THE EVOLUTION AND ECOLOGY OF CAPITELLA (ANNELIDA, CAPITELLIDAE) IN THE GULF OF MEXICO

A Dissertation

by

JUSTIN LEE HILLIARD

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Anja Schulze
Committee Members,	Jaime Alvarado-Bremer
	Michelle Lawing
	Wesley Highfield
Intercollegiate Faculty Chair,	Anja Schulze

May 2020

Major Subject: Marine Biology

ABSTRACT

Members of genus *Capitella* (Annelida: Capitellidae) are often found in high abundance in areas of anthropogenic disturbance, such as fish farms and waste-water outflow sites and are frequently cited as ecological indicators of organic pollution. Understanding evolutionary relations and habitat preferences between *Capitella* species that can tolerate these, and other, harsh environments and those that cannot is important for their continued and improved use as indicators of pollution.

Five previously undocumented species of *Capitella* and *Capitella nonatoi*, a species described from Brazil, were detected in the Gulf of Mexico (GoM) by DNA barcoding. Most of the new species were found in a single location, with one being distributed throughout the GoM. Two of the proposed species are supported by distinctive life history characteristics. These findings underscore the potential to uncover large amounts of biodiversity in the GoM, a region subject to many anthropogenic and natural disturbances. Additionally, support was found for a single evolutionary origin of acicular spines in *Capitella*, which seems to be a morphology unique to Western Atlantic estuarine waters.

Comparing six species abundance modeling techniques using internal validation metrics with six capitellids in Tampa Bay, Florida indicated that none of the assessed models works best for all species. However, Hurdle and GAM-Tweedie models had good performance overall. This was attributed to how these models handle zero-inflation, which every species had. Species rarity was influential and required consideration. For example, *Capitella aciculata* was found to be a very rare species and this restricted model specification, resulting in the removal of one of the covariates. Assessment of environmental term importance indicated that depth and bay segment/region are important across all species, with higher abundance in shallow, near-shore regions of the bay.

Investigation of the evolution of hypoxia inducible factor (HIF), a key transcription factor in the cellular oxygen-sensing pathway consisting of an alpha and beta subunit, revealed high diversity across Annelida. Some recognized groups of annelids were supported by both gene phylogenies. However, neither of the two genes mirrored current hypotheses of annelid phylogenetics but HIFβ reflected current annelid phylogeny hypotheses more closely, indicating stronger conservation of this gene. Additionally, the protein domains of the two genes were recovered with varying degrees of success. This was attributed to loss of low-quality data during transcriptome assembly and high divergence of the domains.

These findings contribute to our understanding of *Capitella* species diversity, patterns of occurrence, and potential for low-oxygen tolerance. A key component to understanding how *Capitella* have come to occupy so many different marine habitats (e.g. sulfide vents, deep-sea wood falls, squid egg masses) lies in understanding their functional response to the low oxygen levels they encounter in some of these habitats. This will provide insights into the evolution of the HIF transcription factor across Annelida and its potential role in speciation across the phylum.

DEDICATION

In memory of my mother. Her immense love and support continue to drive me to new accomplishments. This is for you, Mom.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Anja Schulze, for her commitment to my graduate education. She has supported me every step of the way from blazingly fast revisions to acquiring grants to support my work and to listening to me and understanding when life happened. Thank you for helping me become the scientist I am today.

To the members of my committee, Dr. Michelle Lawing, Dr. Wesley Highfield, and Dr. Jaime R. Alvarado Bremer, thank you for your time and support. You all provided insightful conversations and challenged me to think more deeply about the methods I was using. I especially appreciate your perspectives of interpretation of my results. Thank you for your guidance and enhancing my critical thinking.

My field and lab work would not have been possible without the support of many people. I will do my best to name you all. Please forgive me if I do not directly thank you. Thank you to Ren Shirai, Heidi Jane deGuzman, Marissa Hajduk, Candace Grimes, MC Hannon, Dr. Anna Armitage, Ashley Whitt, Caroline Weaver, Nathan Reed, David Karlen, Jennifer and Scott Davenport, David Graff, Kevin Cunniff, Nuria Méndez, Tama, Pepe, Zach Hancock, Hillsborough County Environmental Protection Commission, and all of the field crews who helped with the Mexico sampling.

I would like to thank Holly Richards and Nicole Kinslow for your outstanding support of the Galveston Graduate Community. You make the experience of graduate school easier to handle and I certainly appreciate everything you do. I would also like to thank Diego Agudelo for supporting my research using the Texas A&M University High Performance Research Cluster (TAMU HPRC). Thank you to everyone who has contributed to the TAMU HPRC. It is a wonderful resource that made some of my research possible.

I would like to thank the Graduate Community at Texas A&M University Galveston Campus (TAMUG), past and present. We have a great community and I made many memories with you all. Thank you.

Many thanks to all my family and friends who supported me through this adventure. I know that I tend to ramble on about my research and it doesn’t always make sense, but your listening means so much to me. I have had several revelations while exploring different ways of communicating my work. Thank you.

Last, I would like to give a big thank you to my partner, Michelle. You have been by my side for this entire adventure. You have patiently listened to and supported me in every way. Thank you, Sweetheart.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Anja Schulze [advisor] and Jaime Alvarado-Bremer of the Department of Marine Biology, Wesley Highfield of the Department of Marine and Coastal Environmental Science, and Michelle Lawing of the Department of Ecosystem Science and Management.

Funding Sources

Graduate study was supported by two fellowships from Texas A&M University. Research funding was made possible in part by the National Science Foundation under Grant Number DEB-1036186 and CONACYT under Grant Number 2017-25.

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iv

ACKNOWLEDGEMENTS v

CONTRIBUTORS AND FUNDING SOURCES vii

TABLE OF CONTENTS viii

LIST OF FIGURES x

LIST OF TABLES xiv

1. INTRODUCTION 1

1.1. References 5

2. CAPITELLA (ANNELIDA: CAPITELLIDAE) SPECIES IN THE GULF OF MEXICO: DELIMITATION, PHYLOGEOGRAPHIC PATTERNS, AND PHYLOGENY OF THE GENUS 9

2.1. Introduction 9
2.2. Materials and Methods 11
2.2.1. Specimen Collection/Storage 11
2.2.2. DNA Purification and Sequencing 14
2.2.3. Sequence Analysis 16
2.2.4. Morphological Observations 18
2.3. Results 20
2.3.1. COI Analysis 20
2.3.2. H3 Analysis 26
2.3.3. Morphology 28
2.4. Discussion 29
2.5. References 32

3. COMPARATIVE SPECIES ABUNDANCE MODELING OF CAPITELLIDAE (ANNELIDA) IN TAMPA BAY, FLORIDA 39

3.1. Introduction 39
3.2. Materials and Methods 42

3.2.1. Data Collection 42
3.2.2. Database Acquisition and Filtering 48
3.2.3. Spatial Statistics 48
3.2.4. Species Abundance Modeling 50
3.2.5. Environmental Factors 52
3.3. Results 53
3.3.1. Spatial Statistics 53
3.3.2. Species Abundance Modeling 63
3.3.3. Environmental Factors 65
3.4. Discussion 69
3.5. References 74

4. ANALYSIS OF ANNELID TRANSCRIPTOMES AND GENOMES REVEALS HIGH DIVERSITY OF THE OXYGEN SENSING TRANSCRIPTION FACTOR HYPOXIA INDUCIBLE FACTOR (HIF) 87

4.1. Introduction 87
4.2. Materials and Methods 90
4.2.1. Public Data Sourcing and Transcriptome Assembly 90
4.2.2. Capitella sp. TV RNA-Sequencing and Transcriptome Assembly 91
4.2.3. Open Reading Frame (ORF) Prediction and Basic Local Alignment Tool (BLAST) Searches 92
4.2.4. Gene Identification by Phylogenetic Analysis 92
4.2.5. Isoform Selection and Gene-Specific Phylogenetic Analyses 93
4.2.6. Domain Annotation 93
4.3. Results 94
4.3.1. BLAST Searches and Gene Identification by Phylogenetic Analysis 94
4.3.2. Gene-Specific Phylogenies 97
4.3.3. Domain Annotation 98
4.4. Discussion 101
4.5. References 103

5. CONCLUSIONS 112

5.1. References 116

APPENDIX A COMPARATIVE SPECIES ABUNDANCE MODELING OF CAPITELLIDAE (ANNELIDA) IN TAMPA BAY, FLORIDA 118

LIST OF FIGURES

Page

Figure 1.1: Modified from Hilliard et al. (2016). Scanning electron micrographs of *Capitella* cf. *capitata* from Galveston Bay, Texas. A. Fragmented specimen with presegmental structures labeled. B. Fragmented specimen with nine thoracic segments numbered and attention drawn to the ventral groove. C. Hooded hooks on right side of segment 11 of specimen shown in image A. D. Genital spines on segments 8 and 9 of specimen shown in image A. E. Mixed neurochaetae (capillary chaetae and hooded hooks) on left side of segment 7. CC, capillary chaeta; F, main fang; GS, genital spine; HH, hooded hook; NuO, nuchal organ; Per, peristomium; Pro, prostomium; VGr, ventral groove. Scale bars: A=500 µm; B=200 µm; C & E=15 µm; D=50 µm. 2

Figure 2.1: Sampling locations throughout the Gulf of Mexico. The shapefiles used to make this map are from version 2.3.7 of the Global Self-consistent, Hierarchical, High-resolution Geography (GSHHG) Database, accessible at http://www.ngdc.noaa.gov/mgg/shorelines/gshhs.html...................................... 12

Figure 2.2: Maximum likelihood phylogenetic tree based on COI sequences. Branch labels are bootstrap support values|posterior probabilities. Details of each clade's geographic origin and species name if known are listed. The clades with species where acicular spines have been documented are indicated by green boxes. 21

Figure 2.3: Median joining haplotype network of COI for *Capitella* sp. TV (Clade 10, Figure 2.2). Locations marked by a triangle in the legends had some specimens with acicular spines. One hash represents a single mutation. Small black circles within the network represent a predicted haplotype that is not present in the data set... 23

Figure 2.4: Median joining haplotype network of COI *Capitella nonatoi* clade. One hash represents a single mutation. Small black circles within the network represent a predicted haplotype that is not present in the data set. 24

Figure 2.5: K2P genetic distances based on COI. Only distances between groups are shown and no within-group values are reported. The color scheme for the distance values can be interpreted as column = upper right and row = lower left. For example, the genetic distance between *Capitella teleta* and *Capitella aracaensis* is 0.193, or 19.3%.. 25

Figure 2.6: Bayesian Inference phylogenetic tree based on H3 sequences. Branch labels are bootstrap support values/posterior probabilities. Branches are only labeled as they are relevant to the Discussion Section. 27

Figure 2.7: Scanning electron micrographs. *Capitella* sp. TF is in all top images (A-C). A. Complete specimen for total length. B. Thoracic region of specimen with chaetigers 1-4 labeled. C. Dorsal view for emphasis of incomplete intersegmental furrows. *Capitella* sp. TV is in all bottom images (D-F). D. Complete specimen from Galveston, TX for total length. E. Thoracic region of specimen from Tampa, FL with chaetigers 1-4 labeled. F. Dorsal view of specimen from Suter Park, Corpus Christi, TX for emphasis of complete intersegmental furrows. Scale bars: A&D=1 mm; B&E=500 µm; C&F=250 µm. 28

Figure 3.1: Light micrographs of Capitellidae species highlighting diagnostic characters. All specimens are from the Hillsborough County Environmental Protection Commission’s samples and were photographed in-house. Capitellids are characterized by having thoracic and abdominal regions, with the number of thoracic chaetigers being the primary diagnostic for genera. The arrangement and type(s) of chaetae are also important. (a): *Capitella capitata* complex specimen at 4X magnification with capillary chaetae highlighted. (b): *Capitella aciculata* specimen at 40X magnification with acicular spine highlighted. Thoracic chaetigers 1-9 are labeled to differentiate thorax from abdomen. (c): *Mediomastus ambiseta* specimen at 4X magnification with abdominal capillary chaetae highlighted. (d): *Mediomastus californiensis* specimen at 40X with long abdominal hooded hooks highlighted. 43

Figure 3.2: Map of Tampa Bay. The seven bay segments are indicated by the pie charts of relative species frequency over all 23 years. Note that the only instance of no species occurrence is *Capitella aciculata* in Lower Tampa Bay. Map created using QGIS Desktop 3.0.2 and Inkscape 0.92.3. The Tampa Bay shapefile was sourced from the Florida Geographic Data Library. 44

Figure 3.3: Violin plots of species abundance. The white points represent the median of each range. Abbreviations: C.cap=*Capitella capitata*; C.acic=*Capitella aciculata*; C.jon=*Capitella jonesi*; H.fili=*Heteromastus filiformis*; M.amb=*Mediomastus ambiseta*; M.cal=*Mediomastus californiensis*. 56

Figure 3.4: Violin plots of species abundance by Bay Segment, ordered by increasing average abundance. Lorenz curves of the cumulative proportion of averaged species abundance are overlaid. The different colored lines correspond to each species unique distribution and the black lines are a representation of a species with abundance equally distributed among Bay Segments. Note that “Species Abundance” scale varies between graphs. 59

Figure 3.5: Bubble plots of species abundance and LISA scores. Bubble plots are colored by species abundance (No./0.04m^2) ranges and size is a function of abundance for that sample. LISA plots are sorted by bay segment. Correlation significance was assessed at alpha = 0.05. Neighborhoods were defined by the five nearest neighbors. Interpretation: a significant high-high correlation indicates that the sample has a high value and is neighbored by other samples with high values. Raw figures generated with GeoDa (Anselin, Syabri, and Kho 2006) and further modified with Inkscape 0.92.3. 60

Figure 3.6: Plots of the slope and intercept for all models using values from Table A1. The axes are unitless as only slope and intercept were used for plot generation. However, the y-axis represents the observed count and x-axis the predicted count (Potts and Elith 2006). **Indicates that the biased, apparent values are used. 64

Figure 3.7: Plots of measures of correlation and error between observed and predicted values for all models using values from Table A1. Outlying values were not plotted for root mean square error and average error for visualization purposes. Excluded points are highlighted in Table A1. **Indicates that the biased, apparent values are used. 65

Figure 3.8: A table graph of term significance for all data-based models. There is a key at the top. If the term was significant (alpha = 0.05) for a given species/model combination, that block was colored. All colors correspond to those used for the models in Figures 7 and 8. Notice that the count and zero (presence/absence) parts of the Hurdle model are separated for all terms except Bay Segment. Significance of Bay Segment was assessed for the Hurdle model as a whole using a likelihood-ratio test. Bay Segment was not investigated for *Capitella aciculata*. The model chosen as best calibrated for each species is highlighted with a bold block. 66

Figure 3.9: Conditional random forest variable importance plots with species models organized in rows and environmental factors in columns. Variable importance represents the permutation importance of each factor, or the average effect on mean squared error when the factor is removed. Each factor’s importance is assessed independently, thus importance is not additive for each model. Higher importance equates to higher error when that factor is removed from the model. Importance is represented by size and color scales. 68

Figure 3.10: Accumulated Local Effects (ALE) plots of terms with top variable importance for each species. ALE scores are interpreted as the difference between the model prediction (at a given instance of x) and the averaged model prediction. Continuous predictors include a rug plot on the x-axis for visualizing observation density. Abbreviations of bay segments and term units are described in the Methods. 69

Figure 4.1: Illustration depicting the domain structure of hypoxia inducible factor with the basic helix-loop-helix, per-arnt-sim, n-terminal transactivation, and c-terminal transactivation domains labeled. Aryl receptor nuclear transferase has the same arrangement but without the transactivation domains. Illustration created with Illustrator for Biological Sequences. 90

Figure 4.2: Maximum likelihood phylogenetic tree of annotated bHLH genes (Graham and Presnell 2017) and extracted annelid and *Crassostrea gigas* sequences. Only support values for branches relevant to HIFα and ARNT identification are shown and they are in the format of SH-aLRT/UFBoot. The black branches are annelid sequences that did not have clear membership to one of the known clades (NPAS1/3, NPAS4, SIM, ARNTL(2), ARNT2). 97

Figure 4.3: Maximum likelihood phylogenetic tree of annelid HIFα genes. Branch support values are in the format of SH-aLRT/UFBoot. Presence of protein domains is indicated following the key in the top right corner. 99

Figure 4.4: Maximum likelihood phylogenetic tree of annelid ARNT genes. Branch support values are in the format of SH-aLRT/UFBoot. Presence of protein domains is indicated following the key in the top right corner. 100

Figure A.1: Plots of the relative influence of each term as determined with the Boosted Regression Trees. pH is the only term labeled for *Capitella aciculata* because every other term had zero influence. 120

LIST OF TABLES

Page

Table 2.1: All sampling localities, figure abbreviations, and GenBank accession numbers for samples collected by the authors in this study (italicized) and Hilliard et al. (2016). 13

Table 2.2: All clade names, geographic locations, and GenBank accession numbers for samples downloaded from GenBank. 18

Table 3.1: Number of samples collected by Bay Segment and Year. From 2007 onward, LTB+MTB and MR+TCB (samples boxed for both combinations) were treated as two, instead of four, reporting units for random sample selection. Additional samples were collected in 2010 in OTB for a special project (highlighted in grey). Decimal ratios were calculated by continuously dividing every number within a year by the number to its right; for a given year, the quotient of HB and OTB was divided by MTB, whose quotient was divided by LTB, and so on. A decimal ratio of 1 would indicate equal sampling effort across bay segments. Abbreviations follow those outlined in the Methods. 46

Table 3.2: Abundance (N), constancy (C), and dominance (D) overall and by Bay Segment per *Capitella, Heteromastus,* and *Mediomastus*. Abbreviations follow those outlined in the Methods. 54

Table 3.3: The variance and mean of each species abundance. 57

Table 3.4: The likelihood ratio test results from comparing the GAM-Poisson and GAM-Negative Binomial models for each species. Significance at α of 0.05 indicates a better fit of the GAM-Negative Binomial model. 57

Table 4.1: Species used in this study, NCBI SRA accession information for RNA Seq data, and whether HIFα and ARNT were found for them. 95

Table A.1: Model calibration metrics between observed values and model predicted values. Cells highlighted in blue indicate that those values are not present in the graphs of Figures 7 and 8. **Apparent (not corrected for optimism) statistics reported. 118

1. INTRODUCTION

Capitellid polychaetes (Annelida, Capitellidae) are commonly found in marine benthos. Compared to most polychaetes they have a reduced morphology (no appendages and rarely any branchiae), are generally red in color, and are deposit feeders. Morphological synapomorphies that define this family include a thoracic region that bears mostly capillary chaetae (Figure 1.1A&E) and an abdominal region bearing hooded hooks (Figure 1.1C&E). Genera and species of Capitellidae are differentiated by traits such as thoracic and abdominal region chaetal formulae, hooded hook morphology, and the presence/absence of other structures, e.g. genital hooks, anal cirri, and branchiae (Blake 2008). *Capitella* have nine thoracic chaetigers (Figure 1.1A-B) with genital hooks on the dorsum of chaetigers eight and nine (Figure 1.1D).

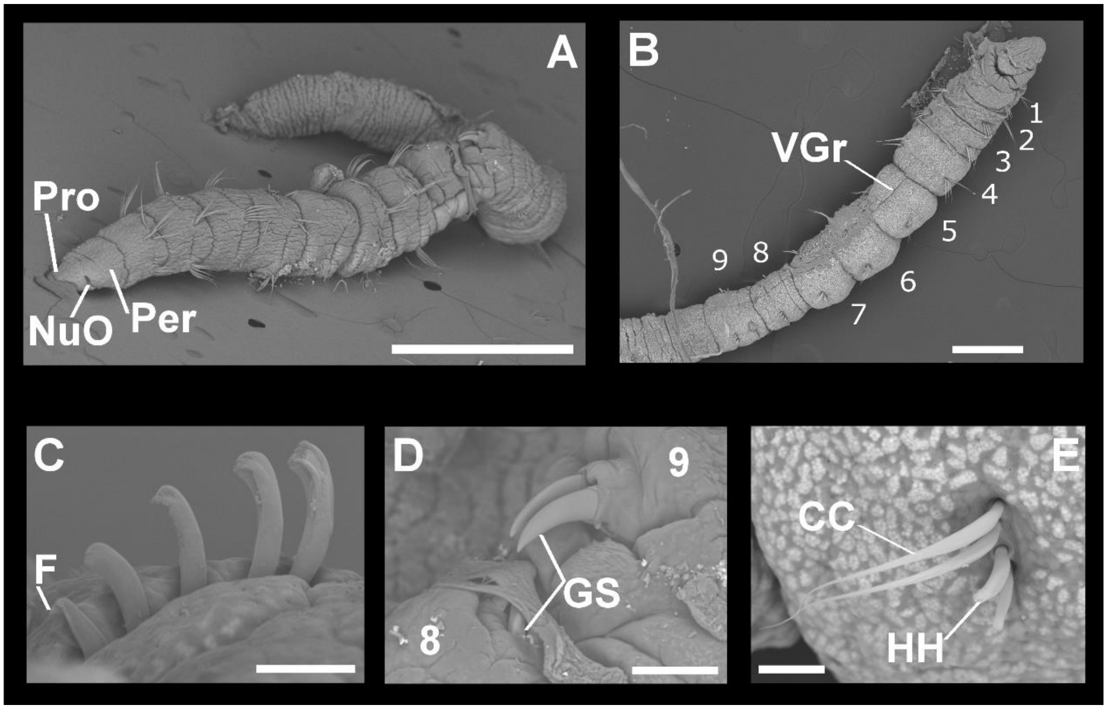


Figure 1.1: Modified from Hilliard et al. (2016). Scanning electron micrographs of *Capitella* cf. *capitata* from Galveston Bay, Texas. A. Fragmented specimen with presegmental structures labeled. B. Fragmented specimen with nine thoracic segments numbered and attention drawn to the ventral groove. C. Hooded hooks on right side of segment 11 of specimen shown in image A. D. Genital spines on segments 8 and 9 of specimen shown in image A. E. Mixed neurochaetae (capillary chaetae and hooded hooks) on left side of segment 7. CC, capillary chaeta; F, main fang; GS, genital spine; HH, hooded hook; NuO, nuchal organ; Per, peristomium; Pro, prostomium; VGr, ventral groove. Scale bars: A=500 µm; B=200 µm; C & E=15 µm; D=50 µm.

The family Capitellidae comprises 44 accepted genera (WoRMS 2020). Capitellids are of interest because they are often regarded as indicators of polluted and/or disturbed marine sediments where they often occur in high abundance. (Dean 2008). This is especially true for *Capitella,* the second largest genus with 31 accepted species (WoRMS 2020). Unfortunately, the extent to their use as such is questionable because their taxonomy and evolutionary history remain largely unresolved (Blake 2008). *Capitella capitata* was originally considered a cosmopolitan species but early work indicated at least six sibling species along the Massachusetts coast alone with few morphological differences but with distinct life histories, reproductive strategies, and allozyme signatures (Grassle and Grassle 1976). Since then, at least 50 sibling species have been documented worldwide (Méndez et al. 2000) based on physiological, reproductive, and developmental characteristics, among other traits (Eckelbarger & Grassle, 1983 & 1987; Gamenick et al. 1998; Grassle et al. 1987). With this is mind, *Capitella* is likely much larger and more diverse than officially indicated on WoRMS (2020). Interestingly, a recent phylogeny only supports monophyly for *Capitella* (Tomioka et al. 2018) and none of the other capitellid genera. These findings in combination with other species presumed cosmopolitan with large distributions highlight the need for a large revision of the family and continued work on species delimitation.

Capitella are found in many diverse habitats. They occur regularly in estuarine sediments, as evidenced by this dissertation. They have been found inhabiting wood and whale-bone falls in the deep-sea (Judge and Barry 2016, Silva et al. 2016). Some species have different sulfide tolerances, with some inhabiting sediment near shallow-water hydrothermal vents (Gamenick et al. 1998). A unique habitat expansion for *Capitella* is squid egg masses. *Capitella ovincola* is one such species that inhabits *Doryteuthis opalescens* egg masses (Zeidberg et al. 2011).

One of the originally detected cryptic species, *Capitella* sp. I (Grassle and Grassle 1976), has been formally described as *Capitella teleta* (Blake et al. 2009) and it is an important model of spiralian/lophotrochozoan development. Its genome has been sequenced (Simakov et al. 2012) and it has become a well-established model organism (Seaver 2016). Notable evo-devo studies include a comprehensive cell fate map (Meyer et al. 2010), annotation and mapping of Hox genes (Fröbius et al. 2008, de Jong and Seaver 2016) and ParaHox genes (Fröbius and Seaver 2006), and several studies on posterior regeneration (Özpolat and Bely 2016, de Jong and Seaver 2017).

