	0.57
ME2.2	I,III	Yes	0.34
ME2.3	I,II,III	Yes	0.38
ME2.4	I,II	Yes	0.4
ME2.5	I,II	Yes	0.4
IRE1.1	II	Yes	1
IRE1.2	I,II	Yes	0.5
IRE2.1	I,II	Yes	0.4
IRE2.2	I,II	Yes	0.6

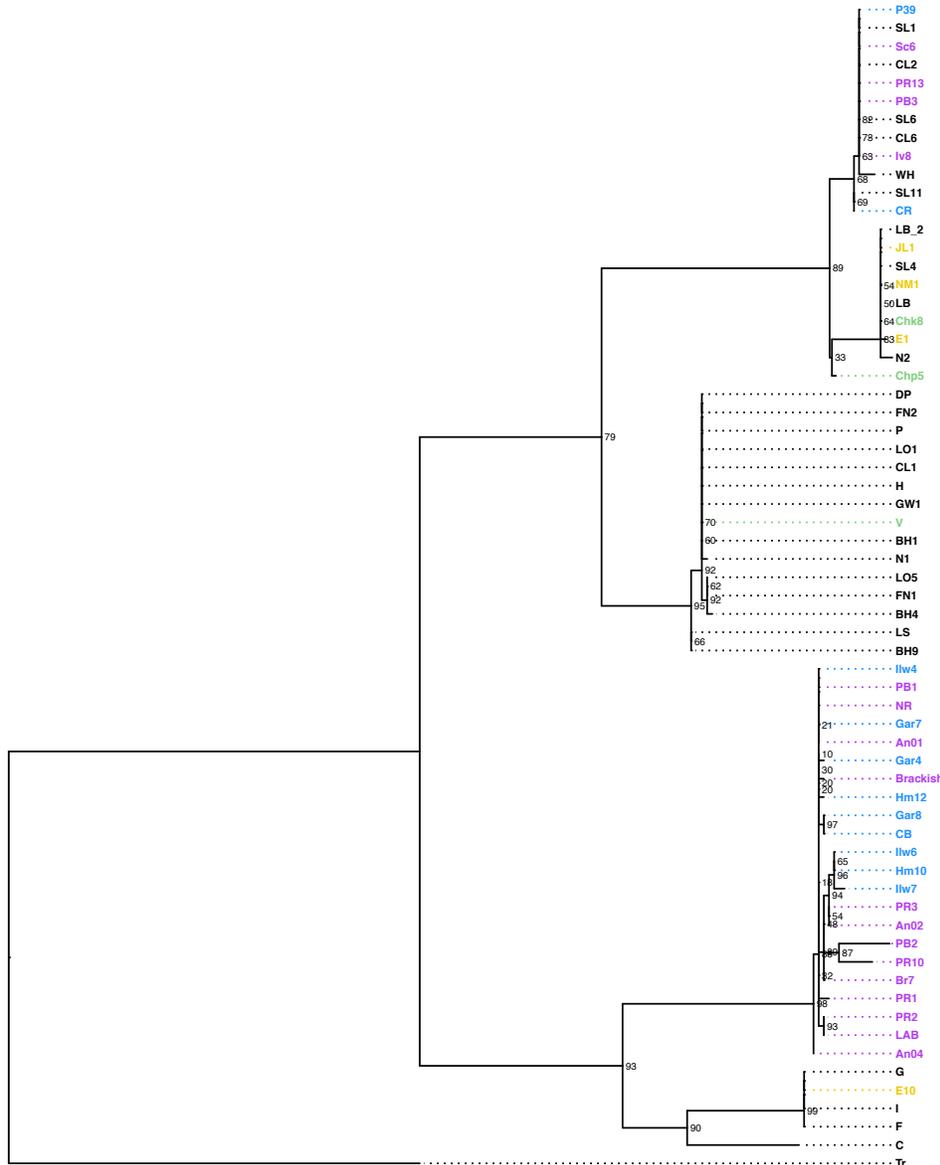
Appendix 17: Phylogenetic tree of *Cordylophora* based on maximum-likelihood analysis of 28S markers

Phylogenetic reconstruction was carried out in IQTree, with numbers at the nodes representing percent bootstrap support out of 10,000 ultrafast bootstrap replicates. Colored taxon names indicate newly sampled estuarine samples, and salinity is represented by colors in heatmap, with white representing values <1.0 ppt. Taxa with black labels are not part of one of the estuary systems, and generally represent individuals from the Great Lakes, Finger Lakes and Europe (see Table 4.1).



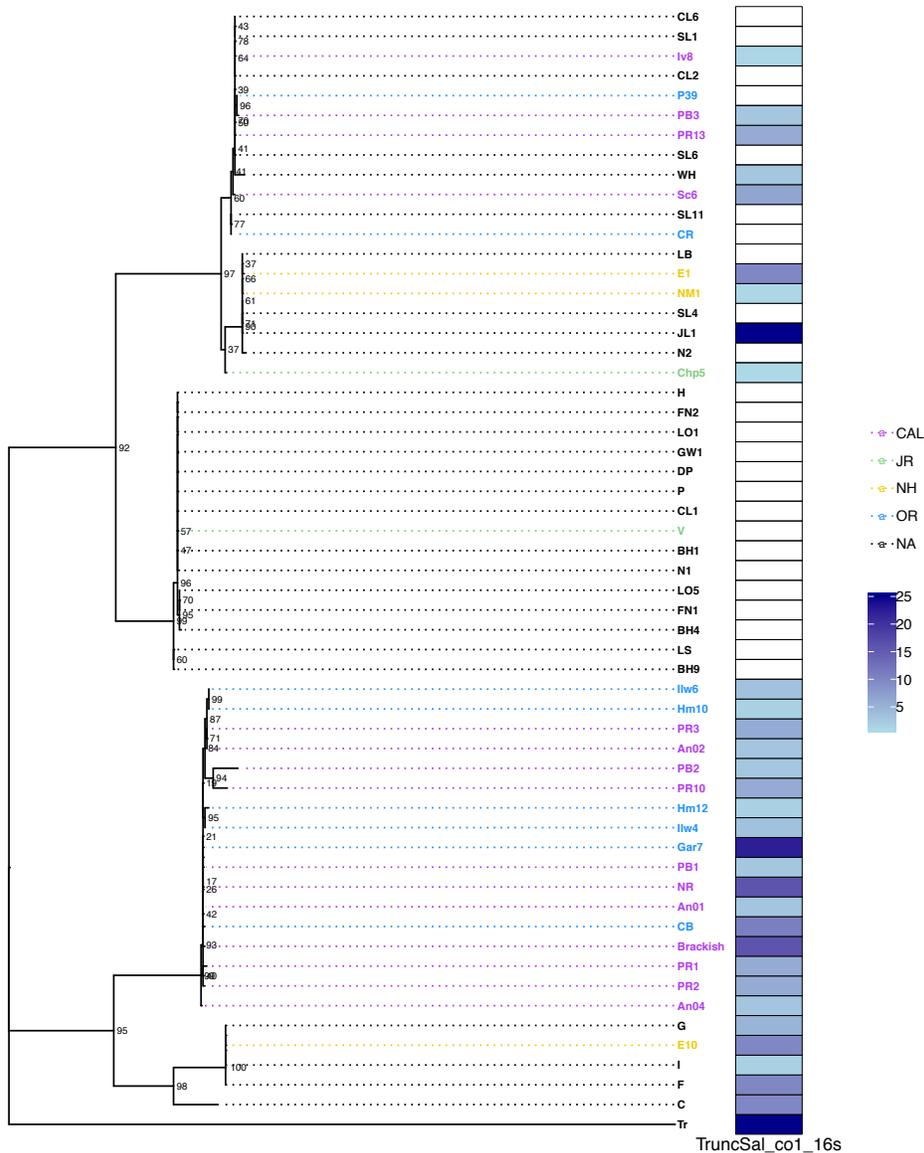
Appendix 18: Phylogenetic tree of *Cordylophora* based on maximum-likelihood analysis of CO1 sequences

Phylogenetic reconstruction was carried out in IQTree, with numbers at the nodes representing percent bootstrap support out of 10,000 ultrafast bootstrap replicates. Colored taxon names indicate newly sampled estuarine samples, and salinity is represented by colors in heatmap, with white representing values <1.0 ppt, colors as in Figure 2. Taxa with black labels are not part of one of the estuary systems, and generally represent individuals from the Great Lakes, Finger Lakes and Europe (see Table 4.1).



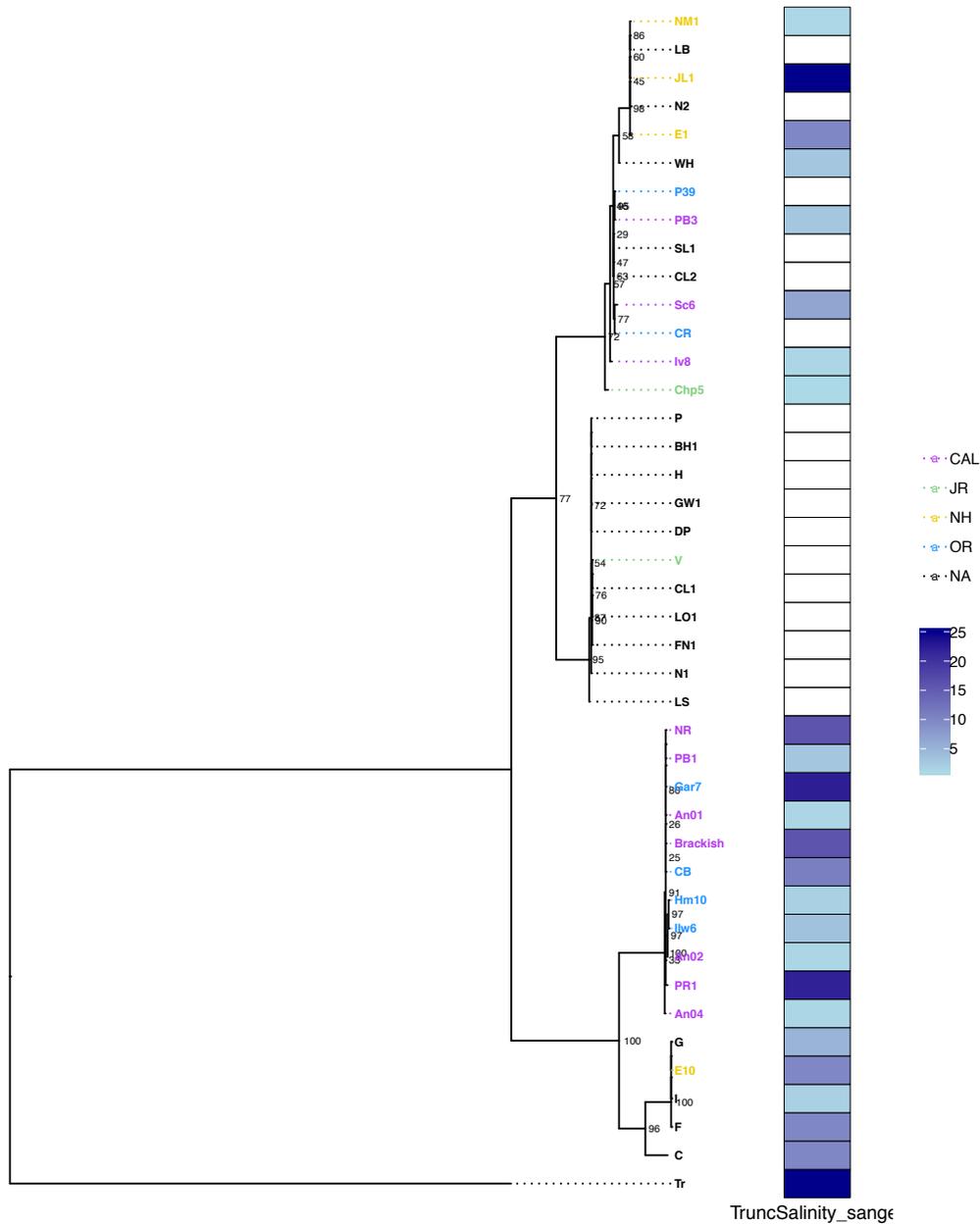
Appendix 19: Phylogenetic tree of *Cordylophora* based on maximum-likelihood analysis of concatenated CO1+16s sequences

Phylogenetic reconstruction was carried out in IQTree, with numbers at the nodes representing percent bootstrap support out of 10,000 ultrafast bootstrap replicates. Colored taxon names indicate newly sampled estuarine samples, and salinity is represented by colors in heatmap, with white representing values <1.0 ppt, colors as in Figure 2. Taxa with black labels are not part of one of the estuary systems, and generally represent individuals from the Great Lakes, Finger Lakes and Europe (see Table 4.1).