Capitella karyotypes have a lot of variability with diploid numbers ranging 12-26 and differences in chromosome morphology between populations of a species (Grassle et al. 1987). This indicates that genome rearrangement and/or duplication events may have been involved in *Capitella* evolution. With such diversity in species and habitat, development of other *Capitella* species as comparative models would be insightful to understand the implications of genome rearrangement for speciation, habitat expansion, and the evolution of segmentation.

This dissertation contributes to the understanding of three aspects of capitellid biology: evolutionary history, ecology, and physiology. Specifically, I have focused on *Capitella* in the Gulf of Mexico (GoM). The first article addresses species delimitation and supports at least five new species from the GoM based on DNA sequence analysis. We also detected a species previously only known from Brazil in the GoM. Thoracic acicular spines were found in a well-supported monophyletic clade, indicating a single origin for this unique morphology.

The second article is focused on the ecology of six capitellids in Tampa Bay, Florida. All taxa have a zero-inflated abundance distribution and there is spatial autocorrelation by bay regions. Lorenz Curves were found to be an effective tool to assess spatial patterns of species abundance across large areas. Bay Segment, Depth, and Dissolved Oxygen were the most important environmental drivers. Modeling was accomplished by comparing six different approaches: GAMs (Poisson, Negative Binomial, Tweedie, and Zero-Inflated Poisson distributions), Hurdle models, and Boosted Regression Trees. There was no model evaluated with top performance for every species. However, GAM-Tweedie and Hurdle models performed well overall and may be useful for studies of other benthic marine invertebrates.

The final article addresses the evolution of the hypoxia inducible factor (HIF) transcription factor and its two subunits, HIFα and aryl receptor nuclear transferase (ARNT), or HIFβ, across Annelida. HIF is a key component of the cellular oxygen sensing pathway. While the recovered gene phylogenies do not directly approximate the known annelid phylogeny, some known groups are supported. Annotation of protein domains with Hidden Markov Model (HMM) profiles resulted in recovery of the known domains of each gene across Annelida but incompletely for each species. Updating HMMs to better reflect diversity within Annelida improved the results.

1.1. References

Blake JA (2008) Family Capitellidae Grube, 1862. In: Taxonomic atlas of the benthic fauna of the Santa Maria Basin and western Santa Barbara Channel, volume 7 – the Annelida part 4. Santa Barbara: Santa Barbara Museum of Natural History p 47-53

Blake JA, Grassle JP, and Ecklebarger KJ (2009) Capitella teleta, a new species designation for the opportunistic and experimental Capitella sp. I, with a review of the literature for confirmed records. Zoosymposia 2: 25-53

de Jong DM and Seaver EC (2016) A Stable Thoracic Hox Code and Epimorphosis Characterize Posterior Regeneration in Capitella teleta. PLOS ONE 11(2): e0149724 de Jong DM and Seaver EC (2017) Investigation into the cellular origins of posterior regeneration in the annelid Capitella teleta. Regeneration (Oxford, England) 5(1): 61-77

Dean HK (2008) The use of polychaetes (Annelida) as indicator species of marine pollution: a review. Revista de Biologia Tropical 56: 11-38

Eckelbarger KJ and Grassle JP (1983) Ultrastructural differences in the eggs and ovarian follicle cells of Capitella (Polychaeta) sibling species. The Biological Bulletin 165: 379-393

Eckelbarger KJ and Grassle JP (1987) Spermatogenesis, sperm storage and comparative sperm morphology in nine species of Capitella, Capitomastus, and Capitellides (Polychaeta: Capitellidae). Marine Biology 95: 415-429

Fröbius AC, Matus DQ, and Seaver EC (2008) Genomic organization and expression demonstrate spatial and temporal Hox gene colinearity in the lophotrochozoan Capitella sp. I. PLOS ONE 3(12): e4004

Fröbius AC and Seaver EC (2006) ParaHox gene expression in the polychaete annelid Capitella sp. I. Development Genes and Evolution 216: 81

Gamenick I, Vismann B, Grieshaber MK, and Giere O (1998) Ecophysiological differentiation of Capitella capitata (Polychaeta). Sibling species from different sulfidic habitats. Marine Ecology Progress Series 175: 155-166

Grassle JP and Grassle JF (1976) Sibling species in the marine pollution indicator Capitella (Polychaeta). Science 192(4239): 567-569

Grassle JP, Gelfman CE, and Mills SW (1987) Karyotypes of Capitella sibling species, and of several species in the related genera Capitellides and Capitomastus (Polychaeta). Bulletin of the Biological Society of Washington 7: 77-88

Hilliard J, Hajduk M, and Schulze A (2016) Species delineation in the Capitella species complex (Annelida: Capitellidae): geographic and genetic variation in the northern Gulf of Mexico. Invertebrate Biology 135(4): 415-422

Judge J and Barry JP (2016) Macroinvertebrate community assembly on deep-sea wood falls in Monterey Bay is strongly influenced by wood type. Ecology 97(11): 3031-3043

Méndez N, Linke-Gamenick I, and Forbes V (2000) Variability in reproductive mode and larval development within the Capitella capitata species-complex. Invertebrate Reproduction and Development 38: 131-142

Meyer NP, Boyle MJ, Martindale MQ, and Seaver EC (2010) A comprehensive fate map by intracellular injection of identified blastomeres in the marine polychaete Capitella teleta. EvoDevo 1(1): 8

Özpolat BD and Bely AE (2016) Developmental and molecular biology of annelid regeneration: a comparative review of recent studies. Current Opinion in Genetics & Development 40: 144-153

Seaver E (2016) Annelid models I: Capitella teleta. Current Opinion in Genetics & Development 39: 34-41

Silva CF, Shimabukuro M, Alfaro-Lucas JM, Fujiwara Y, Sumida PYG, and Amaral ACZ (2016) A new Capitella polychaete worm (Annelida:Capitellidae) living inside whale bones in the abyssal South Atlantic. Deep Sea Research Part I: Oceanographic Research Papers 108: 23-31

Simakov O, Marletaz F, Cho SJ, Edsinger-Gonzales E, Havlak P, Hellsten U, Kuo DH, Larsson T, Lv J, Arendt D, et al. (2013) Insights into bilaterian evolution from three spiralian genomes. Nature 493(7433) 526-531

Tomioka S, Kakui K, and Kajihara H (2018) Molecular Phylogeny of the Family Capitellidae (Annelida). Zoological Science 35(5): 436-445

WoRMS Editorial Board (2020). World Register of Marine Species. Available from http://www.marinespecies.org at VLIZ. Accessed 2020-03-18

Zeidberg LD, Isaac G, Widmer CL, Neumeister H, and Gilly WF (2011) Egg capsule hatch rate and incubation duration of the California market squid, Doryteuthis (= Loligo) opalescens: insights from laboratory manipulations. Marine Ecology 32: 468-479

2. CAPITELLA (ANNELIDA: CAPITELLIDAE) SPECIES IN THE GULF OF MEXICO: DELIMITATION, PHYLOGEOGRAPHIC PATTERNS, AND PHYLOGENY OF THE GENUS

2.1. Introduction

Capitellid polychaetes (Annelida: Capitellidae) are common members of the world's marine benthos, from shallow estuaries to the deep sea. *Capitella capitata* (Fabricius et al., 1780), the most commonly reported species, is frequently cited as a bioindicator (Dean, 2008). However, its utility as a sentinel species is questionable, since it is part of a large complex of morphologically similar but genetically, reproductively and developmentally distinct species (Grassle & Grassle, 1976). Méndez et al. (2000) characterized 50+ putative *Capitella* species on the basis of life history characteristics alone. Other differences among sibling species include physiology, karyotypes, chromosomal count, egg envelope ultrastructure, and mature sperm morphology (Eckelbarger & Grassle, 1983 & 1987; Gamenick et al., 1998; Grassle et al., 1987).

Recent attempts have been made to understand the relationships within this species complex globally with molecular sequence data, mostly using the mitochondrial cytochrome *c* oxidase subunit I (COI) gene (Hilliard et al., 2016; Livi et al., 2017; Man-Ki et al., 2018; Silva et al., 2017; Tomioka et al., 2016). According to the COI data, *Capitella teleta* (formerly known as *Capitella* sp. I, according to Grassle & Grassle 1976) is one of the most widespread species, occurring in Japan (Tomioka et al., 2016), South Korea (Man-Ki et al., 2018) as well as in the Mediterranean Sea around Italy (Livi et al., 2017).

New species have been described from the coast of Brazil (Silva et al., 2017) and speculated around Japan (Tomioka et al., 2016) and Italy (Livi et al., 2017).

Hilliard et al. (2016) found two morphotypes of *Capitella* in the northern Gulf of Mexico (GoM). Based on the presence of acicular spines, one of these matches the description of *C. aciculata* (Hartman 1959), while the second one, without acicular spines, fits the general description of *C. capitata*. However, COI sequence data could not confirm the separation of the two morphotypes as separate clades (Hilliard et al. 2016); instead they appear to constitute a single, morphologically plastic species. In addition, the sequence data suggest a genetic break between the eastern and western populations in the northern GoM.

To document developmental differences between the western and eastern GoM, we recently described differences in the early development and adult morphology between *Capitella* sp. from Tampa Bay, FL and Tamiahua Lagoon, Veracruz, Mexico (Méndez et al., 2019). Populations with acicular spines also occur along the Brazilian coast but were sufficiently genetically distinct from the GoM populations that Silva et al. (2017) described them as a new species, *Capitella neoaciculata*.

For the present study, we examined additional samples from the US and Mexican coasts of the GoM. Our objectives are 1) to determine molecular support for the Tampa Bay and Tamiahua Lagoon populations (Méndez et al. 2019) and identify possible geographic patterns, 2) to delimit species boundaries and re-evaluate the significance of acicular spines as diagnostic characters, and 3) to re-analyze the phylogeny of the genus *Capitella* using the newly generated data in conjunction with publicly available data.

2.2. Materials and Methods

2.2.1. Specimen Collection/Storage

Capitella specimens were collected from shallow mud/sand flats, canals, and the edge of red mangrove swamps along the GOM coast and in Miami, Florida (Figure 2.1 and Table 2.1). Sediment was collected with a shovel at a depth not exceeding 15 cm and sieved through a 0.5- or 1-mm sieve. Some specimens were hand picked off of the mesh and stored in a centrifuge tube with ambient water while others were later picked from retained material (stored in a plastic storage container with ambient water). Most adult *Capitella* spp. (identified by the presence of nine thoracic chaetigers and/or genital spines in the 8th and/or 9th chaetigers) were processed within several hours of collection by relaxing them in a sea water:7% magnesium chloride (or sulfate) (1:1) solution and then fixed and stored in 95% ethanol. Some specimens from Apollo Beach Preserve in Tampa Bay and from Tamiahua Lagoon in Veracruz were maintained in culture to document their early development (Méndez et al. 2019). Some of the cultured specimens were sequenced for this study as well.

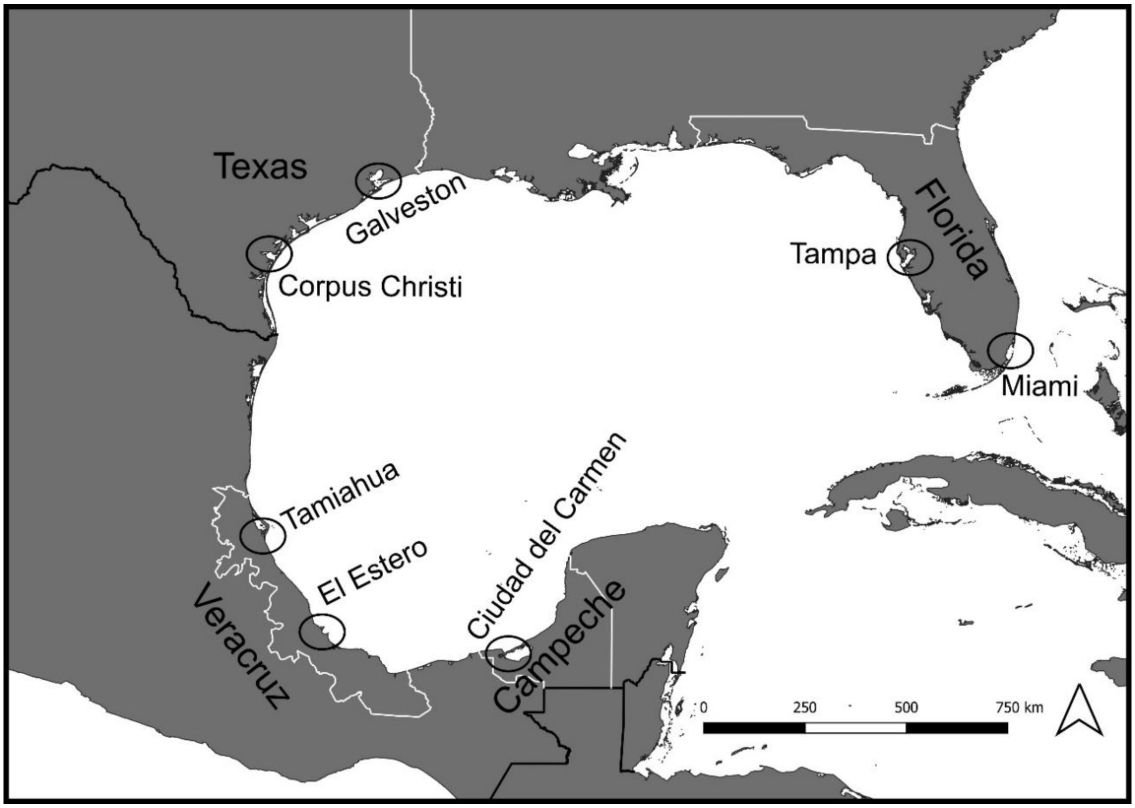


Figure 2.1: Sampling locations throughout the Gulf of Mexico. The shapefiles used to make this map are from version 2.3.7 of the Global Self-consistent, Hierarchical, High-resolution Geography (GSHHG) Database, accessible at http://www.ngdc.noaa.gov/mgg/shorelines/gshhs.html.

Table 2.1: All sampling localities, figure abbreviations, and GenBank accession numbers for samples collected by the authors in this study (italicized) and Hilliard et al. (2016).

Location			**Geographic coordinates**	**GenBank Accession**	
				COI	**H3**

TEXAS	Galveston Bay		29° 13’ 43.20” N
			94° 54’ 57.04” W
				KX961408 – KX961417;
				MT246752 – MT246770
					MT221461; MT221470;
					MT221471; MT221474;
					MT221481; MT221490;
					MT221515 – MT221518

	Christmas Bay		29° 2’ 56.21” N 94° 9’ 54.71” W	KX961404	–
	East Matagorda Bay		28° 45’ 37.87” N 95° 39’ 19.24” W	KX961405	*MT221535*
	Goose Island		28° 7’ 42.30” N 96° 59’ 23.41” W	KX961406	*MT221541*
	Corpus Christi Bay	Oso Bay	27° 43’ 5.05” N 97° 19’ 52.71” W	KX961403; KX961407	*MT221458; MT221542*
		Suter Park	27° 42' 16.4" N 97° 20' 04.1" W	*MT246688 – MT246710*	*MT221459; MT221460;* *MT221462 – MT221469*
		Fish Pass	27°41'41.3" N 97°11'02.0" W	–	*MT221472; MT221473*
FLORIDA	Tampa Bay	Admiral Farragut Academy	27° 46’ 40.90” N 82° 44’ 52.53” W	KX961418 – KX961426	–
		Apollo Beach Preserve	27° 47’ 34.00” N 82° 25’ 3.89” W	*MT246684 – MT246687;* *MT246737 – MT246751*	*MT221500 – MT221514*
		Crystal Beach, Saint Joseph Sound	28° 5’ 30.91” N 82° 46’ 52.84” W	KX961427	–

Table 2.1 Continued

	Location	Geographic coordinates	GenBank Accession	
			COI	H3
FLORIDA	Rookery Bay	26° 0’ 51.24” N 81° 45’ 14.28” W	KX961428 – KX961433	*MT221529 – MT221534*
	Matheson Hammock Park, Miami	25° 40' 21.5"N 80° 15' 37.6"W	*MT246678 – MT246683*	*MT221536 – MT221540*
MEXICO	Tamiahua, Veracruz	21° 16’ 11.05” N 97° 26’ 37.72” W	*MT246711 – MT246721*	*MT221475 – MT221480;* *MT221482 – MT221485*
	El Estero, Veracruz	19° 05' 31.3"N 96° 06' 11.8"W	*MT246722 – MT246729*	*MT221486 – MT221489;* *MT221491 – MT221493*
	Ciudad del Carmen, Campeche	18° 38' 54.7"N 91° 50' 24.8"W	*MT246730 – MT246736*	*MT221495 – MT221499*
	Mangrove/Marsh, Campeche	18°44'28.1"N 91°32'38.8"W	–	*MT221494*

2.2.2. DNA Purification and Sequencing

Tissue samples were isolated from each specimen by removing two or three segments from either the posterior end of a fragment or from the middle of the abdominal region in entire specimens. DNA was extracted and purified using the Qiagen DNeasy Blood and Tissue Kit following the manufacturer’s protocols but using half the specified amount of several reagents (Buffer AL, proteinase K, 100% EtOH, and Buffers AW1 and AW2) with final elutions of 100 µl. COI was amplified by polymerase chain reaction (PCR) using the thermocycler protocol of Carr et al. (2011) and two primer sets: polyLCO (5’-GAYTATWTTCAACAAATCATAAAGATATTGG-3’) and polyHCO (5’-TAMACTTCWGGGTGACCAAARAATCA-3’) (Carr et al., 2011); Mega-COIF (5’-

TAYTCWACWAAYCAYAAAGAYATTGG-3’) and Mega-COIR (5’-TAKACTTCTGGRTGMCCAAARAATC-3’) (Schult et al., 2016). Histone H3 (hereafter H3) was amplified using H3aF and H3aR (Colgan et al., 1998). The thermocycler protocol included one cycle at 95°C for 180 s; 35 cycles of denaturing at 95°C for 30 s, annealing at 54°C for 30 s, and extension at 72°C for 60 s; a final extension at 72°C for 300 s; and an indefinite hold at 12°C. All PCR amplifications were carried out with 25 µl reaction volumes consisting of 16.8 µl autoclaved Milli-Q® (EMD Millipore) water, 1.25 µl of each primer (10 µM), 0.2 µl Taq DNA Polymerase (Qiagen), 2 µl dNTP mix (10 mM) (Qiagen), 2.5 µl 10X Coral Load Buffer (Qiagen), and 1 µl of template DNA. Amplified products were diluted 1:9 in Milli-Q water.

Cycle sequencing was conducted with the BigDye® Direct Cycle Sequencing Kit (Life Technologies, now Thermo Fisher Scientific) using 1 µl Big Dye® Terminator v3.1 Cycle Sequencing RR-100, 2 µl Big Dye® Terminator v1.1/3.1 Sequencing Buffer (5X), 1 µl primer (10 µM), 1 µl template DNA, and 5 µl autoclaved Milli-Q® water. The thermocycler protocol of Hilliard et al. (2016) was followed. Reaction products were cleaned with a ZR DNA Sequencing Clean-Up Kit™ (Zymo Research) following the manufacturer’s protocols using half the designated volume of all reagents with final elution in 15 µl of Hi-di formamide (Thermo Fisher Scientific). Products were analyzed with an ABI 3130 Genetic Analyzer (Applied Biosystems, now Thermo Fisher Scientific). Resulting electropherograms and base calls were edited in Sequencher™ 4.8 (Gene Codes Corporation) by assembling the forward and reverse reads of each specimen. Final sequences were exported as FASTA files.

2.2.3. Sequence Analysis

Capitella spp. COI sequences were downloaded from GenBank (Table 2.2). *Notomastus profondus* (Eisig, 1887) and *Heteromastus filiformis* (Claparède, 1864) were used as outgroups in the COI analysis and a *Heteromastus* sp. was used as an outgroup in the H3 analysis (Table 2.2). Initial multiple sequence alignment by nucleotides was completed using the MUSCLE algorithm (Edgar, 2004) with default settings via MEGAX (Kumar et al., 2018). The alignments were further refined and finalized with MAFFT v7.428 (Katoh et al., 2002 & 2005; Katoh & Standley, 2013) using the global alignment algorithm with a maximum of 16 iterative refinement cycles. MEGAX was used for alignment visualization and adjustment. Primer regions were deleted. Model selection was completed using jModeltest 2.1.10 v20160303 (Darriba et al., 2012) with 88 candidate models (11 substitution schemes, equal/unequal base frequencies, with/without invariable sites, with/without rate variation among sites with four categories) and a BIONJ base tree for likelihood calculations. Phylogenetic analyses were performed using RaxML HPC (8.2.11) (Stamatakis, 2014) for maximum likelihood (ML) and MrBayes (3.2.6) (Ronquist et al., 2012) for Bayesian Inference (BI). Average genetic distances between and within populations were calculated using the K2P model for comparability to other studies (Kimura, 1980).

For COI, the ML analysis was completed with the GTR+CAT+I substitution model (Stamatakis 2006), rapid bootstrapping (Stamatakis et al., 2008) with a random seed of 333 and automatic bootstopping (Pattengale et al., 2009), a final search for the best-scoring tree, and *N. profondus* and *H. filiformis* as outgroups. The BI analysis was completed with the GTR+I+G substitution model (Tavaré, 1986), 7000000 total generations, a 1000-

generation sampling frequency, a 5000-generation diagnostic frequency, a 25% burn-in fraction, and a Metropolis coupling heating coefficient (λ) of 0.05. Median joining haplotype networks (Bandelt et al., 1999) were generated using PopART (Leigh & Bryant, 2015).

ML analysis of H3 was completed with the Jukes Cantor substitution model (Jukes & Cantor 1969), rapid bootstrapping (Stamatakis et al., 2008) with a random seed of 333 and automatic bootstopping (Pattengale et al., 2009), a final search for the best-scoring tree, and *Heteromastus* sp. as the outgroup. The BI analysis was completed with the Jukes Cantor substitution model (Jukes & Cantor, 1969), 9000000 total generations, a 1000-generation sampling frequency, a 5000-generation diagnostic frequency, and a 25% burn-in fraction.

Bayesian Poisson Tree Process (Zhang et al., 2013) was used to delimit species using the COI ML tree. A seed of 333 was set and Markov Chain Monte Carlo was run for 6000000 generations. The input tree was rerooted on the longest branch and the *Heteromastus filiformis* and *Notomastus profondus* were specified as outgroups.

Table 2.2: All clade names, geographic locations, and GenBank accession numbers for samples downloaded from GenBank.

Gene	Clade	Location	GenBank Accession
COI	**INDO-PACIFIC**	Indo-Pacific	KF737175; KF815717; JX676137; JX676150; JX676169; JX676171; JX676173; JX676174; JX676178; JX676179
	CANADA	Hudson Bay (Canada Lineage 1)	GU672406; GU672407; HQ023469; HQ023470
		Hudson Bay (Canada Lineage 2)	HQ023471 – HQ023474
	***Capitella aracaensis*, *C. biota*, *C. neoaciculata*, *C. nonatoi*, and *C. capitata* (Greenland)**	Brazil and Greenland	PopSet: 1043110615
	ITALY	Italy	PopSet: 1021313792
	***Capitella teleta* complex**	Japan, South Korea, and Italy (see above)	LC120630-LC120653; KX286328; KX286329; KX298243 – KX298247
	OUTGROUPS	Portugal	KR916855; KR916899
H3	**OUTGROUP**	Japan	LC208097

2.2.4. Morphological Observations

All specimens used for sequence analyses were morphologically identified with light microscopy. Presence/absence of acicular spines was diagnostic for assigning specimens to *C.* cf. *capitata* or *C.* cf. *aciculata*, respectively. Specimens from laboratory-maintained cultures of the Galveston (n=3), Corpus Christi (Suter Park) (n=4), and Tampa (Apollo Beach Preserve) (n=7) populations were used for scanning electron microscopy (SEM). This was done for comparison between populations and to *C. capitata* from Greenland, the type locality, and *C. teleta* (Blake, 2009; Blake et al., 2009). All cultures were started with individuals identified as *C.* cf. *capitata*. Individuals with *C.* cf. *aciculata* morphology were not included to prevent any confounding effects.