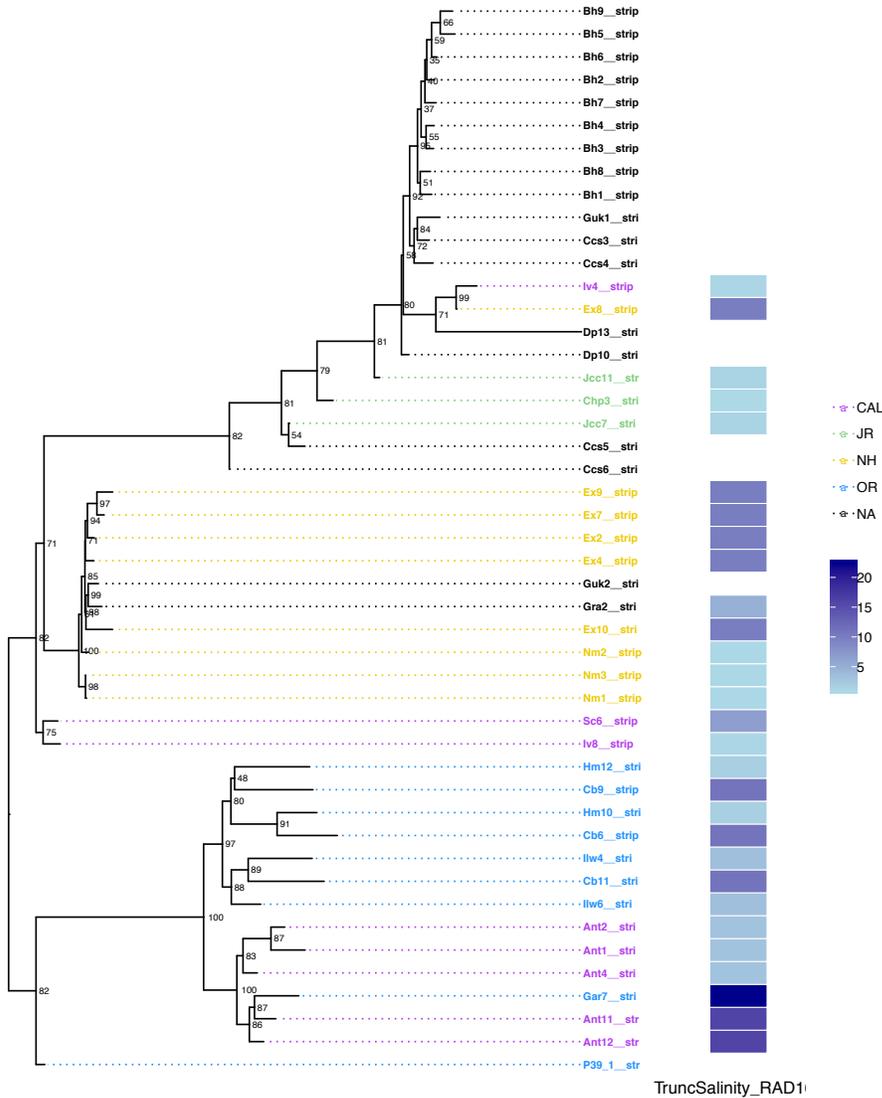
Appendix 20: Phylogenetic tree of *Cordylophora* based on maximum-likelihood analysis of concatenated 16s+28s+CO1 sequences

Phylogenetic reconstruction was carried out in IQTree, with numbers at the nodes representing percent bootstrap support out of 10,000 ultrafast bootstrap replicates. Colored taxon names indicate newly sampled estuarine samples, and salinity is represented by colors in heatmap, with white representing values <1.0 ppt, colors as in Figure 2. Taxa with black labels are not part of one of the estuary systems, and generally represent individuals from the Great Lakes, Finger Lakes and Europe (see Table 4.1).



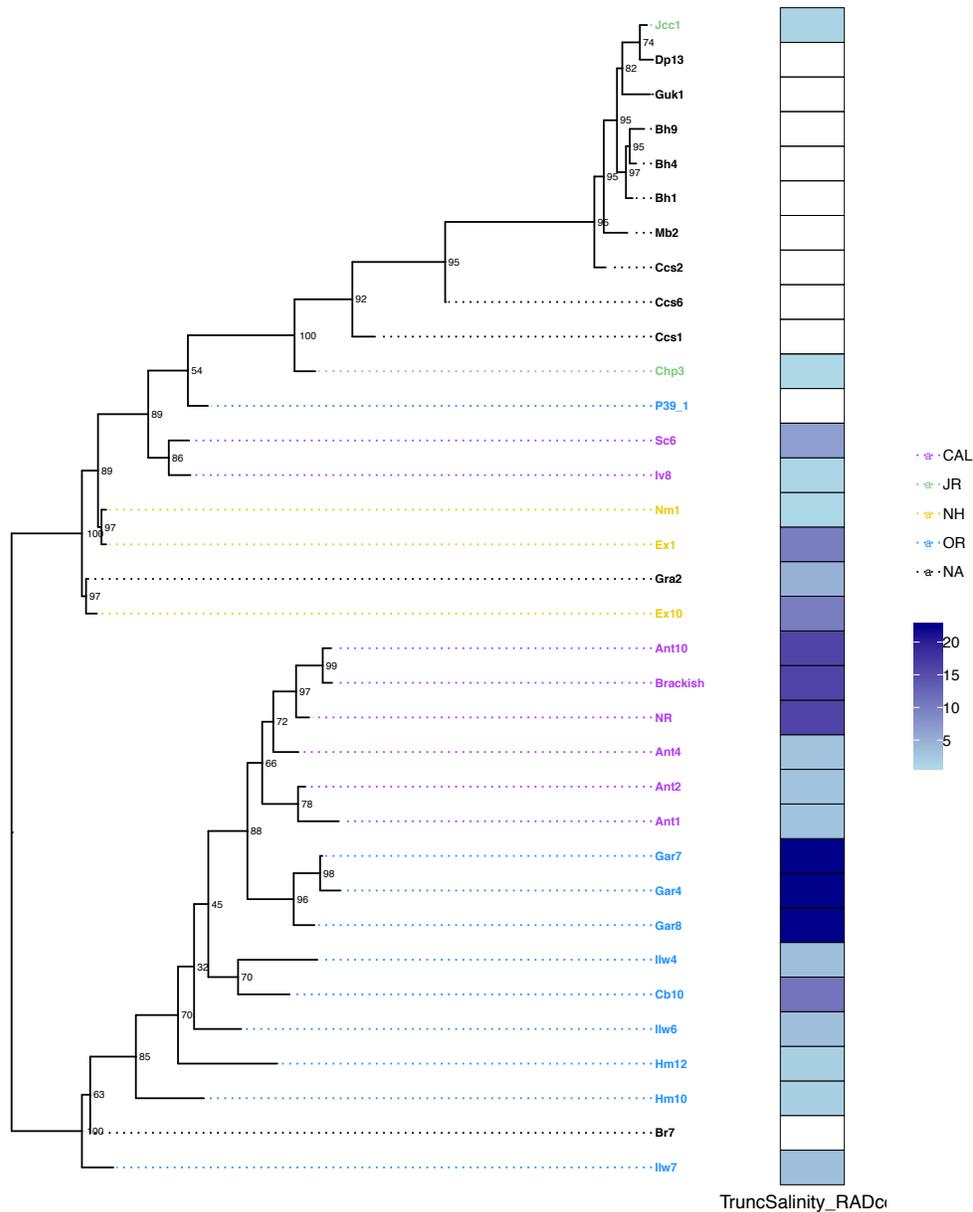
Appendix 21: Phylogenetic tree of *Cordylophora* based on maximum-likelihood analysis combined RAD sequence and Sanger data

Phylogenetic reconstruction was carried out in IQTree, with numbers at the nodes representing percent bootstrap support out of 10,000 ultrafast bootstrap replicates. Colored taxon names indicate newly sampled estuarine samples, and salinity is represented by colors in heatmap, with white representing values <1.0 ppt, colors as in Figure 2. Taxa with black labels are not part of one of the estuary systems, and generally represent individuals from the Great Lakes, Finger Lakes and Europe (see Table 4.1).



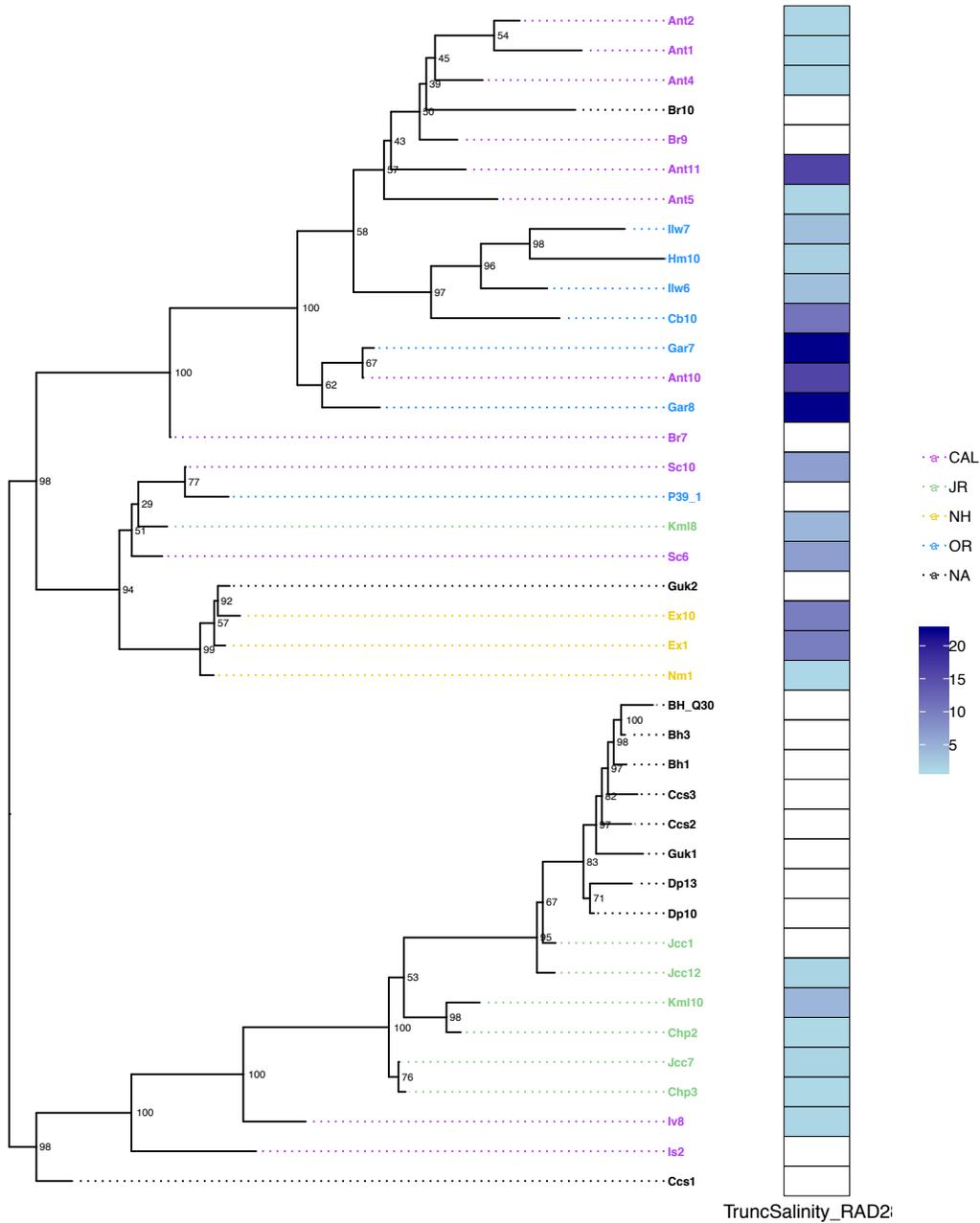
Appendix 21B: Phylogenetic tree of the *Cordylophora* species complex based on maximum-likelihood analysis combined RAD sequence and CO1 sequence data.