For scanning electron microscopy (SEM), all worms were relaxed in a 1:1 solution of sea water:7% magnesium sulfate for at least 30 minutes and stored at 4°C in 2.5% glutaraldehyde in Millonig's Phosphate Buffer (1 part 25% glutaraldehyde, 4 parts 0.34M NaCl, and 5 parts Millonig's Phosphate Buffer (0.46M sodium phosphate monobasic, pH 7.4)) until processing for SEM analysis. Each worm was rinsed (1 part Millonig's Phosphate Buffer and 1 part 0.6M NaCl) three times for five minutes. Samples were then rinsed briefly with distilled water followed by dehydration in an ethanol series (30% for 10 min, 50% for 10 min, 2 x 70% for 10 min, 2 x 80% for 10 min, 2 x 95% for 10 min, and 2 x 100% for 10 min). Samples were transferred to a dish, excess ethanol removed, and covered with hexamethyldisilazane and left for several minutes until dry. Worms were mounted on SEM stubs using carbon tabs, and sputter-coated with gold/palladium. SEM was conducted at Texas A&M University at Galveston, TX with a Hitachi TM3000 scanning electron microscope.

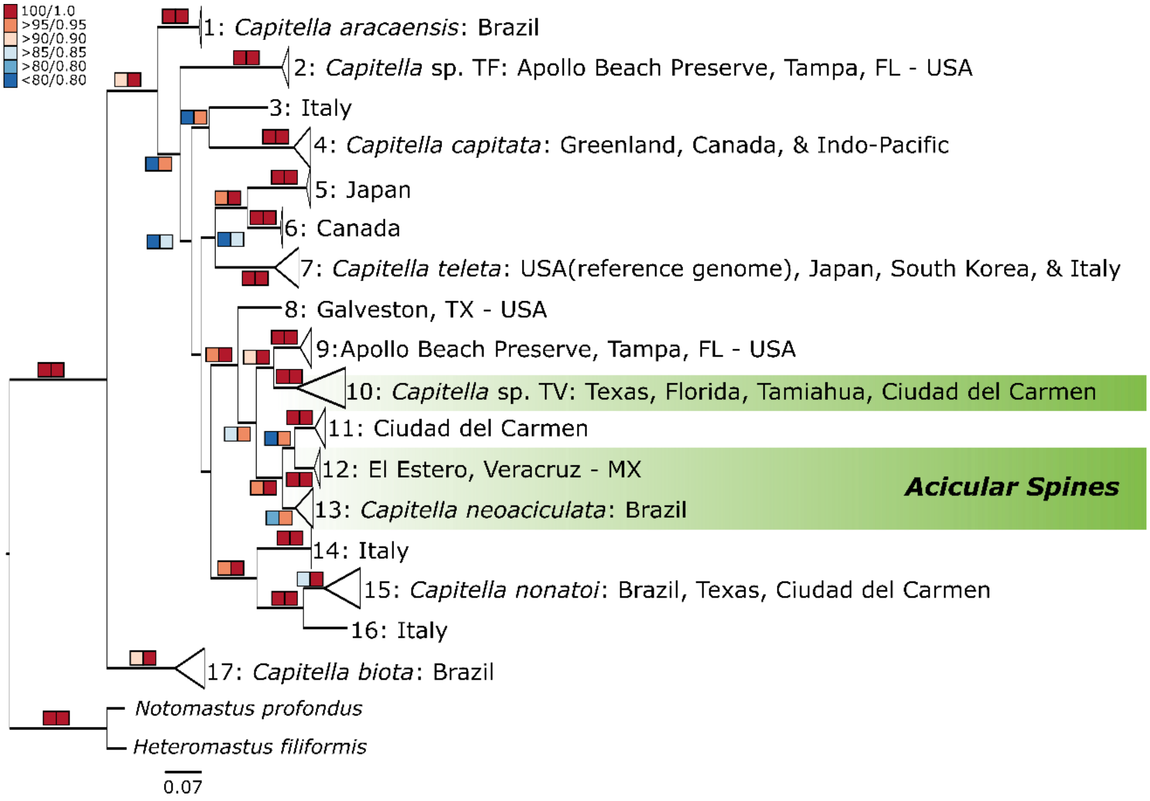
2.3. Results

2.3.1. COI Analysis

A total of 112 specimens were successfully sequenced for COI. Analysis of COI sequences revealed a total of 17 *Capitella* clades (Figure 2.2). Seven of the clades (2, 3, 8, 9, 11, 12, 16) are limited to a single location and three (3, 8, 16) only contain a single specimen. The remaining six encompass specimens from at least two locations. As clades 1, 3, 4, 5, 6, 7, 13, 14, 16, and 17 do not contain any sequences from the Gulf of Mexico, the geographic origins for these clades are listed in Figure 2.2 as countries, but the sequences originated from multiple locations in those countries.

The GoM specimens fall into seven different clades in the phylogenetic tree (clades 2, 8, 9, 10, 11, 12 and 15). Only one of them, clade 15, can be assigned to a named species, *Capitella nonatoi* (Silva et al. 2017), originally described from Araçá Bay, São Paulo, Brazil and here reported from Ciudad del Carmen, Campeche and Texas (Galveston and Suter Park). Specimens from Tamiahua, Veracruz form a clade with specimens from Ciudad del Carmen, Campeche; Texas; and Florida (Clade 10). This clade is referred to as *Capitella* sp. TV following Méndez et al. (2019). Some of the specimens from Apollo Beach Preserve, Tampa Bay culture form a distinct clade (Clade 2) and are referred to as *Capitella* sp. TF after Méndez et al. (2019). *Capitella* sp. TF has morphological differences from *Capitella* sp. TV (see below for more details). The remaining four clades with samples from the GoM cannot be linked to any previously reported species.

Acicular spines were found in three clades, 10, 12, and 13 (Figure 2.2, green boxes). Clades 10 and 13 contain specimens that identified as both *Capitella* cf. *capitata* and *Capitella* cf. *aciculata*. Clade 13 is a species recently described from Brazil (Silva et al. 2017) with all specimens having acicular spines.



100/1.0
>95/0.95
>90/0.90
>85/0.85
>80/0.80
<80/0.80
1: *Capitella aracaensis*: Brazil
2: *Capitella* sp. TF: Apollo Beach Preserve, Tampa, FL - USA
3: Italy
4: *Capitella capitata*: Greenland, Canada, & Indo-Pacific
5: Japan
6: Canada
7: *Capitella teleta*: USA(reference genome), Japan, South Korea, & Italy
8: Galveston, TX - USA
9:Apollo Beach Preserve, Tampa, FL - USA
10: *Capitella* sp. TV: Texas, Florida, Tamiahua, Ciudad del Carmen
11: Ciudad del Carmen
12: El Estero, Veracruz - MX
13: *Capitella neoaciculata*: Brazil
Acicular Spines
14: Italy
15: *Capitella nonatoi*: Brazil, Texas, Ciudad del Carmen
16: Italy
17: *Capitella biota*: Brazil
Notomastus profondus
Heteromastus filiformis
0.07

Figure 2.2: Maximum likelihood phylogenetic tree based on COI sequences. Branch labels are bootstrap support values|posterior probabilities. Details of each clade's geographic origin and species name if known are listed. The clades with species where acicular spines have been documented are indicated by green boxes.

Figure 2.3 shows that the *Capitella* sp. TV clade is split into Western and Eastern regions, with all Texas and Mexico samples in one group and all Florida samples in the other, respectively. The two groups are separated by three mutations. Specimens with acicular spines were found throughout (Figure 2.3). Within the Eastern Region are all of the samples from Miami and nine of the samples from Apollo Beach Preserve, including field collected and culture specimens. The remaining eight field-collected specimens from Apollo Beach Preserve form a distinct clade sister to *Capitella* sp. TV (Clade 9, Figure 2.2).

The *Capitella nonatoi* clade (Clade 15, Figure 2.2) includes specimens from Corpus Christi, Galveston and Ciudad del Carmen, in addition to the sequences from Brazil. Reconstruction of haplotype networks reveals geographic separation between the Gulf of Mexico and Brazil, with some haplotypes shared between Ciudad del Carmen and Corpus Christi (Figure 2.4). K2P genetic distances for COI between all populations are reported in Figure 2.5. Multiple samples were collected from Suter Park, Corpus Christi Bay (Jan 11, 2017 and May 29, 2017) and Galveston Bay (July 10, 2015 and May 23, 2018) Texas over five years, but *C. nonatoi* specimens were only found during May sampling. *Capitella* sp. TV was present during the other samplings as well as the other locations throughout both bays.

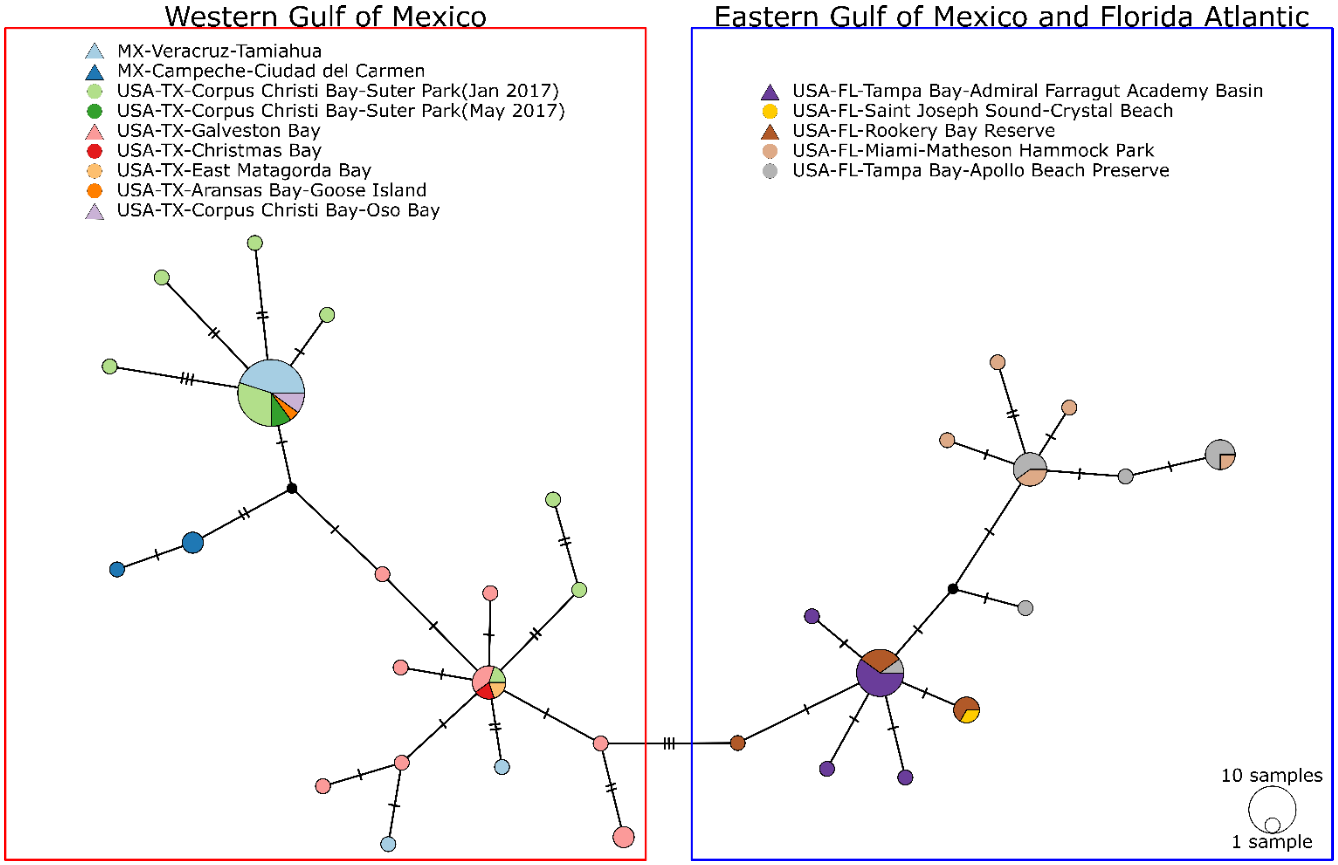

Western Gulf of Mexico
MX-Veracruz-Tamiahua
MX-Campeche-Ciudad del Carmen
USA-TX-Corpus Christi Bay-Suter Park(Jan 2017)
USA-TX-Corpus Christi Bay-Suter Park(May 2017)
USA-TX-Galveston Bay
USA-TX-Christmas Bay
USA-TX-East Matagorda Bay
USA-TX-Aransas Bay-Goose Island
USA-TX-Corpus Christi Bay-Oso Bay
Eastern Gulf of Mexico and Florida Atlantic
USA-FL-Tampa Bay-Admiral Farragut Academy Basin
USA-FL-Saint Joseph Sound-Crystal Beach
USA-FL-Rookery Bay Reserve
USA-FL-Miami-Matheson Hammock Park
USA-FL-Tampa Bay-Apollo Beach Preserve
10 samples
1 sample

Figure 2.3: Median joining haplotype network of COI for *Capitella* sp. TV (Clade 10, Figure 2.2). Locations marked by a triangle in the legends had some specimens with acicular spines. One hash represents a single mutation. Small black circles within the network represent a predicted haplotype that is not present in the data set.

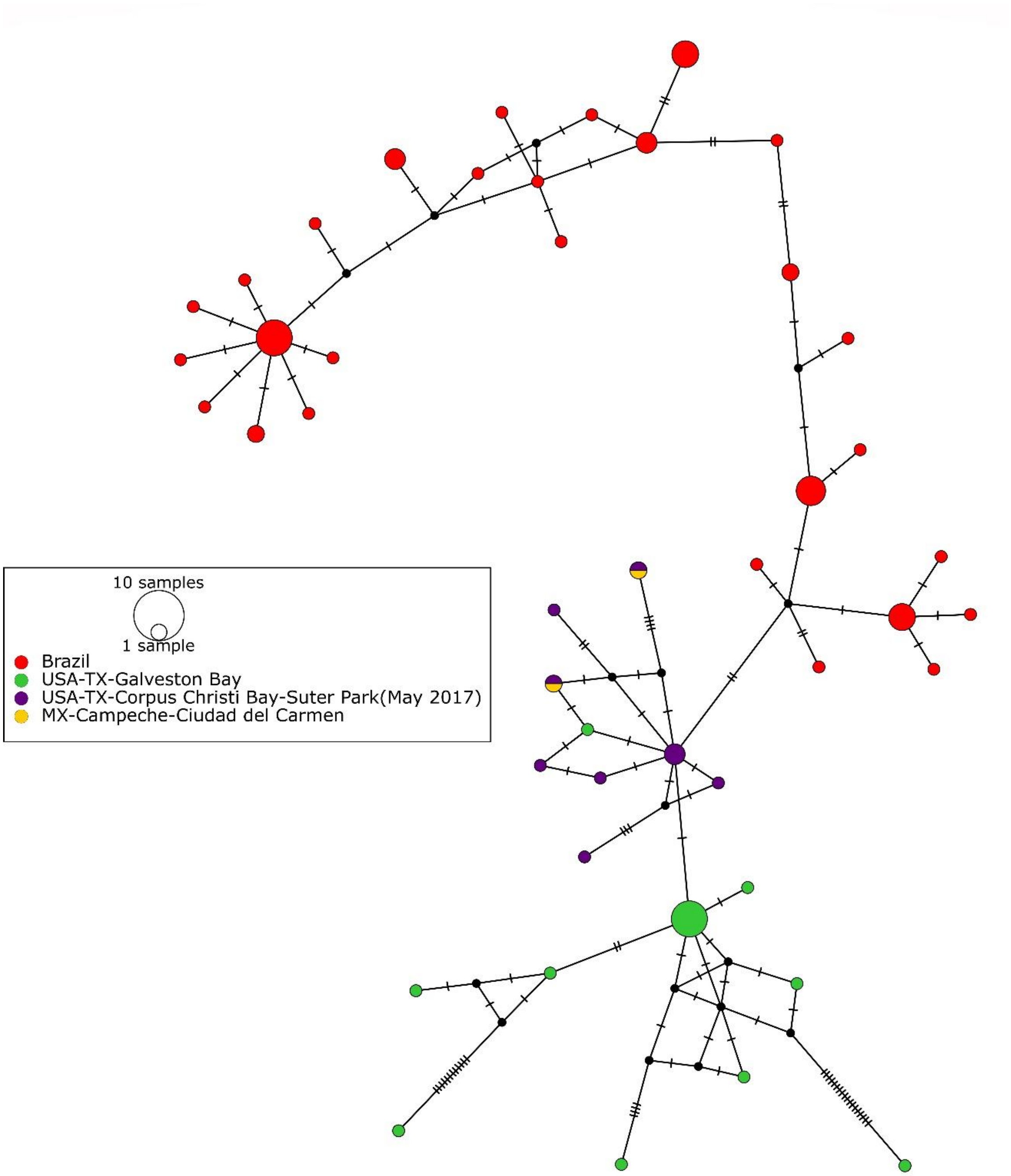
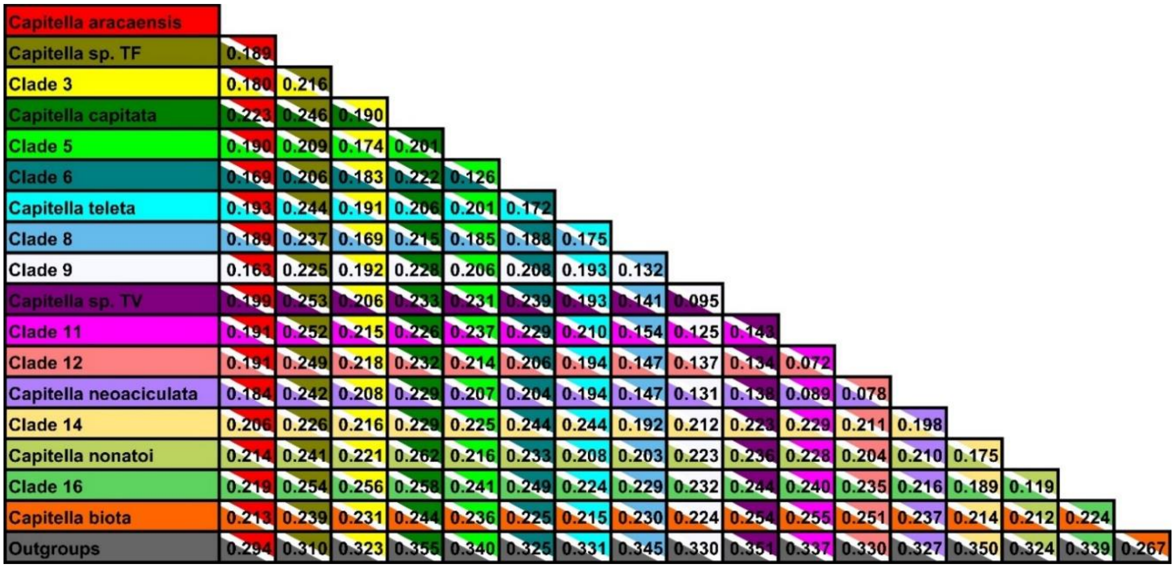

10 samples
1 sample
Brazil
USA-TX-Galveston Bay
USA-TX-Corpus Christi Bay-Suter Park(May 2017)
MX-Campeche-Ciudad del Carmen

Figure 2.4: Median joining haplotype network of COI *Capitella nonatoi* clade. One hash represents a single mutation. Small black circles within the network represent a predicted haplotype that is not present in the data set.

Our species delimitation analysis using bPTP clearly separated the 17 *Capitella* clades (Figure 2.2). While 13 of the clades were confirmed as species under both maximum likelihood and Bayesian posterior probability, four clades (Clades 11, 13, 14 and 17) were additionally split up into several species, depending on the criterion used. For Clade 11, both solutions delimited two species (each represented by a single specimen). Clade 13 contains a single species under the maximum likelihood criterion but eight species under Bayesian posterior probability. Clade 14 includes two species under maximum likelihood and forms a single species under posterior probability. Clade 17 (*Capitella biota*) was delimited as six species under maximum likelihood and as eight under posterior probability.



Capitella aracaensis																	
Capitella sp. TF	0.189																
Clade 3	0.180	0.216															
Capitella capitata	0.223	0.246	0.190														
Clade 5	0.190	0.209	0.174	0.201													
Clade 6	0.169	0.206	0.183	0.222	0.126												
Capitella teleta	0.193	0.244	0.191	0.206	0.201	0.172											
Clade 8	0.189	0.237	0.169	0.215	0.185	0.188	0.175										
Clade 9	0.163	0.225	0.192	0.228	0.206	0.208	0.193	0.132									
Capitella sp. TV	0.199	0.253	0.206	0.233	0.231	0.239	0.193	0.141	0.095								
Clade 11	0.191	0.252	0.215	0.226	0.237	0.229	0.210	0.154	0.125	0.143							
Clade 12	0.191	0.249	0.218	0.232	0.214	0.206	0.194	0.147	0.137	0.134	0.072						
Capitella neoaciculata	0.184	0.242	0.208	0.229	0.207	0.204	0.194	0.147	0.131	0.138	0.089	0.078					
Clade 14	0.206	0.226	0.216	0.229	0.225	0.244	0.244	0.192	0.212	0.223	0.229	0.211	0.198				
Capitella nonatoi	0.214	0.241	0.221	0.262	0.216	0.233	0.208	0.203	0.223	0.236	0.228	0.204	0.210	0.175			
Clade 16	0.219	0.254	0.256	0.258	0.241	0.249	0.224	0.229	0.232	0.244	0.240	0.235	0.216	0.189	0.119		
Capitella biota	0.213	0.239	0.231	0.244	0.236	0.225	0.215	0.230	0.224	0.254	0.255	0.251	0.237	0.214	0.212	0.224	
Outgroups	0.294	0.310	0.323	0.355	0.340	0.325	0.331	0.345	0.330	0.351	0.337	0.330	0.327	0.350	0.324	0.339	0.267

Figure 2.5: K2P genetic distances based on COI. Only distances between groups are shown and no within-group values are reported. The color scheme for the distance values can be interpreted as column = upper right and row = lower left. For example, the genetic distance between *Capitella teleta* and *Capitella aracaensis* is 0.193, or 19.3%.

2.3.2. H3 Analysis

Analysis of H3 did not reveal much, if any, geographic structure. The two Ciudad del Carmen specimens that clustered within *Capitella nonatoi* by COI analysis formed a well-supported clade with two specimens from Fish Pass, Corpus Christi, Texas (Figure 2.6). We did not successfully sequence COI for the Fish Pass specimens. Several of the samples from Apollo Beach Preserve formed supported clades, but the clade membership does not align with that of COI.