Phylogenetic reconstruction was carried out in IQTree, with numbers at the nodes representing percent bootstrap support out of 10,000 ultrafast bootstrap replicates. Colored taxon names indicate newly sampled estuarine samples, and salinity is represented by colors in heatmap, with white representing values <1.0 ppt, colors as in Figure 2. Taxa with black labels are not part of one of the estuary systems, and generally represent individuals from the Great Lakes, Finger Lakes and Europe (see Table 1).



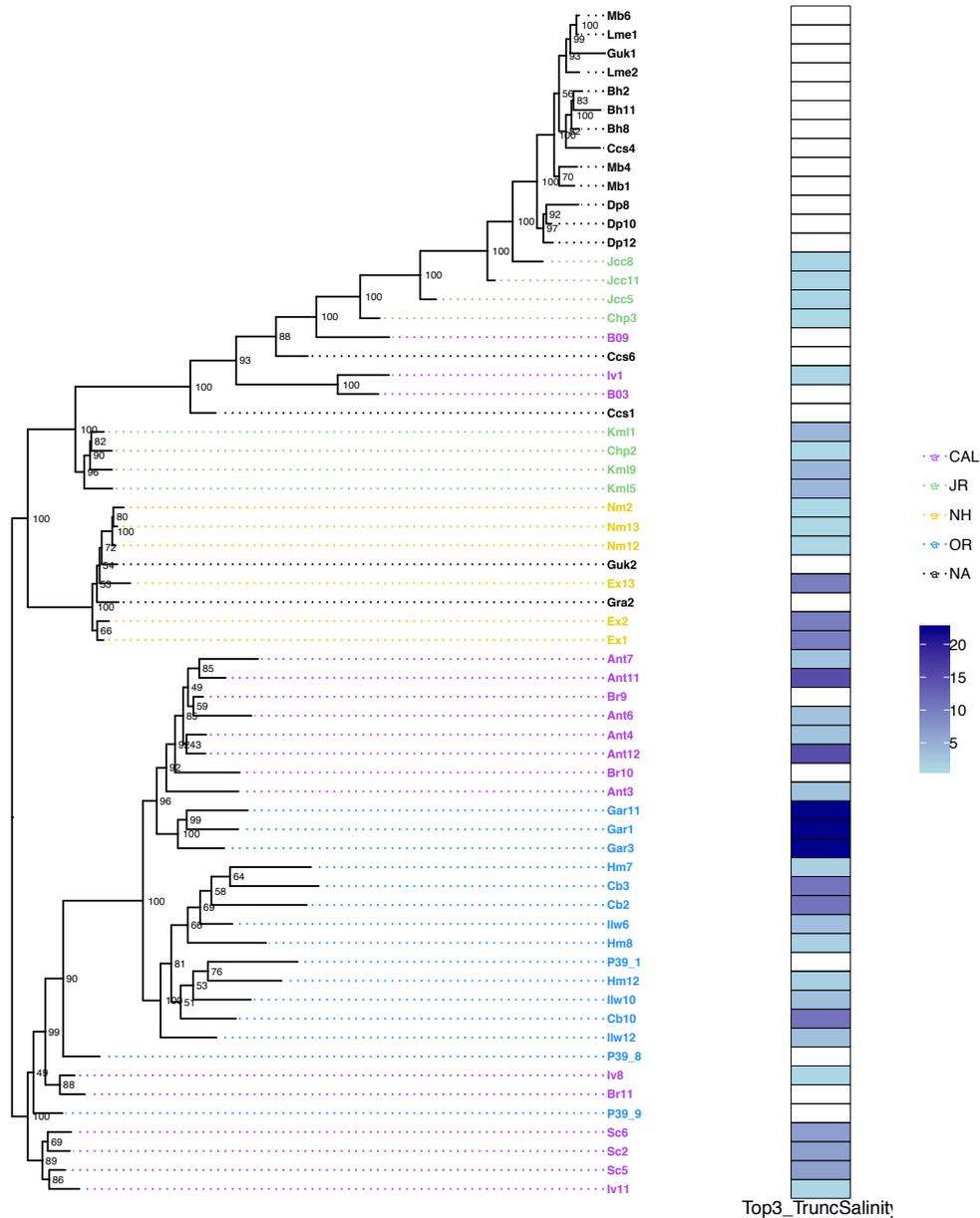
Appendix 21C: Phylogenetic tree of the *Cordylophora* species complex based on maximum-likelihood analysis combined RAD sequence and 28s sequence data.

Phylogenetic reconstruction was carried out in IQTree, with numbers at the nodes representing percent bootstrap support out of 10,000 ultrafast bootstrap replicates. Colored taxon names indicate newly sampled estuarine samples, and salinity is represented by colors in heatmap, with white representing values <1.0 ppt, colors as in Figure 2. Taxa with black labels are not part of one of the estuary systems, and generally represent individuals from the Great Lakes, Finger Lakes and Europe (see Table 1).