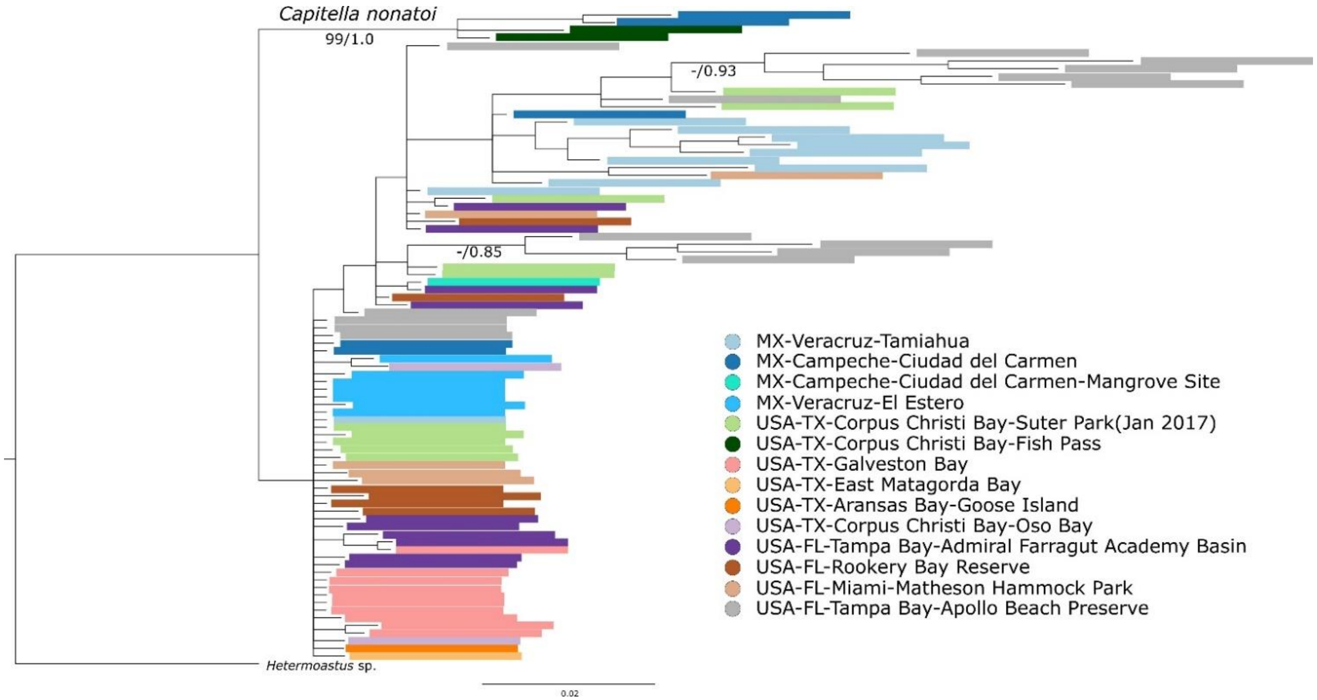

Capitella nonatoi
99/1.0
-/0.93
-/0.85
Hetermoastus sp.
0.02
MX-Veracruz-Tamiahua
MX-Campeche-Ciudad del Carmen
MX-Campeche-Ciudad del Carmen-Mangrove Site
MX-Veracruz-El Estero
USA-TX-Corpus Christi Bay-Suter Park(Jan 2017)
USA-TX-Corpus Christi Bay-Fish Pass
USA-TX-Galveston Bay
USA-TX-East Matagorda Bay
USA-TX-Aransas Bay-Goose Island
USA-TX-Corpus Christi Bay-Oso Bay
USA-FL-Tampa Bay-Admiral Farragut Academy Basin
USA-FL-Rookery Bay Reserve
USA-FL-Miami-Matheson Hammock Park
USA-FL-Tampa Bay-Apollo Beach Preserve

Figure 2.6: Bayesian Inference phylogenetic tree based on H3 sequences. Branch labels are bootstrap support values/posterior probabilities. Branches are only labeled as they are relevant to the Discussion Section.

2.3.3. Morphology

Scanning electron micrographs show that *Capitella* sp. TF (Figure 2.7 A-C) is not as long as *Capitella* sp. TV (Figure 2.7 D-F) in terms of total body length and thoracic length. *Capitella* sp. TF has dorsally incomplete intersegmental furrows in chaetigers 1-4 (Figure 2.7 B-C) whereas *Capitella* sp. TV has complete furrows (Figure 2.7 E-F). Thoracic chaetigers 1-4 of *Capitella* sp. TF also appear more inflated than those of *Capitella* sp. TV (Figure 2.7 B & E).

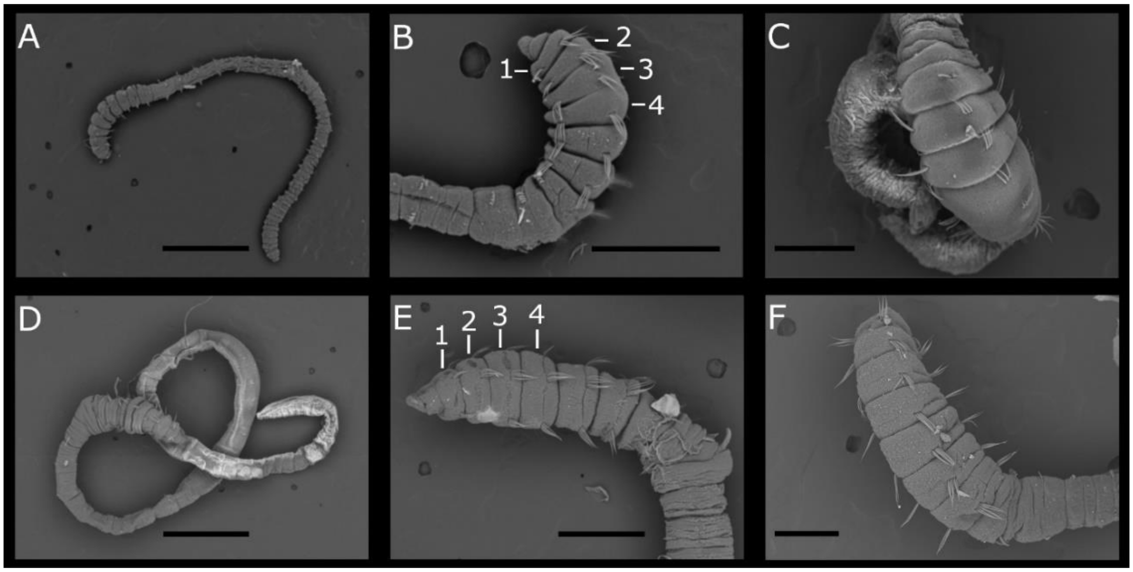


Figure 2.7: Scanning electron micrographs. *Capitella* sp. TF is in all top images (A-C). A. Complete specimen for total length. B. Thoracic region of specimen with chaetigers 1-4 labeled. C. Dorsal view for emphasis of incomplete intersegmental furrows. *Capitella* sp. TV is in all bottom images (D-F). D. Complete specimen from Galveston, TX for total length. E. Thoracic region of specimen from Tampa, FL with chaetigers 1-4 labeled. F. Dorsal view of specimen from Suter Park, Corpus Christi, TX for emphasis of complete intersegmental furrows. Scale bars: A&D=1 mm; B&E=500 µm; C&F=250 µm.

2.4. Discussion

Our results show that *Capitella* has high diversity along the GoM coast. This diversity remains largely undocumented, as the majority of specimens used in this study could not be assigned to any named species. Our phylogenetic analysis and the Bayesian Poisson Tree Process species delimitation supported at least five new *Capitella* spp. (Clades 2, 9, 10, 11, and 12 Figure 2.2). *Capitella* species are generally not localized to a particular location in the GoM and most locations host several species. Furthermore, the proportional contributions of different species at a particular location may be temporally variable, depending on season, environmental conditions, or random factors (Grassle and Grassle, 1976; Gamenick et al., 1998). For example, *Capitella nonatoi* was only collected in May 2017 and 2018 in Suter Park, Corpus Christi and Galveston Bay, respectively. In comparison, *Capitella* sp. TV was the only species found at these locations at other times of the year and it was found in Suter Park in May 2017.

The acicular spine morphology occurs in Clades 10, 12, and 13 (Figure 2.2, green boxes) which, with Clade 11, form a well-supported derived clade in our phylogeny, suggesting a single evolutionary origin of acicular spines. Within this acicular clade, only *Capitella neoaciculata* (Clade 13) is currently known from outside of the GoM (Brazil) and it is monophyletic with species from southern Veracruz and Campeche, Mexico. This would suggest that speciation within this clade may coincide with the opening of the GoM and Caribbean Sea and that this morphology evolved during the Jurassic Period (Ross and Scotese, 1988). Clitellates evolved as early as the Triassic Period (Manum et al., 1991) and the divergence of Capitellidae and Echiura from Clitellata (Weigert and Bleidorn, 2016; Helm et al., 2018) predates this. Further work using the fossil records of clitellates will be necessary to establish a molecular clock for Capitellidae and test this hypothesis. While some of these species' ranges may extend into the Caribbean, there are likely undiscovered *Capitella* species with acicular spines as well.

Capitella capitata was described from Greenland and has been redescribed with material from there (Blake, 2009). Blake (2009) hypothesized that *Capitella capitata* is widely distributed in Arctic and Subarctic waters and this is supported by the current analysis (Figure 2.2) and the original report of the *Capitella capitata* COI sequence from Greenland (Silva et al., 2017). Interestingly, there is support for samples from the Indo-Pacific region being true *Capitella capitata* (Figure 2.2). This is likely the result of an introduction by anthropogenic activity. *Capitella teleta* is also widely distributed and has been detected in the Mediterranean Sea and waters of South Korea and Japan (Figure 2.2). It was hypothesized that this, too, is due to anthropogenic activity (Tomioka et al., 2016). Continued sampling around the world will shed light on these patterns.

Our data clearly support *Capitella* sp. TV (Clade 10) and *Capitella* sp. TF (Clade 2) as separate species. Méndez et al. (2019) recently distinguished these two species by their larval development, based on samples collected from Tamiahua, Veracruz (abbreviated as TV) and Tampa Bay, FL (abbreviate as TF). In addition, they can be distinguished by their overall size (Figure 2.7) and pigmentation patterns in *Capitella* sp. TF (Méndez et al., 2019).

Three sympatric species occurred at Apollo Beach Preserve in Tampa Bay, Florida. At the time of collection, we did not detect any obvious morphological differences among the specimens and some of the larger specimens were preserved for morphological and molecular studies. The smaller specimens and the remaining larger specimens were kept in culture under the assumption that they all belonged to a single species. Our COI sequence analysis revealed that *Capitella* sp. TV was present both in the culture and the field-preserved samples. *Capitella* sp. TF, a smaller species, was only detected in the culture. A third species, comprising larger specimens (Clade 9, Figure 2.2), was only sequenced from the field-preserved material (although it is possible that some specimens are present in the culture but have not been sequenced). The three species may have different ecological characteristics and/or occupy different depths (Gamenick et al., 1998) or they may maintain sympatry through differing seasonal dynamics (Grassle and Grassle, 1976). However, our sampling design did not address this.

Capitella sp. TV from Miami were collected at Matheson Hammock Park, the type locality for *Capitella caribaeorum* (Warren & George, 1986). They are approximately the same size (*C. caribaeorum* < 20 mm, *Capitella* sp. TV ≤ 24 mm), but our specimens from this location were collected differently. While Warren & George (1986) described *C. caribaerum* from a culture stemming from intertidally collected decaying mangrove leaves, our specimens were retrieved by sieving sediment from sand flats with interspersed algal mats.

We further considered the possibility that our *Capitella* sp. TF (Clade 2), represents *Capitella caribaeorum*, but differences in size (mid-thoracic width: *Capitella caribaeorum*: 0.7 mm; Capitella sp. TF: ~0.25 mm) and life cycle make this unlikely as well. *Capitella caribaeorum* larvae metamorphose in the brood tube and emerge as juvenile worms (George, 1975) whereas *Capitella* sp. TF larvae emerge from brood tubes as swimming metatrochophores, settling to metamorphose after three days (Méndez et al., 2019).

H3 had limited utility for species delimitation in *Capitella*, as the phylogenetic analysis did not support most of the clades revealed in the COI analysis. The only species that was supported as monophyletic in the H3 analysis was *C. nonatoi*. The average K2P distance between *C. nonatoi* H3 and all other *Capitella* is 8.54%. Average K2P distances among Clades 9-12 range 0.517-11.5%, likely reflecting more recent speciation. Although H3 has been successfully used to delimit three cryptic *Heteromastus* species from Korea (Man-Ki et al., 2019), the marker may lack sufficient variation to resolve relationships among closely related *Capitella* species.

In conclusion, at least five undescribed species of *Capitella* were present in our samples as revealed by COI analysis. Two of them, *Capitella* sp. TV and *Capitella* sp. TF, also differ in their larval development. *Capitella* sp. TV is widespread throughout the GoM, ranging to at least the southern Atlantic coast of Florida, while the range of *Capitella* sp. TF still remains to be determined. Our phylogenetic tree supports a single origin of acicular spines. Our analyses included our own sequences in combination with publicly available sequences from multiple locations worldwide. Many parts of the world still remain unstudied with regard to *Capitella* diversity. Increased sampling efforts would lead to a more global picture of *Capitella* phylogeny and biogeography. Finally, multiple lines of evidence are useful when delimiting species. Culturing of *Capitella* is important to link developmental characteristics to the genetic signatures and morphology.

2.5. References

Bandelt H, Forster P, and Röhl A (1999) Median-Joining networks for inferring intraspecific phylogenies. Molecular Biology and Evolution 16(1): 37-48

Blake JA (2009) Redescription of Capitella capitata (Fabricius) from west Greenland and designation of a neotype (Polychaeta, Capitellidae). Zoosymposia 2: 55-80

Blake JA, Grassle JP, and Ecklebarger KJ (2009) Capitella teleta, a new species designation for the opportunistic and experimental Capitella sp. I, with a review of the literature for confirmed records. Zoosymposia 2: 25-53

Carr CM, Hardy SM, Brown TM, Macdonald TA, and Hebert PDN (2011) A tri-oceanic perspective: DNA barcoding reveals geographic structure and cryptic diversity in Canadian polychaetes. PLOS ONE 6(7): e22232

Claparède É (1864) Glanures zootomiques parmi les annélides de Port-Vendres (Pyrénées Orientales). Mémoires de la Société de Physique et d'Histoire Naturelle de Genève 17(2): 509-510

Colgan DJ, McLauchlan A, Wilson GDF, Livingston SP, Edgecombe GD, Macaranas J, et al. (1998) Histone H3 and U2 snRNA DNA sequences and arthropod molecular evolution. Australian Journal of Zoology 46, 419-437.

Darriba D, Taboada GL, Doallo R, and Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. Nature Methods 9(8): 772

Dean HK (2008) The use of polychaetes (Annelida) as indicator species of marine pollution: a review. Revista de Biologia Tropical 56: 11-38

Eckelbarger KJ and Grassle JP (1983) Ultrastructural differences in the eggs and ovarian follicle cells of Capitella (Polychaeta) sibling species. The Biological Bulletin 165: 379-393

Eckelbarger KJ and Grassle JP (1987) Spermatogenesis, sperm storage and comparative sperm morphology in nine species of Capitella, Capitomastus, and Capitellides (Polychaeta: Capitellidae). Marine Biology 95: 415-429

Eisig H (1887) Monographie der Capitelliden des Golfes von Neapel und der angrenzenden meeres-abschnitte nebst untersuchungen zur vergleichenden anatomie und physiologie. Fauna und Flora des Golfes von Neapel und der angrenzenden Meeres-Abschnitte, 16. Berlin: Verlag Von R. Friedländer & Sohn

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32(5): 1792-1797

Fabricius O, Berry SS, and Roper CFE (1780) Fauna Groenlandica, systematice sistens animalia groenlandiae occidentalis hactenus indagata, quoad nomen specificium, triviale, vernaculumque, synonyma auctorum plurimum, descriptionem, locum, victum, generationem, mores, usum capturamque singuli, pro ut detegendi occasio fuit, maximaque parte secundum proprias observationes. Hafniae [= Copenhagen] & Lipsiae [= Leipzig]: Ioannis Gottlob Rothe

Gamenick I, Vismann B, Grieshaber MK, and Giere O (1998) Ecophysiological differentiation of Capitella capitata (Polychaeta). Sibling species from different sulfidic habitats. Marine Ecology Progress Series 175: 155-166

George JD (1975) The culture of benthic polychaetes and harpacticoid copepods on agar. In Persoone G and Jaspers E (Eds.) Proceedings of 10th European Symposium on Marine Biology, Ostend, Belgium 1: 143-159

Grassle JP and Grassle JF (1976) Sibling species in the marine pollution indicator Capitella (Polychaeta). Science 192(4239): 567-569

Grassle JP, Gelfman CE, and Mills SW (1987) Karyotypes of Capitella sibling species, and of several species in the related genera Capitellides and Capitomastus (Polychaeta). Bulletin of the Biological Society of Washington 7: 77-88

Hartman O (1959) Capitellidae and Nereidae (marine annelids) from the Gulf side of Florida, with a review of freshwater Nereidae. Bulletin of Marine Science of the Gulf and Caribbean 9: 153-168

Helm C, Beckers P, Bartolomaeus T, Drukewitz, S.H., Kourtesis, I., Weigert, A., Purschke, G., Worsaae, K., Struck, T.H., and Bleidorn, C (2018) Convergent evolution of the ladder-like ventral nerve cord in Annelida. Frontiers in Zoology 15: 36

Hilliard J, Hajduk M, and Schulze A (2016) Species delineation in the Capitella species complex (Annelida: Capitellidae): geographic and genetic variation in the northern Gulf of Mexico. Invertebrate Biology 135(4): 415-422

Jukes TH and Cantor CR (1969) Evolution of protein molecules. In H.N. Munro Editor, Mammalian Protein Metabolism. Academic Press, New York p 21-132

Katoh K, Misawa K, Kuma K, and Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30(14): 3059-3066

Katoh K, Kuma K, Toh H, and Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Research 33(2): 511-518

Katoh K and Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution 30(4): 772-780

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16(2): 111-120

Kumar S, Stecher G, Li M, Knyaz C, and Tamura K (2018) MEGA X: Molecular evolutionary genetics analysis across computing platforms. Molecular Biology and Evolution 35(6): 1547-1549

Leigh JW and Bryant D (2015) popart: full-feature software for haplotype network construction. Methods in Ecology and Evolution 6: 1110-1116

Livi S, Tomassetti P, Vani D, and Marino G (2017) Genetic evidences of multiple phyletic lineages of Capitella capitata (Fabricius 1780) complex in the Mediterranean Region. Journal of Mediterranean Ecology 15: 5-11

Man-Ki J, Wi JH, and Suh H (2018) A reassessment of Capitella species (Polychaeta: Capitellidae) from Korean coastal waters, with morphological and molecular evidence. Marine Biodiversity 48(4): 1969-1978

Man-Ki J, Soh HY, and Suh H (2019) Three new species of Heteromastus (Annelida, Capitellidae) from Korean waters, with genetic evidence based on two gene markers. ZooKeys 869: 1-18

Méndez N, Linke-Gamenick I, and Forbes V (2000) Variability in reproductive mode and larval development within the Capitella capitata species-complex. Invertebrate Reproduction and Development 38: 131-142

Méndez N, Hilliard J, and Schulze A (2019) Early development of two Capitella species (Annelida: Capitellidae) from the Gulf of Mexico. Journal of the Marine Biological Association of the United Kingdom 1-12

Pattengale ND, Alipour M, Bininda-Emonds ORP, Moret BME, and Stamatakis A (2009) How many bootstrap replicates Are necessary? In S. Batzoglou Editor, Research in Computational Molecular Biology. Springer, Berlin & Heidelberg p 184-200

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, et al. (2012) MRBAYES 3.2: Efficient Bayesian phylogenetic inference and model selection across a large model space. Systematic Biology 61(3): 539-542

Ross MI, and Scotese CR (1988) A hierarchical tectonic model of the Gulf of Mexico and Caribbean region. Tectonophysics 155: 139-168

Schult N, Pittenger K, Davalos S, and McHugh D (2016) Phylogeographic analysis of invasive Asian earthworms (Amynthas) in the northeast United States. Invertebrate Biology 135(4): 314-327

Silva CF, Seixas VC, Barroso R, Di Domenico M, Amaral ACZ, and Paiva P (2017) Demystifying the Capitella capitata complex (Annelida, Capitellidae) diversity by morphological and molecular data along the Brazilian coast. PLOS ONE 12(5): e0177760

Stamatakis A (2006) Phylogenetic models of rate heterogeneity: a high performance computing perspective. Proceedings of IPDPS2006, HICOMB Workshop. Rhodos, Greece

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9): 1312-1313

Stamatakis A, Hoover P, and Rougemont J (2008) A rapid bootstrap algorithm for the RAxML web servers. Systematic Biology 57(5): 758-771

Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. In R.M. Miura Editor, Some Mathematical Questions in Biology – DNA Sequence Analysis. Providence: American Mathematical Society p 57-86

Tomioka S, Kondoh T, Sato-Okoshi W, Ito K, Kakui K, and Kajihara H (2016) Cosmopolitan or cryptic species? A case study of Capitella teleta (Annelida: Capitellidae). Zoological Science 33(5): 545-554

Warren LM and George JD (1986) Capitella caribaeorum sp. nov., a new capitellid polychaete from the Caribbean. Bulletin of the British Museum of Natural History 50(2): 117-125

Weigert A and Bleidorn C (2016) Current status of annelid phylogeny. Organisms Diversity & Evolution 16: 345-362

Zhang J, Kapli P, Pavlidis P, and Stamatakis A (2013) A general species delimitation method with applications to phylogenetic placements. Bioinformatics 29(22): 2869-2876

3. COMPARATIVE SPECIES ABUNDANCE MODELING OF CAPITELLIDAE (ANNELIDA) IN TAMPA BAY, FLORIDA

3.1. Introduction

Species abundance modeling is used in ecology for understanding biogeographic patterns and predicting impacts of climate change. These models are not to be confused with those commonly termed species distribution models (SDM) and ecological niche models (ENM) (Elith and Leathwick 2009, Sillero 2011, Peterson and Soberón 2012). While some of these methods use abundance data, they often use presence-only/-absence data and have prediction of species' distributions, and often maps, as an end-goal. Research using species abundance models for marine organisms has been focused on theoretical ecology, conservation planning, and climate change (Robinson et al. 2017). Taxonomic representation has been varied but biased toward vertebrates, with nearly 50% of studies focused on fish, birds, and mammals (Robinson et al. 2017).

Generalized linear models (GLMs) (Nelder & Wedderburn 1972, McCullagh & Nelder 1983) and generalized additive models (GAMs) (Hastie & Tibshirani 1986, 1990) are regression techniques frequently used for modeling species abundance data (Guisan et al. 2002). GLMs and GAMs are similar in that they both allow for non-Gaussian response distributions and use a monotonic function, often logarithmic, to link the response and predictors. The difference is that GAMs utilize smoothing functions on the predictors to determine their individual relationships with the response (Guisan et al. 2002, Zuur et al. 2009). When there is overdispersion due to zero-inflation, generalized linear modeling can be extended to a two-part, or hurdle, model (Cragg 1971). This approach first fits the abundance data as presence/absence with a binary response and then truncates the data and fits non-zero abundance with a Poisson or Negative Binomial response (Zuur et al. 2009).

A newer method being applied to abundance modeling is boosted regression trees (BRTs). This approach combines classification and regression trees (Breiman et al. 1984) with boosting algorithms (Freund & Schapire 1996). Some advantages include its ability to accommodate nonparametric data sets and fit complex interactions (De'ath 2007, Elith et al. 2008). For marine species abundance modeling, GLMs and GAMs were used 18% of the time while BRTs were used in only 4.2% of papers reviewed (Robinson et al. 2017). Hegel et al. (2010) provide a general overview of various other modeling strategies. There have been several comparisons of modeling strategies and while some target marine organisms (Connolly et al. 2009, Shelton et al. 2014), they are often focused on terrestrial organisms (e.g. Potts & Elith 2006, Baldridge et al. 2016) and vertebrates (Oppel et al. 2012). Such a comparative study of modeling strategies has not been completed for capitellid polychaetes.

Capitellids occur ubiquitously throughout the world's oceans. They have been reported from river mouths, estuaries, sea grass beds, deep sea sediments, and even wood and bones from whale falls in the deep sea (Judge & Barry 2016, Silva et al. 2016). This is especially the case for the best-known genus of the family, *Capitella*. Cryptic species of *C. capitata* were initially reported off the coast of Massachusetts primarily on the basis of life history characteristics and allozyme data (Grassle & Grassle 1976). Since then, there have been 50+ putative species described worldwide on the basis of life history alone (Méndez et al. 2000). Recent efforts have aimed to understand this species complex using the mitochondrial cytochrome *c* oxidase subunit I gene (COI) from the coasts of Brazil, Japan, Korea, Italy, and the Gulf of Mexico (Hilliard et al. 2016, Tomioka et al. 2016, Livi et al. 2017, Man-Ki et al. 2017, Silva et al. 2017). Some DNA barcoding with COI has been done on other genera (Carr et al. 2011, Lobo et al. 2016) and a phylogeny of the family indicates monophyly only for *Capitella* and a need to revise other genera (Tomioka et al. 2018).

Species in the *Capitella* species complex, as well as other capitellids, such as *Heteromastus filiformis, Mediomastus ambiseta,* and *M. californiensis,* have all shown utility as bioindicators (reviewed in Dean 2008). However, resolving species boundaries and understanding ecological drivers of abundance are necessary steps to effectively use them as such. All of these species, as well as *C. aciculata* and *C. jonesi,* occur throughout Tampa Bay, Florida. Tampa Bay is on the west central Florida coast (27°27'-28°3' N; 82°20'-82°44' W), in a biogeographic transition zone between the Northern Gulf of Mexico and Floridian ecoregions, creating a very diverse system (Yates & Greening 2011 and references therein, Spalding et al. 2007). Tampa Bay has an average depth of 4 m and surface area of nearly 1,036 km^2 (Morrison & Yates 2011). The shorelines are characterized by tidal flats and mangroves (Glick & Clough 2006).

The Environmental Protection Commission of Hillsborough County (EPCHC) has been continuously surveying the benthos of Tampa Bay since 1993. This dataset provides a unique opportunity for spatial modeling of benthic organisms in an estuarine system. We sought to conduct a meta-analysis using EPCHC data on *C. capitata* complex, *C. aciculata*, *C. jonesi*, *H. filiformis*, *M. ambiseta*, and *M. californiensis* (Figure 3.1) in this study. One goal of this study is to explore the data to understand spatial patterns inherent to each species. A second goal is to model environmental drivers of species abundance and ask whether there is one modeling strategy that works well for all species. Modeling is accomplished by comparing six different approaches: GAMs (Poisson, Negative Binomial, Tweedie, and Zero-Inflated Poisson distributions), Hurdle models, and BRTs. The third goal is to use Random Forest models (Breiman 2001), another classification and regression method, for environmental driver evaluation. These results can be used to inform future studies of benthic invertebrate spatial ecology in general as well as population genetics, speciation, phylogeography, and toxicology of capitellids in the Gulf of Mexico.

3.2. Materials and Methods

3.2.1. Data Collection

The EPCHC has been collecting benthic samples throughout the bay since 1993 following a program designed by the Tampa Bay Estuary Program to monitor large changes throughout the bay using robust randomized sampling (Squires et al. 1994). Tampa Bay is divided into the seven segments of Hillsborough Bay (HB), Boca Ciega Bay (BCB), Terra Ceia Bay (TCB), Manatee River (MR), Lower Tampa Bay (LTB), Middle Tampa Bay (MTB), and Old Tampa Bay (OTB) (Figure 3.2). Hexagon grids are overlaid to further divide regions. Smaller hexagons are used in smaller regions (HB, BCB, TCB and MR) to increase the number of samples. A number of hexagons are randomly selected for sampling each year (July – October) and a random point is generated within each hexagon. The same general strategy has been followed even though some aspects of the design (number of samples, reporting period, etc.) have changed over time.

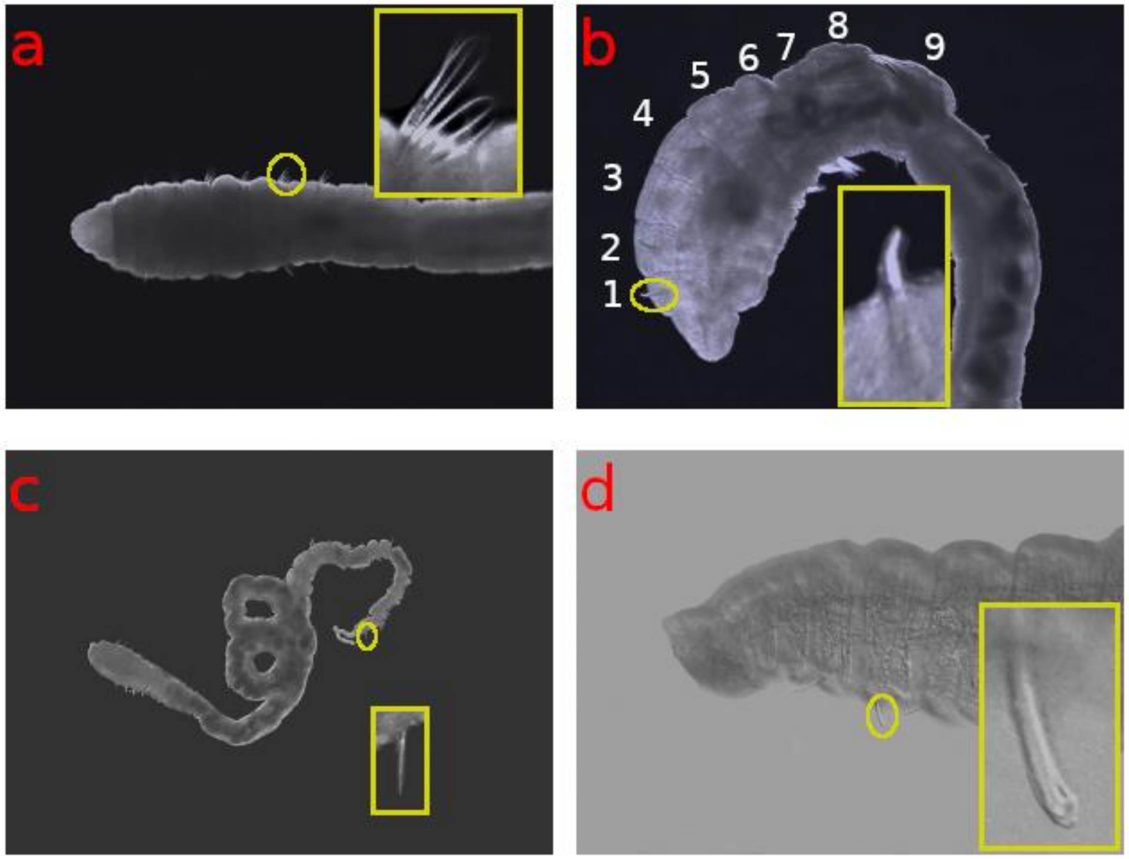


Figure 3.1: Light micrographs of Capitellidae species highlighting diagnostic characters. All specimens are from the Hillsborough County Environmental Protection Commission's samples and were photographed in-house. Capitellids are characterized by having thoracic and abdominal regions, with the number of thoracic chaetigers being the primary diagnostic for genera. The arrangement and type(s) of chaetae are also important. (a): *Capitella capitata* complex specimen at 4X magnification with capillary chaetae highlighted. (b): *Capitella aciculata* specimen at 40X magnification with acicular spine highlighted. Thoracic chaetigers 1-9 are labeled to differentiate thorax from abdomen. (c): *Mediomastus ambiseta* specimen at 4X magnification with abdominal capillary chaetae highlighted. (d): *Mediomastus californiensis* specimen at 40X with long abdominal hooded hooks highlighted.

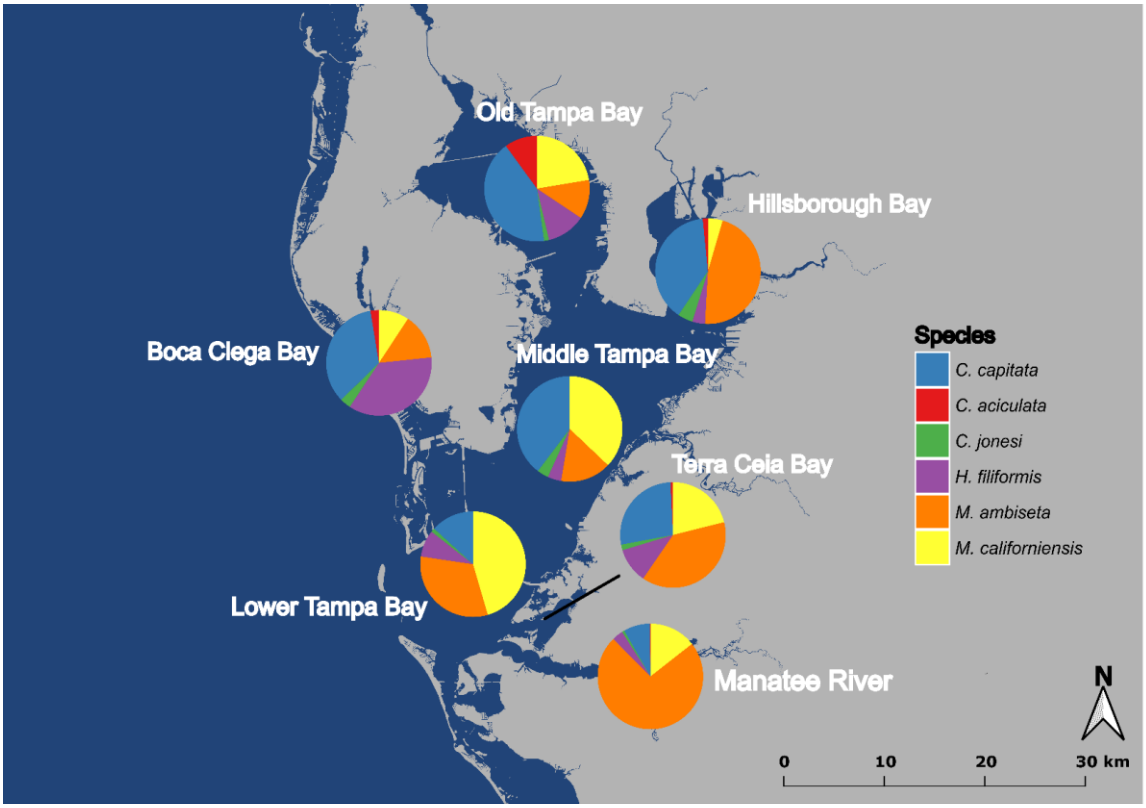


Figure 3.2: Map of Tampa Bay. The seven bay segments are indicated by the pie charts of relative species frequency over all 23 years. Note that the only instance of no species occurrence is *Capitella aciculata* in Lower Tampa Bay. Map created using QGIS Desktop 3.0.2 and Inkscape 0.92.3. The Tampa Bay shapefile was sourced from the Florida Geographic Data Library.

Prior to 2007 there was a very large sampling effort with up to 134 samples collected in 1995 (Table 3.1). However, the effort was not consistent with as few as 78 samples collected in 2006 (Table 3.1). Substantially fewer samples have been collected per year since 2007 but the sampling effort has become more consistent by bay segment and overall, with about 44 samples collected per year (Table 3.1). This is also evidenced by the sample decimal ratio, which increases and becomes more consistent from 2007 onward (Table 3.1).

From 2007 onward MR+TCB and LTB+MTB were treated as single reporting units by EPCHC for the random sample selection (Table 3.1). We did not combine these bay regions for modeling and kept them as separate bay segment categories for consistency and to avoid added complexity. Despite the change in reporting units, the number of samples for each bay segment remained relatively constant (Table 3.1).

Infaunal samples are collected with a single benthic grab using a Young Modified Van Veen grab (0.04 m2), sifted through a 500 micron mesh sieve, and bulk preserved. A 10% formalin solution was used for preservation prior to 2012 and NOTOXhisto™ (Scientific Device Laboratory) has been used since. After 72+ hours, samples are washed and transferred to 70% isopropanol for storage and identification. Surface and bottom water quality (pH, temperature, dissolved oxygen, and salinity) and depth are recorded at time of collection. A sample is also collected for calculation of the silt/clay fraction. More sampling design details, including a map of the hexagon grids, are available in EPCHC's Benthic Report (Karlen et al. 2015).

Table 3.1: Number of samples collected by Bay Segment and Year. From 2007 onward, LTB+MTB and MR+TCB (samples boxed for both combinations) were treated as two, instead of four, reporting units for random sample selection. Additional samples were collected in 2010 in OTB for a special project (highlighted in grey). Decimal ratios were calculated by continuously dividing every number within a year by the number to its right; for a given year, the quotient of HB and OTB was divided by MTB, whose quotient was divided by LTB, and so on. A decimal ratio of 1 would indicate equal sampling effort across bay segments. Abbreviations follow those outlined in the Methods.

	HB	OTB	MTB	LTB	MR	TCB	BCB	Bay-Wide	Decimal Ratio
1993	19	17	20	17	11	7	0	91	
1994	19	17	20	17	10	7	0	90	
1995	29	23	21	22	11	7	21	134	9.71E-07
1996	27	15	24	24	13	8	21	132	1.13E-06
1997	22	16	22	21	13	8	21	123	1.24E-06
1998	26	16	20	17	13	7	21	120	1.07E-06
1999	23	19	21	19	13	8	21	124	9.48E-07
2000	22	19	23	17	13	8	27	129	6.30E-07
2001	25	18	26	12	9	5	23	118	9.91E-07
2002	25	8	21	9	7	4	9	83	8.50E-06
2003	28	9	22	12	7	3	10	91	1.03E-05
2004	25	9	22	11	10	1	10	88	2.22E-05
2005	24	10	22	11	6	5	10	88	6.94E-06
2006	24	8	19	8	5	5	9	78	9.75E-06
2007	9	7	7	1	5	4	10	43	1.41E-03

Table 3.1 Continued

	HB	OTB	MTB	LTB	MR	TCB	BCB	Bay-Wide	Decimal Ratio
2008	9	7	5	3	6	3	11	44	1.28E-03
2009	9	7	6	2	5	4	11	44	1.28E-03
2010	9	22	5	3	5	4	11	59	4.08E-04
2011	9	7	5	3	7	2	11	44	1.28E-03
2012	9	7	5	3	7	2	11	44	1.28E-03
2013	9	6	5	3	6	3	11	43	1.50E-03
2014	9	7	5	3	5	4	11	44	1.28E-03
2015	9	7	5	3	6	3	11	44	1.28E-03

3.2.2. Database Acquisition and Filtering

The EPCHC maintains a Microsoft Access database on an FTP site (ftp://ftp.epchc.org/EPC_ERM_FTP/Benthic_Monitoring/). “EPC DataSubmittals_####.zip” is the relevant file and contains all data from 1993 through present. Data used in this study span 1993 to 2015 and were downloaded in December 2017. Species identifications were performed by different agencies using the most current identification literature at the time. The EPCHC is continually verifying identifications and posting data updates. Therefore, minor changes to our dataset have occurred since we completed the analyses and will continue to occur as specimens are re-examined.

The two relevant tables within the database are “Biology”, with records of species abundance data, and “DataSpreadsheet”, with all details related to field work and the measured environmental variables. These tables were joined and filtered for bottom measurements, year, hexagon ID, station number ID, latitude, longitude, and absolute abundance (Number/0.04m2) of *Capitella capitata* complex (hereafter *C. capitata*), *C. aciculata*, *C. jonesi*, *Heteromastus filiformis*, *Mediomastus ambiseta*, *M. californiensis*, and all other Capitellidae entries. The final data frame for modeling consisted of 1788 observations of the six listed species’ abundance data, latitude, longitude, bay segment, year, temperature (°C), salinity (psu), pH, dissolved oxygen (mg/L), silt clay fraction, and depth (meters).

3.2.3. Spatial Statistics

Unless otherwise stated, all analyses were performed with R, version 3.4.4 (R Core Team 2018) and all graphics were generated with base-R graphics and the package ggplot2 (Wickham 2009). Inkscape was used to further modify figures for final presentation.

Spatial patterns were explored for data description and to inform statistical model construction. Species constancy (presence in samples) and dominance (proportion of total abundance) were calculated (Carmo et al. 2013). Constancy is a percent calculated as: C = (p*100)/N, where p is the number of samples the species was present in and N is the total number of samples. Categories include C ≥ 50%, Constant; C = 25 – 50%, Accessory; and C ≤ 25%, Rare. Dominance is also a percent calculated as: D = (i/t)*100, where i is the species of interest's abundance and t is total abundance. Categories include D ≥ 10%, Eudominant; D = 5 – 10%, Dominant; D = 2 – 5%, Subdominant; D = 1 – 2%, Recessive; and D ≤ 1%, Rare. Pie charts of species dominance were plotted and overlaid onto a map of Tampa Bay to illustrate the seven regions relative species dominance (Figure 3.2). To assess abundance as a function of space, Lorenz curves (Lorenz 1905, Burt et al. 2009), a graphical way to assess equality, were manually fit and overlaid with violin plots for each species as a function of bay segment. Abundance was averaged by bay segment for Lorenz curves to account for unequal sample sizes. A table was generated to assess variation in sampling spatially and temporally. Local Indicator of Spatial Association (LISA), or the Local Moran's I (Anselin 1995), was calculated and plotted using the software GeoDa (Anselin et al. 2006). The K-Nearest Neighbors method was used to define neighborhoods with five neighbors (six total including the sample being considered). GeoDa was also used to generate bubble plots of species abundance.

Species abundance structures were assessed for overdispersion by comparing their mean and sample variances. When variance equals mean, a Poisson model is appropriate. Overdispersion is present when the variance is larger than the mean and indicates that another distribution may be more appropriate. This was also assessed by a likelihood-ratio test between Poisson and Negative Binomial models using the r-package lmtest, version 0.9-36 (Zeileis & Hothorn 2002). Zero-inflation of species abundance was assessed by visualization with violin plots using the r-package vioplot (Adler 2005).

3.2.4. Species Abundance Modeling

GAMs were fit for each species using all terms as covariates with r-package mgcv, version 1.8-23 (Wood 2011, Wood et al. 2016, Wood 2017). To not assume a linear relationship between each predictor and the response, a smoothing function was applied to all continuous covariates to generate a data-driven structure. Instead of using Year as a categorical term, the covariate "Total Samples/Year" was calculated and used to determine if species abundance is a function of sampling effort. The models took the form of: Species ~ s(Temperature) + s(Salinity) + s(pH) + s(Dissolved Oxygen) + s(Depth) + s(Silt Clay Fraction) + s(Total Samples/Year) + Bay Segment. All GAMs were fit using the "REML" smoothing method and a logarithmic link function. The only exception is that the Zero-Inflated Poisson distribution models had to be fit with an identity link function.

The Hurdle model was fit using r-package pscl, version 1.5.2 (Zeileis et al. 2008, Jackman 2017) with all covariates in the binomial and negative binomial parts of the model. Link functions used were logarithmic for the negative binomial model and logit for the binomial model. BRTs were fit by first using the 'gbm.step' function (available in the supplementary materials of Elith et al. 2008, used for this study, and the dismo r-package) to determine an optimal number of trees. A seed of 37 was set to permit reproducibility, a Poisson distribution was used for the response, tree complexity (or interaction depth) was set to one to not allow any interactions, bag fraction was 0.5, and learning rate (or shrinkage) was started at 0.01 and adjusted until an optimal tree count of at least 1,000 was reached (Elith et al. 2008). An exception was that *C. aciculata* was fit with a bag fraction of 1.0 due to algorithm convergence problems.

All models had static structure; all terms were used in every model with no term selection procedure or term interactions. This is because the goal was to assess model specification "out-of-the-box" and not refine any particular model to optimize its fit. Predictor significance was assessed at alpha of 0.05 for all models except for BRTs, for which the relative influence of each term is estimated (Friedman 2001, Friedman & Meulman 2003). Significance of bay segment in Hurdle models was assessed with a likelihood ratio test (r-package lmtest), resulting in a single significance value for the entire model.

Model evaluation and selection of a "best" model was completed by testing internal predictive performance, comparing the model's predicted values against the observed values. Statistics used were the Pearson correlation coefficient, Spearman rank correlation, root mean square error (RMSE), average (mean absolute) error (AVE), slope, and intercept (Potts & Elith 2006). We also followed the methods of Potts and Elith (2006) to correct the calibration statistics by estimating bias using the 0.632+ bootstrap method (Efron & Tibshirani 1997, Steyerberg et al. 2001) with 200 iterations. R-code was sourced from the online supplementary material of Zuur et al. (2009) and modified to shorten processing time.

A data set with as many zeros as *C. aciculata* (16 records of presence, data not shown), especially an entire bay segment with all zeros (LTB), presented unique problems. Algorithm convergence failures and matrix singularities were routinely encountered and required more exploration of model parameterization to resolve this issue. A consequence is that we could not use the 0.632+ bootstrap corrections and had to remove bay segment. Instead of removing bay segment from analyses, removing LTB samples would have also worked. We instead chose to keep LTB samples for their information in other cofactors. Another consequence is that stochasticity could not be included in the BRT and bag fraction had to be fixed at 1.0. These problems are not unrelated as bootstrapping for estimate correction and introducing stochasticity into the BRT both require subsampling the data frame. Limited presence records for *C. aciculata* results in a higher (1.05x10^-7) probability of a data frame with all zeros being built. In comparison, *C. capitata* has a 2.87x10^-141 probability of this happening.

3.2.5. Environmental Factors

Analyses for this section were performed with R, version 3.6.0 (R Core Team 2019). Random forest models were built to assess environmental drivers of species abundance independent of the model specification comparisons. We have allowed interactions in the random forest as our goal is not to compare this to the other model specifications but use it to understand environmental driver importance, and inclusion of interactions can aid this. As total samples/year is not an environmental term, it was not included in this analysis. We specifically used conditional random forests because they reduce variable-selection bias due to differences in variable types and structures (Strobl et al. 2009). Models were fit with r-package party, version 1.3-3 (Hothorn et al. 2006, Strobl et al. 2007,2008) and r-package caret, version 6.0-85 (Kuhn et al. 2020). Plots of variable importance were created.

In the interest of space, we chose to only assess the relationship between each species and its most important variable. Accumulated local effects (ALE) plots (Apley and

Zhu 2016) were generated using r-package iml, version 0.9.0 (Molnar et al. 2018). ALE plots average the effects of a factor on model predictions while holding all other factors constant. Changes in model predictions are averaged in windows, measuring local effects. This approach avoids data extrapolation that can lead to bias when predictors are correlated. See Molnar (2019) for further reading.

3.3. Results

3.3.1. Spatial Statistics

Constancy and dominance of the six species are variable throughout all bay segments. *Capitella capitata* was most dominant overall (D = 33.60%) (Table 3.2). *C. aciculata* was the least dominant overall at 2.72% (Table 3.2) and was not found at all in LTB. Its peak dominance was in OTB where it comprised 9.97% of abundance (Figure 3.2 and Table 3.2). *C. jonesi* also occurred in a small portion of the samples (C = 5.22%) but was found throughout the bay with dominance ranging 0.84 – 4.63% (Figure 3.2 and Table 3.2). *C. capitata* and *Mediomastus* spp. were the most constant throughout the bay and had dominance that ranged anywhere from 4.6 to 71.48% of species abundance (Figure 3.2 and Table 3.2). *Heteromastus filiformis* has dominance throughout the bay at 3.88 – 11.87% but is its most dominant in BCB (D = 36%) (Figure 3.2 and Table 3.2).

Table 3.2: Abundance (N), constancy (C), and dominance (D) overall and by Bay Segment per *Capitella*, *Heteromastus*, and *Mediomastus*. Abbreviations follow those outlined in the Methods.

	TOTAL			HB			OTB			MTB		
	N	C	D	N	C	D	N	C	D	N	C	D
C. capitata	2760	16.56	33.34	1082	20.54	38.99	618	18.08	42.13	392	9.50	39.60
C. aciculata	233	0.90	2.81	44	0.74	1.59	147	1.48	10.02	1	0.30	0.10
C. jonesi	250	5.32	3.02	129	6.44	4.65	23	5.90	1.57	35	5.34	3.54
H. filiformis	949	8.45	11.46	108	7.43	3.89	175	11.07	11.93	41	5.93	4.14
M. ambiseta	2706	13.21	32.69	1286	12.62	46.34	176	12.92	12.00	154	6.23	15.56
M. californiensis	1381	11.81	16.68	126	3.22	4.54	328	9.59	22.36	367	13.35	37.07

Table 3.2 Continued

LTB			MR		
N	C	D	N	C	D
66	10.05	13.23	65	9.47	8.16
0	0.00	0.00	1	0.59	0.13
6	2.28	1.20	6	1.18	0.75
41	3.20	8.22	27	7.10	3.39
159	10.05	31.86	583	23.67	73.15
227	23.74	45.49	115	11.24	14.43

TCB			BCB		
N	C	D	N	C	D
111	18.48	27.21	426	26.10	31.72
3	1.09	0.74	37	2.03	2.76
7	5.43	1.72	44	7.80	3.28
44	9.78	10.78	513	14.58	38.20
157	20.65	38.48	191	16.27	14.22
86	18.48	21.08	132	13.22	9.83

Species abundance distributions indicate that there is zero-inflation for all species (Figure 3.3). There is also evidence of statistical overdispersion, indicating that distributions other than Poisson are appropriate (Table 3.3). Comparing GAM-Poisson and GAM-Negative Binomial models with a likelihood ratio test indicates that a Negative Binomial distribution describes all species better (Table 3.4). These results led to use of a Negative Binomial distribution for the Hurdle model.