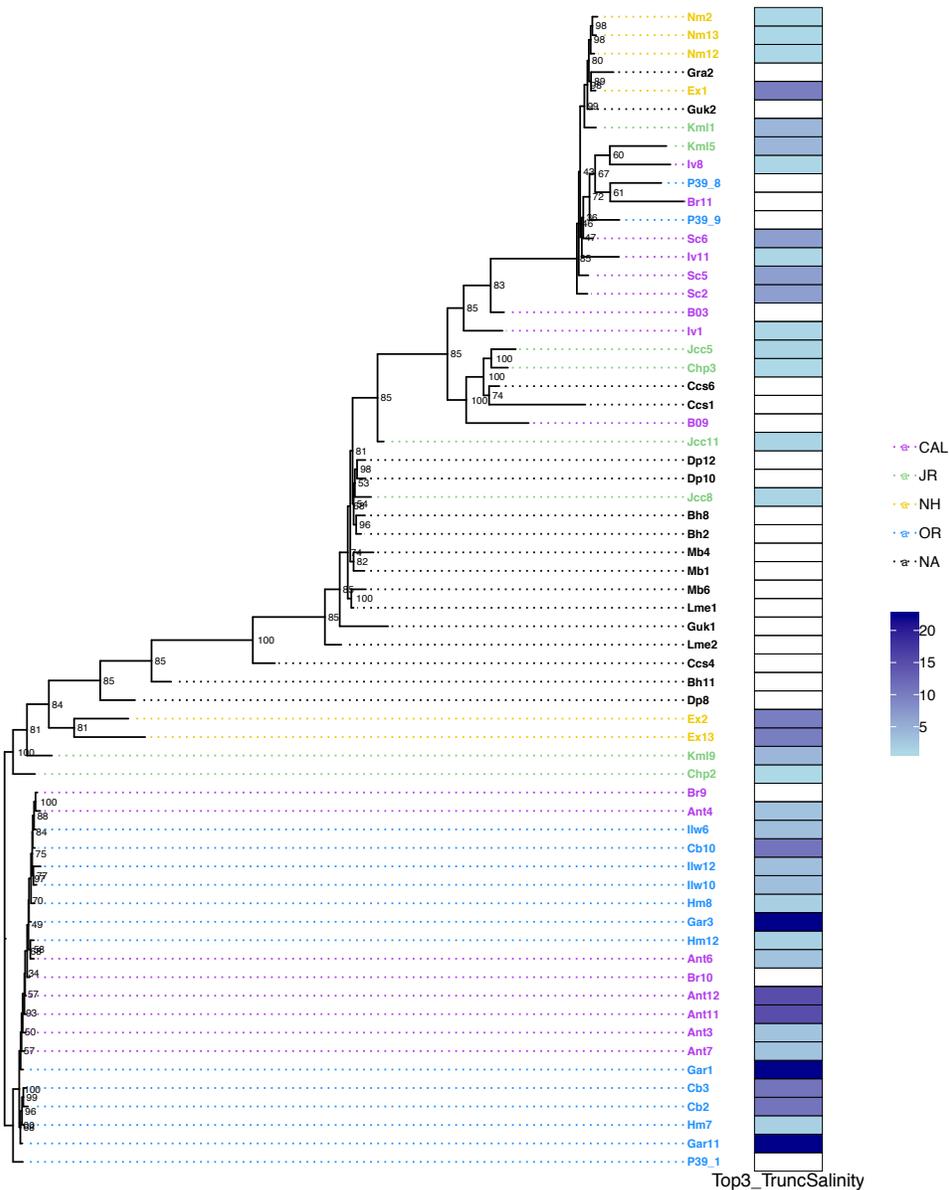
Appendix 22: RAD-seq maximum-likelihood phylogeny of *Cordylophora* based on an alignment of unlinked nucleotide sites for three colonies per locality

Phylogenetic reconstruction was carried out in IQTree, with numbers at the nodes representing percent bootstrap support out of 10,000 ultrafast bootstrap replicates. Colored taxon names indicate newly sampled estuarine samples, and salinity is represented by colors in heatmap, with white representing values <1.0 ppt, colors as in Figure 4.5.