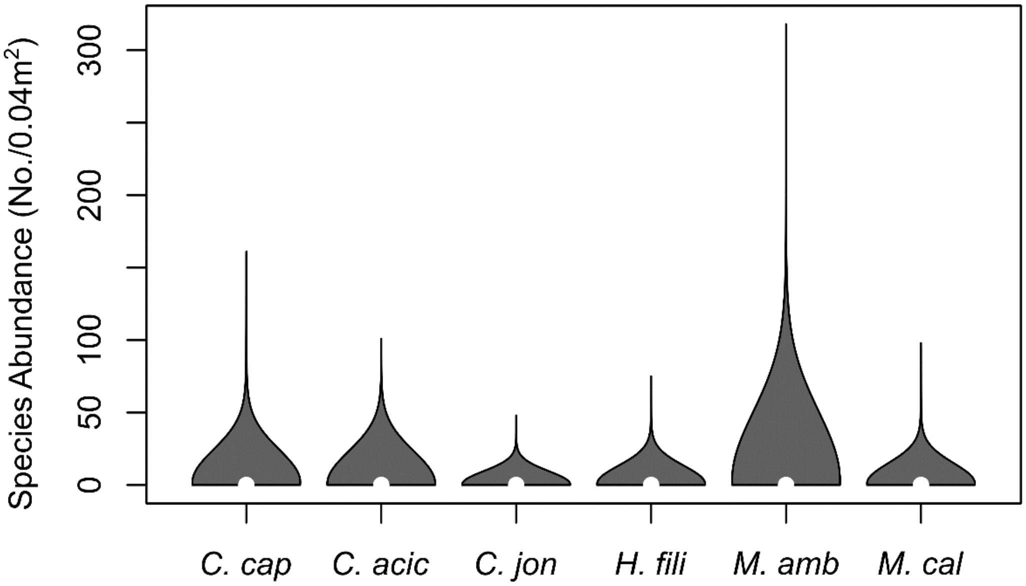


Figure 3.3: Violin plots of species abundance. The white points represent the median of each range. Abbreviations: C.cap=*Capitella capitata*; C.acic=*Capitella aciculata*; C.jon=*Capitella jonesi*; H.fili=*Heteromastus filiformis*; M.amb=*Mediomastus ambiseta*; M.cal=*Mediomastus californiensis*.

Table 3.3: The variance and mean of each species abundance.

	Variance	Mean
C. capitata	65.13	1.54
C. aciculata	7.38	0.13
C. jonesi	2.15	0.13
H. filiformis	13.01	0.53
M. ambiseta	126.45	1.51
M. californiensis	21.47	0.77

Table 3.4: The likelihood ratio test results from comparing the GAM-Poisson and GAM-Negative Binomial models for each species. Significance at α of 0.05 indicates a better fit of the GAM-Negative Binomial model.

	Chi-Square	Degrees of Freedom	p-value
C. capitata	5970.7	33.564	<2.2E-16
C. aciculata	53.749	35.572	0.02885
C. jonesi	363.24	38.151	<2.2E-16
H. filiformis	1337.5	34.464	<2.2E-16
M. ambiseta	5387.6	36.59	<2.2E-16
M. californiensis	3972.6	36.034	<2.2E-16

The Lorenz curves illustrate zero inflation and spatial autocorrelation (Figure 3.4). The violin plots for every species have a similar shape to those in Figure 3.3, indicating zero inflation. Spatial autocorrelation is indicated by steep and changing slopes in the Lorenz curves. For example, the violin plot for *H. filiformis* (Figure 3.4d) at BCB shows a very large abundance and the steep slope between LTB and BCB indicates that BCB has a large portion, or unequal share, of all *H. filiformis* abundance in Tampa Bay. A contrasting example is *M. californiensis* (Figure 3.4f) whose violin plots appear more equal and Lorenz curve is closer to the line of equal distribution. However, all species show some degree of spatial autocorrelation.

Looking at bubble plots of abundance and LISA plots, organized by bay segment (Figure 3.5), there is evidence of spatial autocorrelation. For example, LISA plots show that areas of species' presence often result in significantly autocorrelated neighborhoods (Figure 3.5). There is also a pattern of species' presence near-shore with few occurrences in the open-water areas of the bay (Figure 3.5).

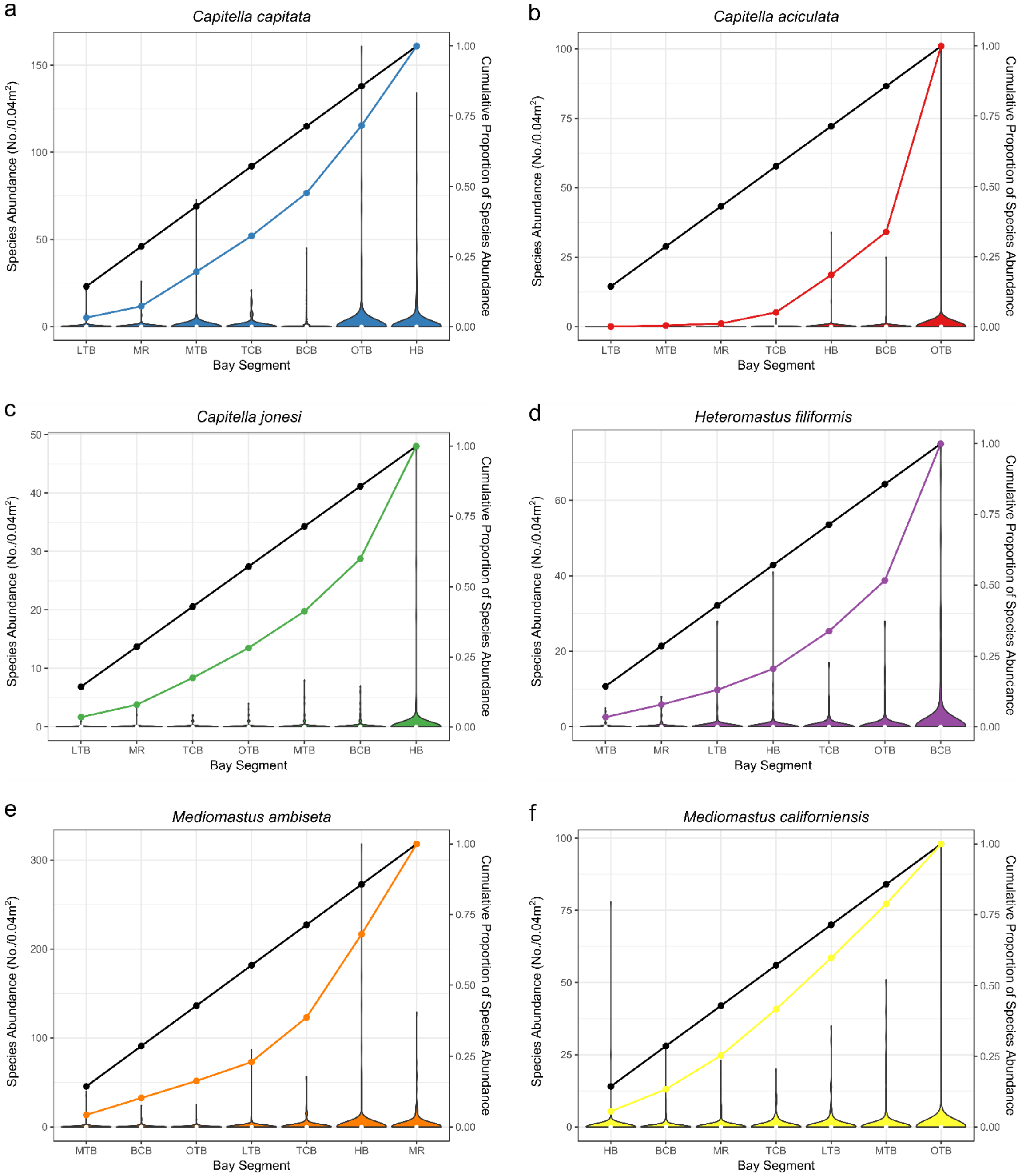


Figure 3.4: Violin plots of species abundance by Bay Segment, ordered by increasing average abundance. Lorenz curves of the cumulative proportion of averaged species abundance are overlaid. The different colored lines correspond to each species unique distribution and the black lines are a representation of a species with abundance equally distributed among Bay Segments. Note that "Species Abundance" scale varies between graphs.

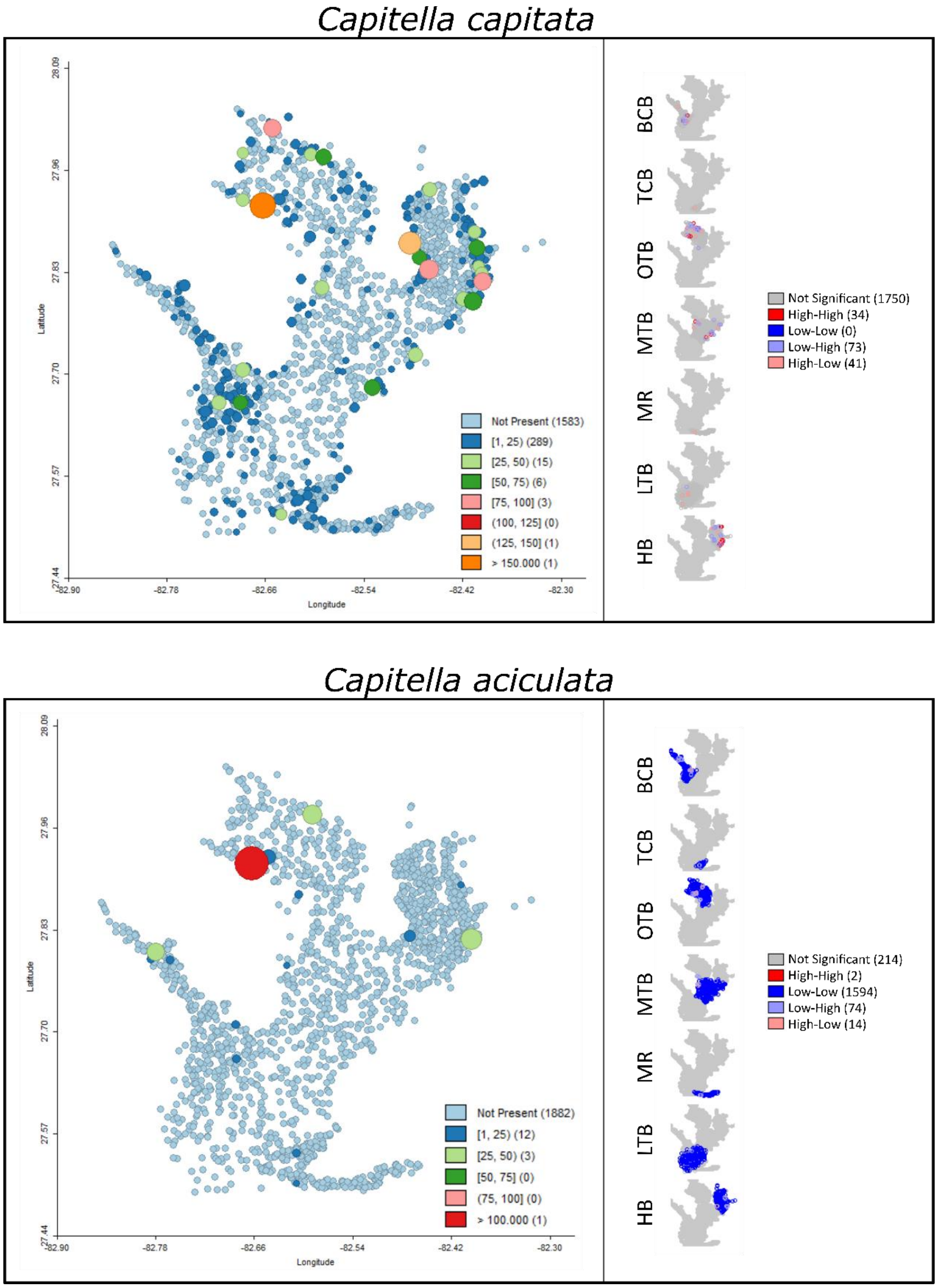


Figure 3.5: Bubble plots of species abundance and LISA scores. Bubble plots are colored by species abundance (No./0.04m^2) ranges and size is a function of abundance for that sample. LISA plots are sorted by bay segment. Correlation significance was assessed at alpha = 0.05. Neighborhoods were defined by the five nearest neighbors. Interpretation: a significant high-high correlation indicates that the sample has a high value and is neighbored by other samples with high values. Raw figures generated with GeoDa (Anselin, Syabri, and Kho 2006) and further modified with Inkscape 0.92.3.

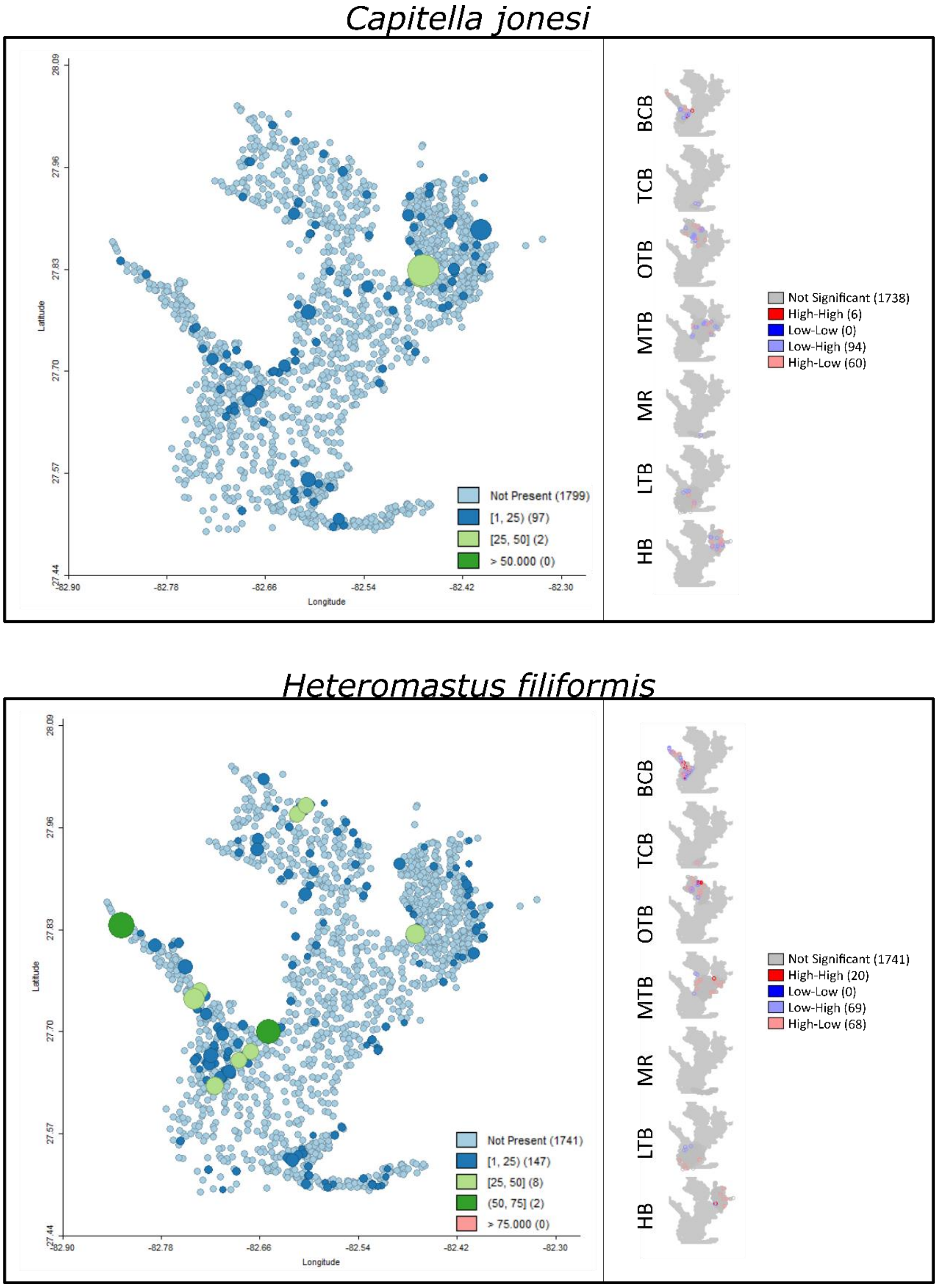
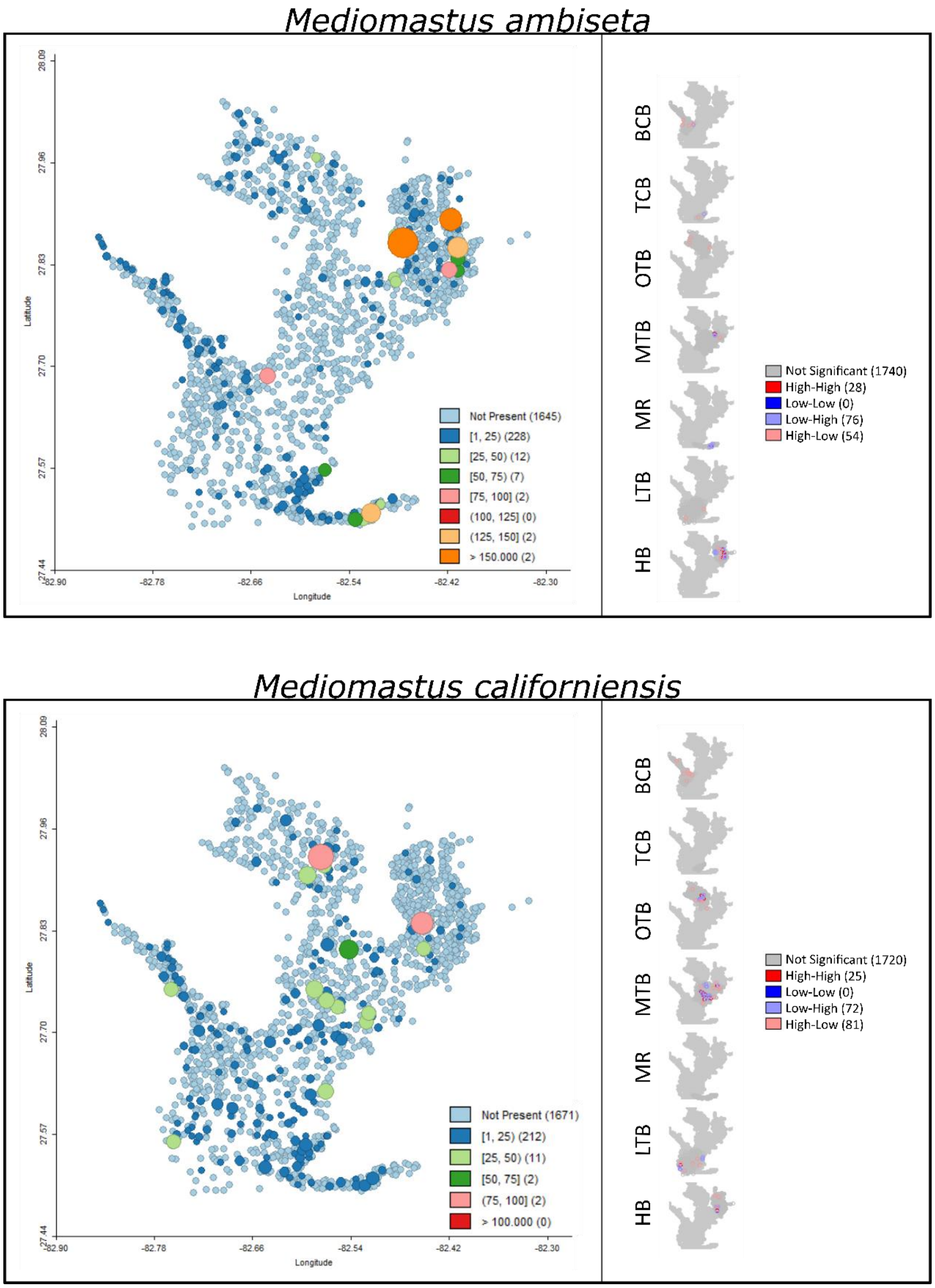


Figure 3.5 Continued



Mediomastus ambiseta

Mediomastus californiensis

Figure 3.5 Continued

3.3.2. Species Abundance Modeling

The 0.632+ bootstrap to correct model optimism was applied to every species except *C. aciculata*. A problem with matrix singularity was encountered during the 0.632+ bootstrap processes and apparent (non-adjusted) statistics of the model fits are presented for this species. Model calibration can be assessed with the slope and intercept (observed count ~ predicted count) (Figure 3.6, Table A.1) (Potts & Elith 2006). It is clear that there is a lot of variation within and between species. All models had a bias (intercept) within ±1.0 except BRT, *C. aciculata*; GAM-Zero-Inflated Poisson, *C. capitata*; and Hurdle, *M. ambiseta*. The consistency/spread (slope) was more variable (Figure 3.6, Table A.1). The models considered best calibrated for each species are: Hurdle, *C. capitata* (m = 0.95, b = 0.04); GAM-Poisson, *C. aciculata* (m = 1.01, b = 0.00); GAM-Tweedie, *C. jonesi* (m = 0.97, b = 0.01); Hurdle, *H. filiformis* (m = 1.01, b = -0.01); GAM- Negative Binomial, *M. ambiseta* (m = 0.98, b = 0.26); and GAM-Tweedie, *M. californiensis* (m = 0.94, b = 0.07) (Figure 3.6, Table A.1).This calibration is reflected in both correlation values, with the best calibrated models generally having the highest, or near highest, values (Figure 3.7, Table A.1). It is also corroborated by the RMSE and AVE values, with selected models having the lowest or relatively low values (Figure 3.7, Table A.1).

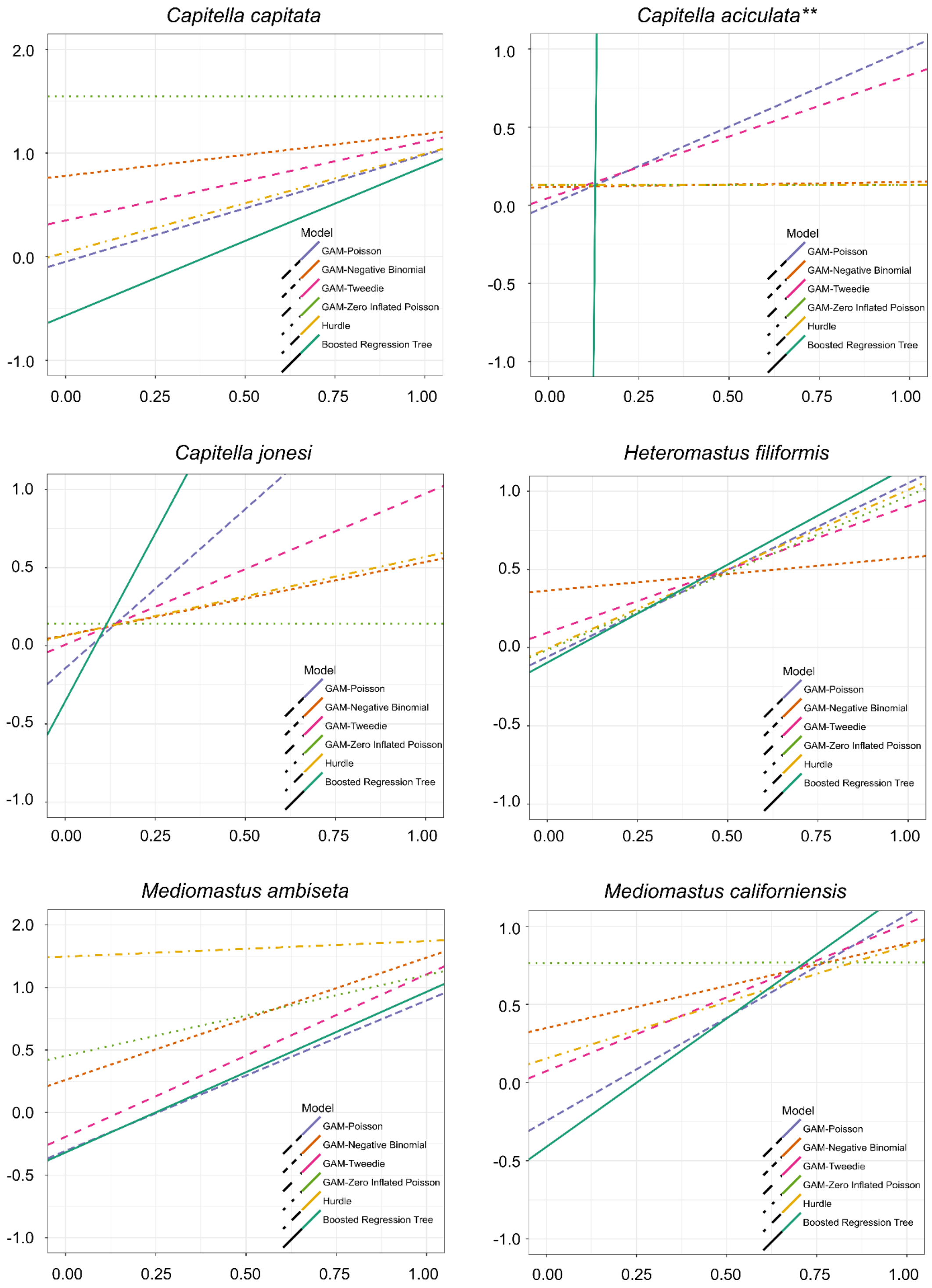


Figure 3.6: Plots of the slope and intercept for all models using values from Table A1. The axes are unitless as only slope and intercept were used for plot generation. However, the y-axis represents the observed count and x-axis the predicted count (Potts and Elith 2006). **Indicates that the biased, apparent values are used.

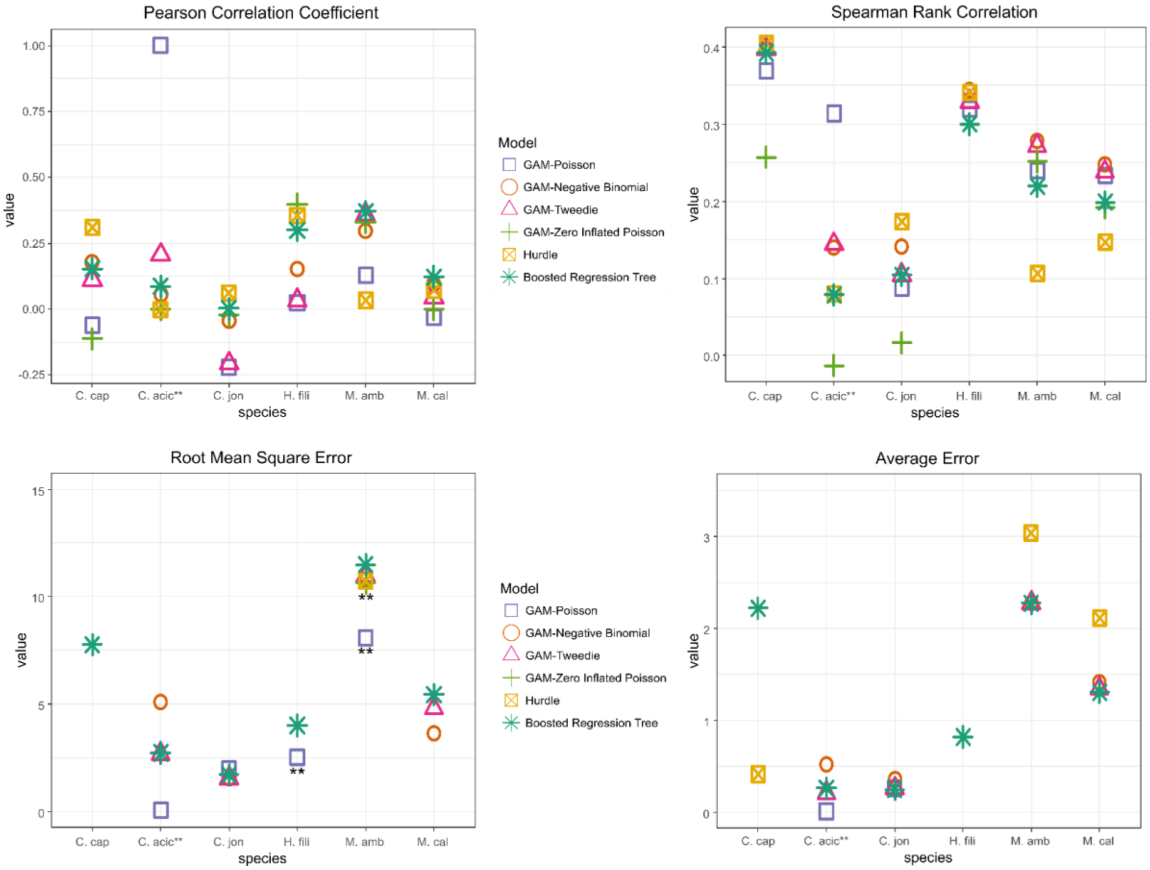


Figure 3.7: Plots of measures of correlation and error between observed and predicted values for all models using values from Table A1. Outlying values were not plotted for root mean square error and average error for visualization purposes. Excluded points are highlighted in Table A1. **Indicates that the biased, apparent values are used.

3.3.3. Environmental Factors

Significance was assessed at alpha 0.05 for all models (Figure 3.8) except for BRTs for which the relative influence of each term is estimated (Friedman 2001, Friedman & Meulman 2003) (Figure A.1).GAM-Poisson and -Zero Inflated Poisson models found all terms significant for every species except for salinity|*H. filiformis* (Figure 3.8). Bay segment was found significant for every species/model combination. Depth was found significant for most species/model combinations and was significant for every best-calibrated model. Total samples had the next overall significance with every species/best-calibrated model combination returning it as significant except for *C. jonesi* (Figure 3.8). This is generally corroborated by the BRTs for which depth or bay segment are within the top two influential terms for most species. The exceptions are *C. aciculata*, which had pH as the only term that influences abundance, and *H. filiformis*, which had depth third but still of strong influence (Figure A.1).

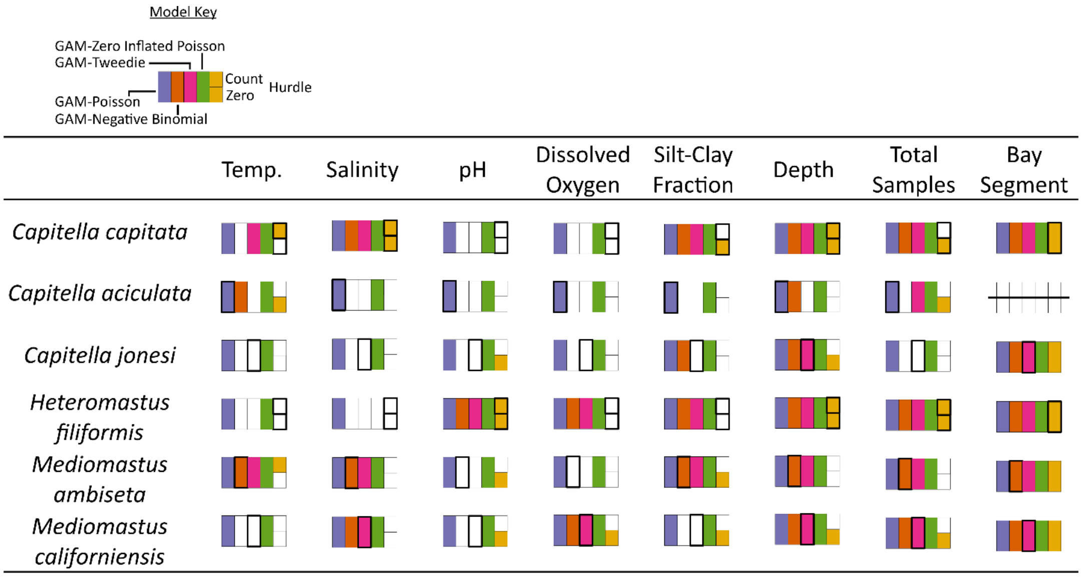


Figure 3.8: A table graph of term significance for all data-based models. There is a key at the top. If the term was significant (alpha = 0.05) for a given species/model combination, that block was colored. All colors correspond to those used for the models in Figures 7 and 8. Notice that the count and zero (presence/absence) parts of the Hurdle model are separated for all terms except Bay Segment. Significance of Bay Segment was assessed for the Hurdle model as a whole using a likelihood-ratio test. Bay Segment was not investigated for *Capitella aciculata*. The model chosen as best calibrated for each species is highlighted with a bold block.

Random forest models indicated that bay segment and/or depth had 50+% importance for every species (Figure 3.9). Bay segment was most important for *C. aciculata*, *C. jonesi*, and *M. ambiseta* while depth was most important for *C. capitata* and *H. filiformis* (Figure 3.9). Dissolved oxygen was most important for *M. californiensis* (Figure 3.9). ALE plots for *C. capitata* and *H. filiformis* indicate a negative relationship of species abundance and depth; higher species abundances are found at shallower depths (<1.9 m and <1.0 m, respectively) (Figure 3.10). Species with bay segment being most important have ALE plots (Figure 3.10) that reflect patterns of spatial autocorrelation observed in Lorenz curves (Figure 3.4). For example, the Lorenz curve for *C. jonesi* indicates a large portion of species abundance in HB and the ALE plot shows a large positive effect of HB. *M. californiensis* has a somewhat sigmoidal relationship with dissolved oxygen, with a large increase in abundance between ~4.4 – 7.5 mg/L (Figure 3.10).

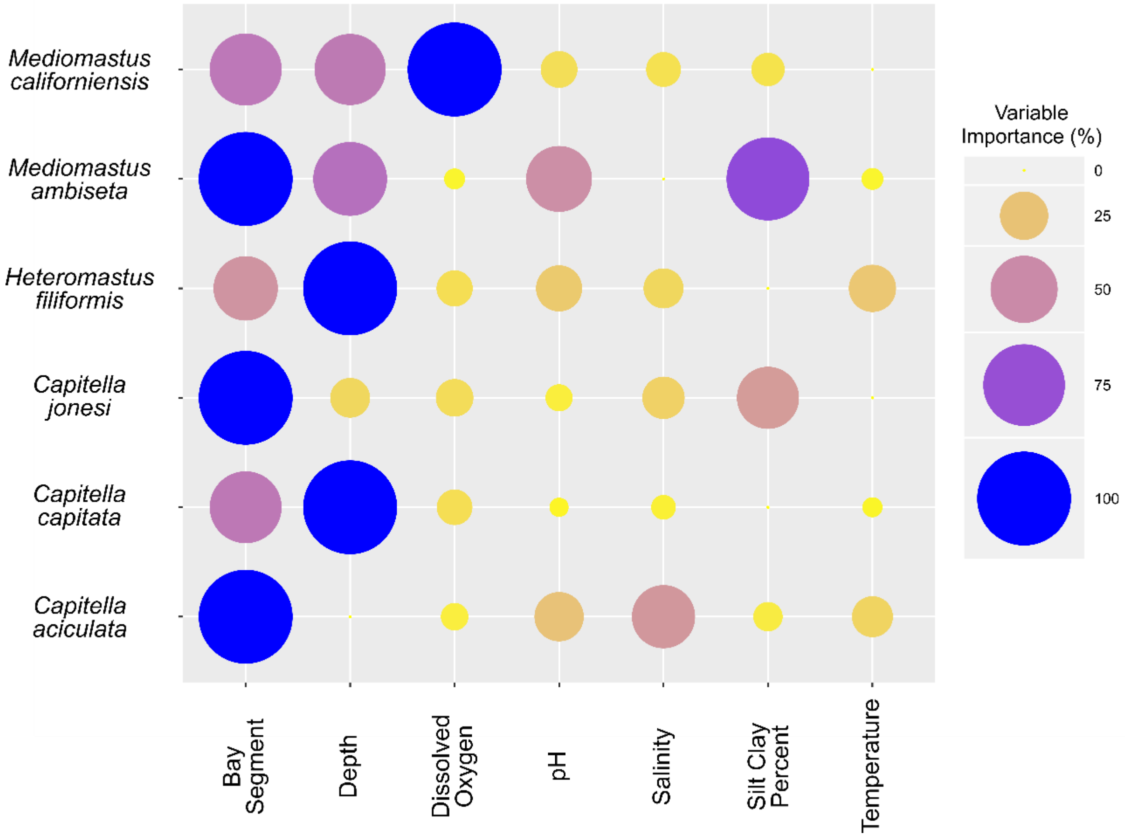


Figure 3.9: Conditional random forest variable importance plots with species models organized in rows and environmental factors in columns. Variable importance represents the permutation importance of each factor, or the average effect on mean squared error when the factor is removed. Each factor's importance is assessed independently, thus importance is not additive for each model. Higher importance equates to higher error when that factor is removed from the model. Importance is represented by size and color scales.

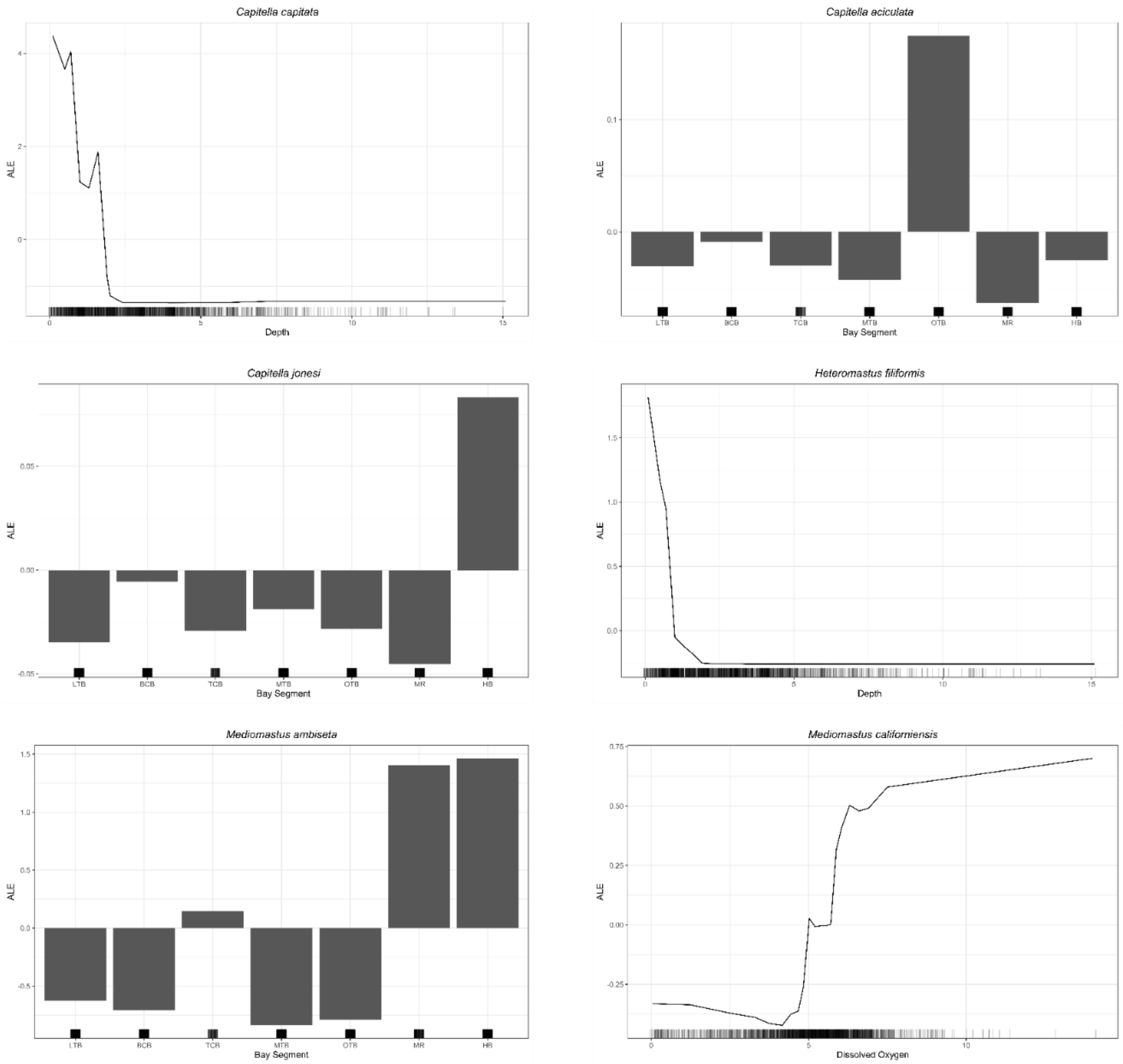


Figure 3.10: Accumulated Local Effects (ALE) plots of terms with top variable importance for each species. ALE scores are interpreted as the difference between the model prediction (at a given instance of x) and the averaged model prediction. Continuous predictors include a rug plot on the x-axis for visualizing observation density. Abbreviations of bay segments and term units are described in the Methods.

3.4. Discussion

Our results demonstrate presence of spatial autocorrelation structured by bay segment for capitellids and variation in model specification across species. Using Lorenz curves in conjunction with violin plots demonstrated an effective way to assess this and overall species abundance structure across different regions of a large area. Interpretation of Lorenz curves is that the further the curve is from the line of perfect equality, the more unequally distributed the species is. One difference from standard Lorenz curves is that ours do not meet the line of equality at the origin. This is because we are using an x-axis with categories that are all assumed to have some abundance in the case of perfect equality. The curve's slope allows for quick assessment of autocorrelation; given a steep slope between two regions, it can be inferred that the region with greater averaged abundance has, proportionally, more of the total species abundance in the entire bay. Some apparent examples are OTB for *Capitella aciculata*, BCB for *Heteromastus filiformis*, and TCB for *Mediomastus ambiseta* (Figure 3.4). Additionally, the slope of the curve to each bay segment reflects the magnitude of effect that bay segment has in ALE plots (results not shown).

Bubble plots of species abundance and LISA plots confirm autocorrelation indicated by Lorenz curves (Figures 3.5). It is important to keep in mind that the small details do not matter much when interpreting the LISA plots as this dataset was collected with large scale patterns in mind. Therefore, what is important is whether or not bay segments appear significantly autocorrelated overall; the fact that one neighborhood is significant and a neighboring group is not has little interpretation because the sampling strategy is not appropriate for such comparisons. For example, *H. filiformis* abundance in BCB (Figure 3.5) has an apparent clustering clear in the bubble plot. LISA plots show several neighborhoods with High-High (high values surrounded by other high values) and Low-High LISA scores (Figure 3.5).

The zero-inflatedness of every species is not surprising. Benthic infaunal invertebrates, especially estuarine polychaetes, are known for having patchy spatial distributions, often associated with grain size or organic matter/food (Warren 1977, James & Gibson 1980, Ansari et al. 1986, Kalejta & Hockey 1991, Sanchez-Moyano & García-Asencio 2009). Larval settlement patterns are sometimes attributed to the presence of conspecific adults (Osman & Whitlatch 1995, Snelgrove et al. 2001), chemical cues in sediment (Qian 1999), and/or subjection to near-bottom flow dynamics, with active selection during passive flow (Butman 1986, 1989, Butman & Grassle 1992, Snelgrove et al. 1993). The presence of some bacteria and their metabolites has even been attributed to inducing larval settlement (Harder et al. 2002, Lau et al. 2003, Chung et al. 2010). *Capitella* are among the better studied marine invertebrates (Grassle & Grassle 1976, Blake et al. 2009, Seaver 2016). One of the cryptic species discovered by Grassle and Grassle (1976) has since been formally described as *C. teleta* (Blake et al. 2009) and has been the subject of several studies on larval settlement (Dubilier 1988, Grassle et al. 1992, Hill & Nelson 1992, Thiyagarajan et al. 2005, Biggers et al. 2012, Burns et al. 2014). There are limited studies on the other genera considered (Hannan 1984, Snelgrove 1994).

Although not considered the best-calibrated model for every species, GAM-Tweedie and/or Hurdle models have intercepts closest to, or near, the origin for all species, indicating low bias in their calibration. One of the two was considered best-calibrated for *C. capitata*, *C. jonesi*, *H. filiformis*, and *M. californiensis* (Table A.1). *C. aciculata* and *M. ambiseta* have GAM-Tweedie slopes within ±0.3 (Table A.1). This is an indication that although these two models may not be deemed the best fit in all cases, they are reasonably calibrated and could be considered a good starting point.

The overall high performance of GAM-Tweedie and Hurdle models may be attributable to their ability to handle the excess zeros in unique ways. Hurdle models have been specifically used for rare species count data (Cunningham and Lindenmayer 2005). This approach splits the dataset into a binary version (presence/absence) and a zero-truncated abundance version. assuming that processes driving presence/absence are separate from those driving abundance (Cragg 1971, Zuur et al. 2009). All other methods keep the dataset whole and assume the processes driving zero-inflation are also driving abundance. A Tweedie distribution allows more flexibility for the species abundance distribution's shape, as this is determined by a power term (p) in the variance function (e.g. p=1 is the Poisson distribution) (Jørgensen 1987). The mgcv r-package has the option to estimate the power term during model fitting, resulting in an automated distribution choice that may fit the data better than the standard Poisson or Negative Binomial distributions.

BRT models have shown equal (Martínez-Rincón et al. 2012) and better (França and Cabral 2015) performance compared to GAMs. Our results indicate better performance of GAMs. This is likely due to the "stump model" restriction (not allowing any interactions), as a key benefit of a BRT is that it can handle very complex interactions that are not possible in the data-based models (De'ath 2007, Elith et al. 2008). It is likely that a well-built BRT could perform just as well as, if not better than, the GAM-Tweedie and Hurdle models. See Elith et al. (2008) for a guide to assembling BRT models and further references on the topic.

General location within a bay and sampling effort are expected to affect how many worms are collected, so the overall significance of bay segment and total samples is not surprising. What was unexpected was the significance of depth, as Tampa Bay is only 4m deep on average (Morrison & Yates 2011). This is not a result of collinearity as standard measures (augmented pairs plots and variance inflation factor values) did not reveal any collinearity among environmental factors (results not shown). Relationships of species abundance and diversity with depth are complex (Houston & Haedrich 1984, Paterson & Lambshead 1995, Sibaja-Cordero et al. 2012). Studies focused on depth gradients in shallow estuarine systems would be useful to better understand this relationship.

We recognize that these data may be confounded by presence of cryptic species. There is little known about how many species comprise the *C. capitata* complex in Tampa Bay, but recent work indicates at least three distinct genetic lineages (Section 2 of this dissertation). It is also possible that *C. aciculata* may not be a distinct species and is part of a *C. capitata* complex lineage (Hilliard et al. 2016). Preliminary work on *H. filiformis* in the Gulf of Mexico indicates the presence of distinct genetic lineages world-wide and likely the presence of another species complex in Capitellidae (JH, pers. obs.). While there has been no work on genetic lineages of *M. ambiseta* and *M. californiensis,* it can be hypothesized that they are also species complexes due to their large geographic range (Blake 2008) and the emerging patterns in *Capitella*. This is, unfortunately, a factor that we cannot control or account for at this time. However, these results still further our general understanding of capitellid ecology.

We have shown that, despite filling a similar ecological niche (burrowing deposit feeders), there is not one model optimal for every species. Much consideration should be given to a taxon’s biology, especially the shape of its distribution in the area of interest, and the structure of the data frame (e.g., sampling design and scale). For example, *C. aciculata* was especially zero-inflated and this required more exploration of model parameterization. This highlights the complex biology of capitellids, as the extreme zero-inflation may be due to this species truly being rare. It may also be that *C. aciculata* is not a unique species (Hilliard et al. 2016) and the records should be combined with *C. capitata* complex until there is further resolution of species boundaries. Consideration of the data structure and the sampling scale and design indicated that spatial autocorrelation needed accounted for on a bay-scale and comparisons at smaller scales were not appropriate. Taking an approach similar to the one presented here allows for systematic comparison of several modeling strategies at once. The model(s) considered best can then be refined. In the case of a benthic infaunal marine invertebrate with zero-inflated presence/absence records, Hurdle and GAM-Tweedie models may be a good place to start if resources are limited.