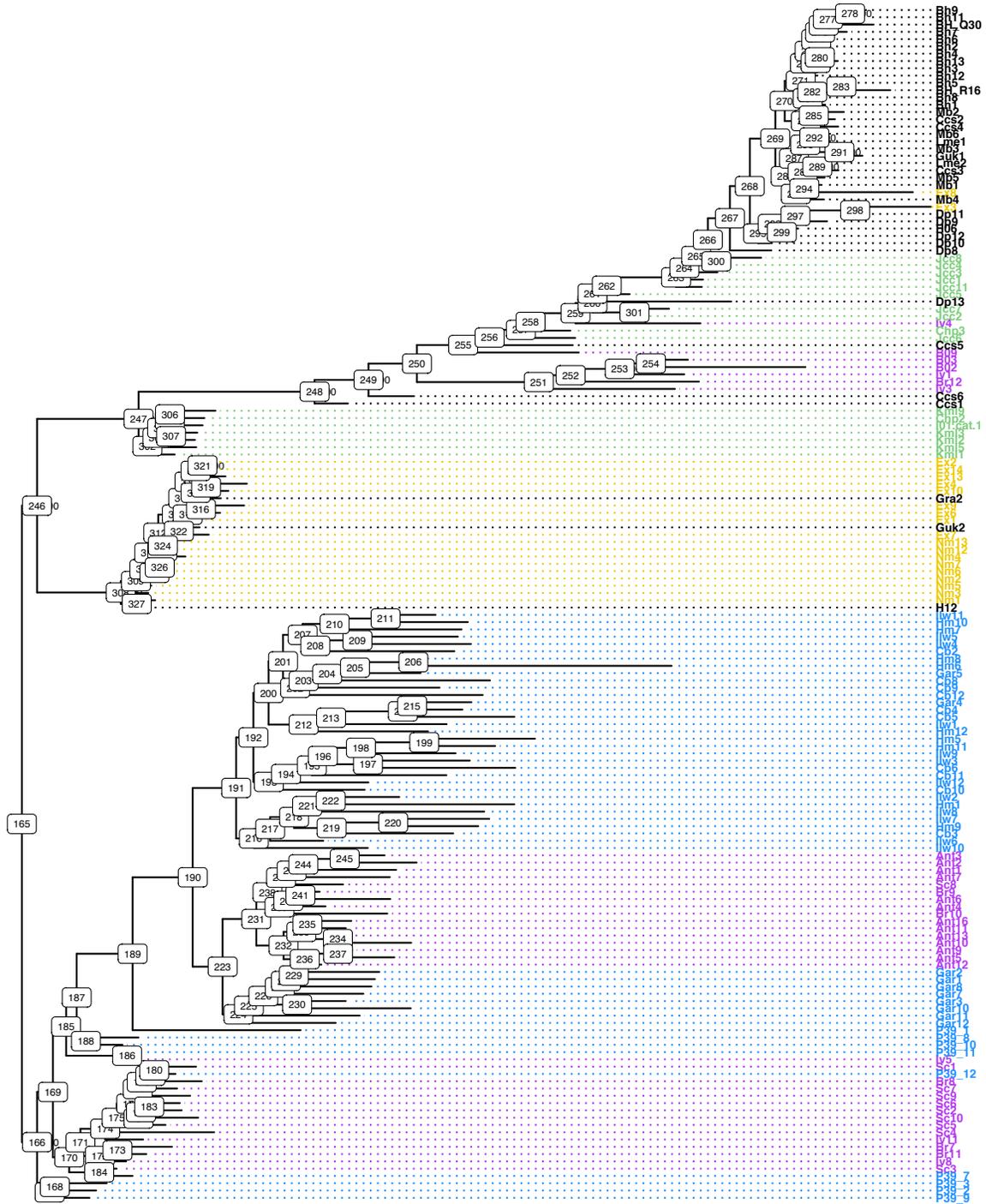


Appendix 23: RAD-seq maximum-likelihood phylogeny of *Cordylophora* based on an alignment of unlinked nucleotide sites for three colonies per locality, 50% missing data

Phylogenetic reconstruction was carried out in IQTree, with numbers at the nodes representing percent bootstrap support out of 10,000 ultrafast bootstrap replicates. Colored taxon names indicate newly sampled estuarine samples, and salinity is represented by colors in heatmap, with white representing values <1.0 ppt, colors as in Figure 4.5.



Appendix 24: RAD-seq maximum-likelihood phylogeny of *Cordylophora* with node numbers for interpretation of Appendix 25



Appendix 25: Summary of Bayesian ancestral character state reconstruction for each internal node of the RAD-seq phylogeny

Analysis was conducted for 10,000 MCMC steps per node and sampled every 100 steps.

Node	Avg. Salinity	Min Salinity	Max Salinity	St. Dev.
165	1.95	-0.56	3.73	1.15
166	2.46	0.46	4.09	0.89
167	1.50	-0.54	3.05	0.76
168	1.33	-0.88	2.88	0.75
169	2.95	0.84	5.02	1.01
170	2.78	0.69	4.62	0.79
171	2.56	0.04	4.09	0.82
172	2.00	0.31	4.35	0.97
173	1.24	-1.81	3.24	0.95
174	2.35	0.62	4.54	0.94
175	3.03	0.35	5.29	1.12
176	3.43	1.44	5.19	1.03
177	3.12	1.44	4.57	0.78
178	3.29	1.44	5.32	0.82
179	2.13	0.85	4.53	0.88
180	2.39	0.88	4.14	0.86
181	4.32	1.44	7.66	1.41
182	4.40	1.44	7.44	1.46
183	4.99	1.44	7.49	1.25
184	3.57	1.44	5.88	1.09
185	2.51	0.79	4.53	1.04
186	0.30	0.21	1.44	0.29
187	2.10	0.34	3.56	0.67
188	1.87	-0.80	3.82	1.00
189	2.45	-2.88	5.50	2.21
190	4.20	0.77	6.83	1.80
191	3.02	0.69	5.52	1.39
192	2.33	-0.14	6.55	1.83
193	2.20	-1.07	6.26	2.29
194	2.53	-1.45	5.82	1.71
195	3.53	1.15	6.43	1.61
196	3.41	-0.22	6.09	1.80
197	6.15	0.90	11.06	2.22
198	1.77	-4.02	6.26	2.48

199	2.26	-3.19	5.18	2.16
200	2.46	-0.09	5.95	1.67
201	2.16	-0.96	5.15	1.58
202	2.70	0.70	5.17	1.18
203	2.59	-0.37	5.68	1.57
204	4.94	0.27	8.36	1.73
205	6.98	1.26	11.01	2.23
206	1.99	1.44	2.36	0.18
207	2.24	-0.59	6.12	1.87
208	3.41	-0.64	7.34	1.94
209	4.64	-0.43	7.06	1.69
210	1.95	-1.11	5.52	1.25
211	1.69	-1.44	4.98	1.43
212	2.44	-0.53	7.00	1.89
213	4.39	-0.05	9.39	2.35
214	8.23	1.44	12.75	3.32
215	9.41	1.16	15.14	3.78
216	3.59	1.22	5.60	1.10
217	3.89	1.15	6.04	1.39
218	4.02	1.44	6.07	1.10
219	4.99	0.32	8.00	1.97
220	3.34	-1.84	6.89	2.57
221	3.90	1.44	6.91	1.48
222	4.28	0.86	6.84	1.40
223	5.94	-0.27	10.55	2.76
224	7.64	0.35	13.33	3.16
225	7.63	0.99	14.49	3.51
226	7.71	-0.37	16.21	4.68
227	9.54	1.44	17.50	4.94
228	10.80	1.44	17.53	4.59
229	11.53	1.44	16.82	3.84
230	10.83	1.35	19.88	6.37
231	4.26	0.78	8.79	2.48
232	5.27	0.34	11.55	3.23
233	7.03	0.86	12.18	3.42
234	12.11	1.44	16.90	5.07
235	8.73	1.44	14.01	2.98
236	7.53	1.44	11.60	2.80
237	3.55	1.19	6.39	1.50

238	4.30	1.44	7.40	1.27
239	3.82	0.94	5.80	1.09
240	3.13	0.80	5.90	1.13
241	2.97	0.29	5.59	1.38
242	4.36	1.44	7.60	1.50
243	3.97	0.00	7.37	1.80
244	3.36	-0.33	6.48	1.44
245	3.57	1.02	7.50	1.64
246	1.93	-2.50	5.01	1.51
247	2.42	-1.16	5.21	1.70
248	1.20	-1.95	3.89	1.52
249	0.11	-2.87	2.35	1.25
250	1.35	-1.22	4.25	1.41
251	-0.44	-4.75	2.89	2.07
252	0.20	-4.15	4.62	2.31
253	-0.55	-3.79	3.76	1.88
254	-0.65	-4.20	2.04	1.65
255	2.12	-0.05	4.79	1.39
256	1.97	-0.69	5.69	1.35
257	3.12	0.85	5.22	1.12
258	2.56	-0.33	5.60	1.49
259	1.86	-0.57	4.38	1.03
260	1.44	1.44	1.44	0.00
261	1.44	1.44	1.44	0.00
262	0.85	-1.96	4.10	1.24
263	2.09	-0.75	5.58	1.30
264	2.22	0.38	5.61	1.27
265	1.75	-0.02	4.14	0.95
266	1.86	-0.44	3.75	0.95
267	1.17	-1.82	3.23	1.22
268	1.60	-0.68	4.35	1.36
269	1.61	-1.07	4.00	1.04
270	0.86	-1.12	3.11	1.13
271	0.96	-1.52	2.69	1.14
272	0.97	-0.89	3.07	0.98
273	0.84	-0.87	3.36	0.88
274	0.51	-0.97	1.59	0.71
275	0.31	-1.54	1.47	0.84
276	0.31	-0.88	1.46	0.70

277	0.69	-0.86	2.42	0.92
278	0.44	-1.59	2.44	0.97
279	0.38	-0.68	1.92	0.60
280	0.30	-1.13	1.74	0.72
281	0.47	-1.84	2.15	1.00
282	0.23	-2.28	1.73	1.08
283	0.63	-2.20	2.56	0.95
284	0.04	-2.25	2.73	1.48
285	-0.37	-2.65	1.91	1.26
286	1.08	-0.21	2.88	0.66
287	0.99	-0.19	2.77	0.71
288	0.98	-0.80	3.00	1.00
289	0.59	-1.40	1.96	0.85
290	0.61	-0.51	2.94	0.87
291	0.27	-1.95	1.76	0.89
292	0.30	-0.59	1.44	0.38
293	0.71	-0.81	2.29	0.86
294	1.02	-1.67	2.91	1.06
295	1.53	-1.73	3.81	1.31
296	0.85	-1.57	3.18	1.30
297	1.05	-1.12	3.31	1.24
298	0.51	0.39	1.44	0.20
299	0.77	-0.87	3.23	0.81
300	1.50	0.21	4.03	0.98
301	1.54	1.44	1.58	0.06
302	2.53	-1.07	4.97	1.42
303	1.73	-0.44	3.67	0.86
304	1.65	0.42	3.63	0.69
305	1.22	0.38	1.97	0.49
306	1.22	0.44	2.00	0.49
307	2.62	0.31	5.59	1.37
308	0.67	-1.71	3.32	1.63
309	0.90	-1.96	3.70	1.19
310	1.21	-0.06	2.81	0.67
311	1.20	-0.02	2.41	0.51
312	1.06	-0.13	2.29	0.67
313	3.28	0.99	5.30	1.13
314	4.86	0.94	7.01	1.57
315	6.23	1.44	7.89	1.71

316	7.80	1.44	10.92	2.40
317	5.56	1.13	8.47	2.04
318	5.66	1.00	8.64	1.96
319	7.00	0.68	11.27	2.51
320	6.86	1.44	10.18	2.45
321	8.28	1.44	10.97	2.04
322	3.05	1.01	4.52	0.81
323	0.85	-0.02	1.87	0.60
324	0.87	-0.18	1.49	0.53
325	0.99	-0.19	2.16	0.55
326	1.21	1.13	1.44	0.06
327	0.27	-1.75	2.37	0.95

Appendix 26: Summary of Evanno method for Clade 1, not including subclade 1B.

# K	REPS	MEAN LNP(K)	STDEV LNP(K)	LN'(K)	LN''(K)	DELTA K
1	10	-73627.72	164.9531	NA	NA	NA
2	10	-48803.65	6.7894	24824.07	30550.52	4499.730078
3	10	-54530.1	16681.271	-5726.45	7342.4	0.440158
4	10	-52914.15	16799.259	1615.95	34157.37	2.033266
5	10	-85455.57	41437.2765	-32541.42	39121.32	0.944109
6	10	-157118.31	53381.8734	-71662.74	40647.85	0.761454
7	10	-188133.2	102422.564	-31014.89	74740.78	0.72973
8	10	-144407.31	77225.4797	43725.89	57468.78	0.744169
9	10	-158150.2	81729.7161	-13742.89	78683.12	0.962724
10	10	-250576.21	181916.8031	-92426.01	NA	NA

Appendix 27: Summary of Evanno method for Clade 1, including subclade 1B.

# K	REPS	MEAN LNP(K)	STDEV LNP(K)	LN'(K)	LN''(K)	DELTA K
1	10	-84912.96	15.1339	NA	NA	NA
2	10	-34544.66	2.9793	50368.3	49265.29	16536.07361
3	10	-33441.65	31.3697	1103.01	12677.75	404.139473
4	10	-45016.39	11245.8388	-11574.74	11898.8	1.058062
5	10	-44692.33	13804.4626	324.06	20280.87	1.469153
6	10	-64649.14	8317.5685	-19956.81	32066.82	3.855312
7	10	-52539.13	12428.2863	12110.01	11184.7	0.899939
8	10	-51613.82	13743.1488	925.31	24426.43	1.777353
9	10	-75114.94	25971.7298	-23501.12	24131.69	0.929152
10	10	-74484.37	17876.8949	630.57	NA	NA