3.5. References

Adler D (2005) vioplot: violin plot. https://github.com/TomKellyGenetics/vioplot

Ansari ZA, Ingole BS, and Parulekar AH (1986) Effect of high organic enrichment of benthic polychaete population in an estuary. Marine Pollution Bulletin 17: 361-365

Anselin L (1995) Local Indicators of Spatial Association-LISA. Geographical Analysis 27: 93-115

Anselin L, Syabri I, and Kho Y (2006) GeoDa: An Introduction to Spatial Data Analysis. Geographical Analysis 38: 5-22

Apley DW and Zhu J (2016) Visualizing the effects of predictor variables in black box supervised learning models. arXiv: 1612.08468

Baldridge E, Harris DJ, Xiao X, and White EP (2016) An extensive comparison of species-abundance distribution models. PeerJ 4: e2823

Biggers WJ, Pires A, Pechenik JA, Johns E, Patel P, Polson T, and Polson J (2012) Inhibitors of nitric oxide synthase induce larval settlement and metamorphosis of the polychaete annelid Capitella teleta. Invertebrate Reproduction & Development 56: 1-13

Blake JA (2008) Family Capitellidae Grube, 1862. In: Taxonomic atlas of the benthic fauna of the Santa Maria Basin and western Santa Barbara Channel, Vol 7. Santa Barbara Museum of Natural History p 47-53

Blake JA, Grassle JP, and Eckelbarger KJ (2009) Capitella teleta, a new species designation for the opportunistic and experimental Capitella sp. I, with a review of the literature for confirmed records. Zoosymposia 2: 25-53

Breiman L, Friedman JH, Olshen RA, and Stone CJ (1984) Classification and Regression Trees, Chapman and Hall/CRC, Boca Raton, FL

Breiman L (2001) Random Forests. Machine Learning 45: 5-32

Burns RT, Pechenik JA, Biggers WJ, Scavo G, and Lehman C (2014) The B Vitamins Nicotinamide (B3) and Riboflavin (B2) Stimulate Metamorphosis in Larvae of the Deposit-Feeding Polychaete Capitella teleta: Implications for a Sensory Ligand-Gated Ion Channel. PLOS ONE 9(11): e109535

Burt JE, Barber GM, and Rigby DL (2009) Describing data with statistics. In: Elementary Statistics for Geographers. The Guilford Press, New York, p 124-146

Butman CA (1986) Larval Settlement of Soft-Sediment Invertebrates: Some Predictions Based on an Analysis of Near-Bottom Velocity Profiles. Elsevier Oceanography Series 42: 487-513

Butman CA (1989) Sediment-trap experiments on the importance of hydrodynamical processes in distributing settling invertebrate larvae in near-bottom waters. Journal of Experimental Marine Biology Ecology 134: 37-88

Butman CA and Grassle JP (1992) Active habitat selection by Capitella sp. I larvae. I. Two-choice experiments in still water and flume flows. Journal of Marine Research 50: 669-715

Carmo RFR, Amorim HP, and Vasconcelos SD (2013) Scorpion diversity in two types of seasonally dry tropical forest in the semi-arid region of Northeastern Brazil. Biota Neotropica 13(2): 340-344

Carr CM, Hardy SM, Brown TM, Macdonald TA, and Hebert PD (2011) A tri-oceanic perspective: DNA barcoding reveals geographic structure and cryptic diversity in Canadian polychaetes. PLOS ONE 6: e22232

Chung HC, Lee OO, Huang Y-L, Mok SY, Kolter R, and Qian P-Y (2010) Bacterial community succession and chemical profiles of subtidal biofilms in relation to larval settlement of the polychaete Hydroides elegans. The ISME Journal 4: 817-828

Connolly SR, Dornelas M, Bellwood DR, and Hughes TP (2009) Testing species abundance models: a new bootstrap approach applied to Indo-Pacific coral reefs. Ecology 90(11): 3138-3149

Cragg JG (1971) Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. Econometrica 39: 829-844

Cunningham RB and Lindenmayer DB (2005) Modeling count data of rare species: some statistical issues. Ecology 86(5): 1135-1142. doi https://doi.org/10.1890/04-0589

Dean HK (2008) The use of polychaetes (Annelida) as indicator species of marine pollution: a review. Revista de Biología Tropical Trop 56: 11-38

De'ath G (2007) Boosted trees for ecological modeling and prediction. Ecology 88: 243-251

Dubilier N (1988) H2S - A Settlement Cue or a Toxic Substance for Capitella sp. I Larvae? The Biological Bulletin 174: 30-38

Efron B and Tibshirani R (1997) Improvements on cross-validation: the .632+ bootstrap method. Journal of the American Statistical Association 92: 548-560

Elith J and Leathwick JR (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics 40: 677-697

Elith J, Leathwick JR, and Hastie T (2008) A working guide to boosted regression trees. Journal of Animal Ecology 77: 802-813

França S and Cabral HN (2015) Predicting fish species richness in estuaries: Which modelling technique to use? Environmental Modelling & Software 66: 17-26. doi https://doi.org/10.1016/j.envsoft.2014.12.010

Freund Y and Schapire RE (1996) Experiments with a new boosting algorithm. In: Saitta L (ed) Machine Learning: Proceedings of the Thirteenth International Conference. Morgan Kaufmann, Bari, Italy

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. The Annals of Statistics 29: 1189-1232

Friedman JH and Meulman JJ (2003) Multiple additive regression trees with application in epidemiology. Statistics in Medicine 22: 1365-1381

Glick P and Clough J (2006) An Unfavorable Tide—Global Warming, Coastal Habitats and Sport fishing in Florida. National Wildlife Federation. https://www.nwf.org/Educational-Resources/Reports/2006/06-01-2006-Unfavorable-Tide

Grassle JP and Grassle JF (1976) Sibling Species in the Marine Pollution Indicator Capitella (Polychaeta). Science 192: 567-569

Grassle JP, Butman CA, and Mills SW (1992) Active habitat selection by Capitella sp. I larvae. II. Multiple-choice experiments in still water and flume flows. Journal of Marine Research 50: 717-743

Guisan A, Edwards TC, and Hastie T (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling 157: 89-100

Hannan CA (1984) Planktonic larvae 05 act like passive particles in turbulent near-bottom flows. Limnology and Oceanography 29: 1108-1116

Harder T, Lau SCK, Dahms HU, and Qian PY (2002) Isolation of bacterial metabolites as natural inducers for larval settlement in the marine polychaete Hydroides elegans (Haswell). Journal of Chemical Ecology 28: 2029-2043

Hastie T and Tibshirani R (1986) Generalized Additive Models. Statistical Science 1: 297-310

Hastie TJ and Tibshirani RJ (1990) Generalized Additive Models. Chapman and Hall/CRC, Boca Raton, FL

Hegel TM, Cushman SA, Evans J, and Huettmann F (2010) Current State of the Art for Statistical Modelling of Species Distributions. In: Cushman SA, Huettmann F (eds)

Spatial Complexity, Informatics, and Wildlife Conservation. Springer, Tokyo, Berlin, Heidelberg, New York, p 273-311

Hill SD and Nelson L (1992) Lindane (1, 2, 3, 4, 5, 6-Hexachlorocyclohexane) Affects Metamorphosis and Settlement of Larvae of Capitella species I (Annelida, Polychaeta). The Biological Bulletin 183: 376-377

Hilliard J, Hajduk M, and Schulze A (2016) Species delineation in the Capitella species complex (Annelida: Capitellidae): geographic and genetic variation in the northern Gulf of Mexico. Invertebrate Biology 135: 415-422

Houston KA and Haedrich RL (1984) Abundance and biomass of macrobenthos in the vicinity of Carson Submarine Canyon, northwest Atlantic Ocean. Marine Biology 82: 301-305

Hothorn T, Buehlmann P, Dudoit S, Molinaro A, and Van Der Laan M (2006) Survival Ensembles. Biostatistics 7(3): 355-373

Jackman S (2017) pscl: classes and methods for R developed in the political science computational laboratory. United States Study Centre, University of Sydney. Sydney, New South Wales, Australia. R package version 1.5.2. https://github.com/atahk/pscl/

James CJ and Gibson R (1980) The distribution of the polychaete Capitella capitata (Fabricius) in dock sediments. Estuarine and Coastal Marine Science 10: 671-683

Jørgensen B (1987) Exponential dispersion models. Journal of the Royal Statistical Society. Series B 49: 127-162

Judge J and Barry JP (2016) Macroinvertebrate community assembly on deep-sea wood falls in Monterey Bay is strongly influenced by wood type. Ecology 97: 3031-3043

Kalejta B and Hockey PAR (1991) Distribution, abundance, and productivity of benthic invertebrates at the Berg River Estuary, South Africa. Estuarine, Coastal and Shelf Science 33: 175-191

Karlen DJ, Dix TL, Goetting BK, Markham SE, Campbell KW, and Jernigan JM (2015) Twenty year trends in the benthic community and sediment quality of Tampa Bay: 1993-2012. Tampa Bay Benthic Monitoring Program Interpretive Report Prepared for: Tampa Bay Estuary Program. http://www.epchc.org/about/publications-reports, listed as: 20 Year Tampa Bay Benthic Report

Kuhn M (2020) caret: Classification and Regression Training. R package version 6.0-85. https://CRAN.R-project.org/package=caret

Lau SC, Harder T, and Qian PY (2003) Induction of larval settlement in the serpulid polychaete Hydroides elegans (Haswell): role of bacterial extracellular polymers. Biofouling 19: 197-204

Livi S, Tomassetti P, Vani D, and Marino G (2017) Genetic evidences of multiple phyletic lineages of Capitella capitata (Fabricius 1780) complex in the Mediterranean Region. Journal of Mediterranean Ecology 15: 5-11

Lobo J, Teixeira MA, Borges LM, Ferreira MS, Hollatz C, Gomes PT, Sousa R, Ravara A, Costa MH, and Costa FO (2016) Starting a DNA barcode reference library for shallow water polychaetes from the southern European Atlantic coast. Molecular Ecology Resources 16: 298-313

Lorenz MO (1905) Methods of measuring the concentration of wealth. Publications of the American Statistical Association 9: 209-219

Man-Ki J, Wi JH, and Suh H-L (2017) A reassessment of Capitella species (Polychaeta:

Capitellidae) from Korean coastal waters, with morphological and molecular evidence. Marine Biodiversity 48: 1969-1978

Martínez-Rincón R, Ortega-García S, and Vaca-Rodríguez JG (2012) Comparative performance of generalized additive models and boosted regression trees for statistical modeling of incidental catch of wahoo (Acanthocybium solandri) in the Mexican tuna purse-seine fishery. Ecological Modelling 233: 20-25. doi https://doi.org/10.1016/j.ecolmodel.2012.03.006

McCullagh P and Nelder JA (1983) Generalized Linear Models Monographs on Statistics and Applied Probability. Chapman and Hall/CRC, Boca Raton, FL

Méndez N, Linke-Gamenick I, and Forbes VE (2000) Variability in reproductive mode and larval development within the Capitella capitata species complex. Invertebrate Reproduction & Development 38: 131-142

Molnar C, Bischl B, and Casalicchio G (2018) iml: An R package for Interpretable Machine Learning. The Journal of Open Source Software 3(26): 786. doi 10.21105/joss.00786

Molnar C (2019) Interpretable machine learning. A Guide for Making Black Box Models Explainable URL https://christophm.github.io/interpretable-ml-book/

Montagna PA, Palmer TA, Kalke RD, and Gossmann A (2008) Suitability of Using a Limited Number of Sampling Stations to Represent Benthic Habitats in Lavaca-Colorado Estuary, Texas. Environmental Bioindicators 3: 156-171

Morrison G and Yates KK (2011) Environmental setting. In: Yates KK, Greening H, Morrison G (eds) Integrating Science and Resource Management in Tampa Bay, Florida. U.S. Geological Survey Circular 1348. https://doi.org/10.3133/cir1348

Nelder JA and Wedderburn RWM (1972) Generalized Linear Models. Journal of the Royal Statistical Society. Series A 135(3): 370-384

Oppel S, Meirinho A, Ramírez I, Gardner B, O'Connell AF, Miller PI, and Louzao M (2012) Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds. Biological Conservation 156: 94-104

Osman RW and Whitlatch RB (1995) The influence of resident adults on larval settlement: experiments with four species of ascidians. Journal of Experimental Marine Biology and Ecology 190: 199-220

Palmer TA, Montagna PA, Pollack JB, Kalke RD, and DeYoe HR (2011) The role of freshwater inflow in lagoons, rivers, and bays. Hydrobiologia 667: 49-67

Paterson GLJ and Lambshead PJD (1995) Bathymetric patterns of polychaete diversity in the Rockall Trough, northeast Atlantic. Deep Sea Research Part I: Oceanographic Research Papers 42: 1199-1214

Peterson AT and Soberón J (2012) Species Distribution Modeling and Ecological Niche Modeling: Getting the Concepts Right. Natureza & Conservação 10(2): 102-107

Potts JM and Elith J (2006) Comparing species abundance models. Ecological Modelling 199: 153-163

Qian P-Y (1999) Larval settlement of polychaetes. Hydrobiologia 402: 239-253

R Core Team (2018) R: A language and environment for statistical computing, Vienna, Austria. https://www.R-project.org

R Core Team (2019) R: A language and environment for statistical computing, Vienna, Austria. https://www.R-project.org

Robinson NM, Nelson WA, Costello MJ, Sutherland JE, and Lundquist CJ (2017) A

Systematic Review of Marine-Based Species Distribution Models (SDMs) with Recommendations for Best Practice. Frontiers in Marine Science 4: 421

Sánchez-Moyano JE and García-Asencio I (2009) Distribution and trophic structure of annelid assemblages in a Caulerpa prolifera bed from southern Spain. Marine Biology Research 5: 122-132

Seaver EC (2016) Annelid models I: Capitella teleta. Current Opinion in Genetics & Development 39: 35-41

Shelton AO, Thorson JT, Ward EJ, and Feist BE (2014) Spatial semiparametric models improve estimates of species abundance and distribution. Canadian Journal of Fisheries and Aquatic Sciences 71(11): 1655-1666

Sibaja-Cordero JA, Cortés J, and Dean HK (2012) Depth diversity profile of polychaete worms in Bahía Chatham, Isla del Coco National Park, Costa Rican Peninsula. Revista de Biología Tropical 60: 293-301

Sillero N (2011) What does ecological modelling model? A proposed classification of ecological niche models based on their underlying methods. Ecological Modelling 222: 1343-1346

Silva CF, Shimabukuro M, Alfaro-Lucas JM, Fujiwara Y, Sumida PYG, and Amaral ACZ (2016) A new Capitella polychaete worm (Annelida: Capitellidae) living inside whale bones in the abyssal South Atlantic. Deep Sea Research Part I: Oceanographic Research Papers 108: 23-31

Silva CF, Seixas VC, Barroso R, Di Domenico M, Amaral ACZ, and Paiva PC (2017) Demystifying the Capitella capitata complex (Annelida, Capitellidae) diversity by morphological and molecular data along the Brazilian coast. PLOS ONE 12:

e0177760

Snelgrove PVR, Butman CA, and Grassle JP (1993) Hydrodynamic enhancement of larval settlement in the bivalve Mulinia lateralis (Say) and the polychaete Capitella sp. I in microdepositional environments. Journal of Experimental Marine Biology and Ecology 168: 71-109

Snelgrove PVR (1994) Hydrodynamic enhancement of invertebrate larval settlement in microdepositional environments: colonization tray experiments in a muddy habitat. Journal of Experimental Marine Biology and Ecology 176: 149-166

Snelgrove PVR, Grassle JP, and Zimmer CA (2001) Adult macrofauna effects on Capitella sp. I larval settlement: A laboratory flume study. Journal of Marine Research 59: 657-674

Spalding MD, Fox HE, Allen GR, Davidson N, et al. (2007) Marine Ecoregions of the World: A Bioregionalization of Coastal and Shelf Areas. Bioscience 57: 573-583

Squires A, Janicki A, Heimbuch D, Wade D, Wilson H, and Robison D (1994) A monitoring program to assess environmental changes in Tampa Bay, Florida. https://tbeptech.org/34mkf/6bb2b_ngbz/vxoi13.dol

Steyerberg EW, Harrell FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, and Habbema JDF (2001) Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. Journal of Clinical Epidemiology 54: 774-781

Strobl C, Boulesteix A, Zeileis A, and Hothorn T (2007) Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. BMC Bioinformatics 8(25). doi https://doi.org/10.1186/1471-2105-8-25

Strobl C, Boulesteix A, Kneib T, Augustin T, and Zeileis A (2008) Conditional Variable Importance for Random Forests. BMC Bioinformatics 9(307). doi https://doi.org/10.1186/1471-2105-9-307

Strobl C, Hothorn T, and Zeileis A (2009) Party on! A New, Conditional Variable-Importance Measure for Random Forests Available in the party Package. The R Journal 1(2): 14-17. doi 10.32614/RJ-2009-013

Thiyagarajan V, Soo L, and Qian P-Y (2005) The role of sediment organic matter composition in larval habitat selection by the polychaete Capitella sp. I. Journal of Experimental Marine Biology and Ecology 323: 70-83

Tomioka S, Kondoh T, Sato-Okoshi W, Ito K, Kakui K, and Kajihara H (2016) Cosmopolitan or Cryptic Species? A Case Study of Capitella teleta (Annelida: Capitellidae). Zoological Science 33: 545-554

Tomioka S, Kakui K, and Kajihara H (2018) Molecular Phylogeny of the Family Capitellidae (Annelida). Zoological Science 35: 436-445

Van Diggelen AD and Montagna PA (2016) Is Salinity Variability a Benthic Disturbance in Estuaries? Estuaries and Coasts 39: 967-980

Warren LM (1977) The ecology of Capitella capitata in British waters. Journal of the Marine Biological Association of the United Kingdom 57: 151-159

Wickham H (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York. https://ggplot2.tidyverse.org

Wood SN (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73: 3-36

Wood SN, Pya N, and Säfken B (2016) Smoothing Parameter and Model Selection for General Smooth Models. Journal of the American Statistical Association 111: 1548-1563

Wood SN (2017) Generalized Additive Models: An Introduction with R, 2nd Ed. Chapman and Hall/CRC, Boca Raton, FL

Yates KK and Greening H (2011) An Introduction to Tampa Bay. In: Yates KK, Greening H, Morrison G (eds) Integrating Science and Resource Management in Tampa Bay, Florida. U.S. Geological Survey Circular 1348. https://doi.org/10.3133/cir1348

Zeileis A and Hothorn T (2002) Diagnostic checking in regression relationships. R News p 7-10

Zeileis A, Kleiber C, and Jackman S (2008) Regression models for count data in R. Journal of Statistical Software, Articles 27(8)

Zuur AF, Ieno EN, Walker NJ, Saveliev AA, and Smith GM (2009) Mixed Effects Models and Extensions in Ecology with R. Springer Science + Business Media, New York

4. ANALYSIS OF ANNELID TRANSCRIPTOMES AND GENOMES REVEALS HIGH DIVERSITY OF THE OXYGEN SENSING TRANSCRIPTION FACTOR HYPOXIA INDUCIBLE FACTOR (HIF)

4.1. Introduction

Oxygen shapes the composition and functioning of most biological communities. Throughout geological time, hyperoxic conditions have led to bursts of diversification (Graham et al. 1995; Berner et al. 2007) and gigantism (Dudley 1998) whereas hypoxia generally leads to the opposite effects (Levin 2003; Heinrich et al. 2011; Friedrich et al. 2014). In marine and estuarine environments, oxygen levels can vary greatly spatially and temporally. Hypoxia zones are natural phenomena, occurring in oxygen minimum zones, enclosed basins or semi-enclosed fjords, but can be exacerbated by anthropogenic activities, in extreme cases leading to “dead zones” (Rabalais et al. 2002; Díaz and Rosenberg 2008). Hypoxia is generally defined as dissolved oxygen (DO) levels below 2 ml O_2/l (Díaz and Rosenberg 1995). This degree of oxygen depletion can lead to severely diminished diversity and biomass, especially when combined with other environmental stressors such as warming temperatures and acidification (Pörtner et al. 2005; Rosa and Seibel 2008; Melzner et al. 2013; Rosa et al. 2013). The threshold value of 2 ml O_2/l is somewhat arbitrary, as different organisms exhibit different sensitivities to diminishing O_2 and the viability and fitness of an organism are often negatively affected at only slightly decreased O_2 levels well above the 2 ml O_2/l threshold (Vaquer-Sunyer and Duarte 2008).

The phylum Annelida comprises a large diversity of species inhabiting terrestrial, freshwater and marine habitats. Polychaetes are a subset of annelids which form an important component of marine benthic communities. Many species survive or even thrive under low oxygen or hypoxia and can tolerate significant fluctuations in DO levels by sophisticated mechanisms for maximizing oxygen uptake, transport and delivery (Kristensen 1983, Hourdez and Lallier 2007, Qu et al. 2015). However, it is not well studied how they sense DO levels and how the physiological coping mechanisms are activated.

Due to its relevance for physiology and medicine, oxygen sensing is well studied in humans and other vertebrates, and to some degree in invertebrate models such as *Drosophila* (Nambu et al. 1996), *Caenorhabditis* (Epstein et al. 2001; Jiang et al. 2001), and even the simple *Trichoplax adhaerens* (Leonarz et al. 2011). In all of these vastly divergent species, the physiological response to low oxygen is triggered by molecular cascades in which Hypoxia Inducible Factors (HIFs) play a key role. HIFs are heterodimeric transcription factors with a basic helix-loop-helix (bHLH) motif which trigger cellular responses to low oxygen (Wang et al. 1995). Under normoxic conditions, one of the subunits, HIFα, is hydroxylated at two proline residues by the enzymes HIF proline hydroxylase (PHD) and Factor Inhibiting HIF1 (FIH1). Both PHD and FIH1 are oxygen dependent, but FIH1 remains functional under lower DO concentrations than PHD, and thus may be particularly important under moderate hypoxia (Kaelin Jr and Ratcliffe 2008).

Proline hydroxylation of HIFα promotes interaction with a cellular protein complex containing von Hippel-Lindau factor (vHL), eventually leading to proteasomal degradation (Kaelin Jr 2005). Under hypoxic conditions, the degradation is disrupted because PHD and FIH1 become non-functional. As a consequence, HIFα accumulates in the cytoplasm and dimerizes with the stable subunit HIFβ (also known as Aryl Hydrocarbon Nuclear Translator, ARNT). The dimer translocates to the nucleus and acts as a transcription factor by binding to short Hypoxia Response Elements (HREs) in the genome, enhancing transcription of downstream genes. Thus, HIFs are only functional as transcription factors under hypoxic conditions.

Three HIFαs have been characterized in metazoans, but only HIF1α and HIF2α have been studied extensively (Kaelin Jr and Ratcliffe 2008). *Drosophila* and *Caenorhabditis* only have a single PHD family member, known as Egl9, whereas three PHDs occur in vertebrates (Kaelin Jr and Ratcliffe 2008). By comparison, very little is known about oxygen sensing in the third branch of bilaterians, the lophotrochozoans or spiralians. The only lophotrochozoan taxa studied to date in this respect are several species of molluscs, including two species of oysters (Piontkivska et al. 2011; Kawabe and Yokoyama 2012), a mussel (Giannetto et al. 2015) and an abalone (Cai et al. 2014). The above studies have characterized HIFα, HIFβ, and PHD in molluscs, but FIH1 has not been identified in any lophotrochozoan. It has primarily been characterized in vertebrates, as well as recently in a freshwater shrimp (Sun et al. 2016).

HIFα and ARNT are part of the basic helix-loop-helix-per-arnt-sim (bHLH+PAS) family, characterized by a bHLH and two PAS domains (Figure 4.1). Additionally, HIFα has two transactivation domains, a c-terminal (CTAD) and n-terminal (NTAD) (Figure 4.1). CTAD contains an asparagine residue that is hydroxylated under normoxia and binds to the p300 transcription coactivator under hypoxia (Lando et al. 2002). NTAD confers target gene binding specificity (Hu et al. 2007).

There have been multiple studies on the evolution of the bHLH super family (Ledent et al. 2002, Simionato et al. 2007) as well as the bHLH+PAS gene family within it (Yan et al. 2014, Graham and Presnell 2017). These studies focused on identifying overarching patterns in metazoans, with hardly any representation of annelids (but see Simionato et al. 2007). This study aims to examine the diversity and evolution of the two subunits of the HIF transcription factor across the phylum Annelida. We expect to find high diversity of these genes as annelids range widely in their activity patterns and feeding modes (e.g. sessile passive filter feeders vs. highly motile predators) and occur in virtually all marine habitats, spanning well oxygenated to hypoxic conditions.

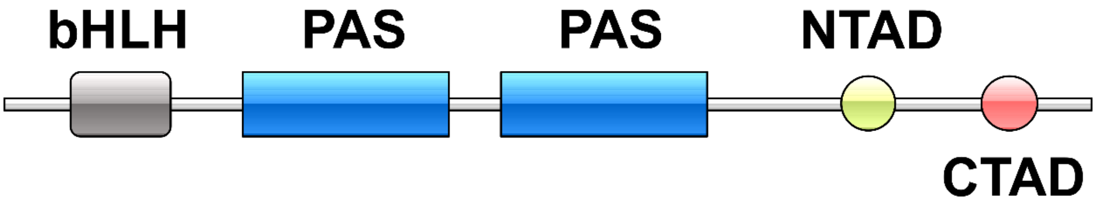


Figure 4.1: Illustration depicting the domain structure of hypoxia inducible factor with the basic helix-loop-helix, per-arnt-sim, n-terminal transactivation, and c-terminal transactivation domains labeled. Aryl receptor nuclear transferase has the same arrangement but without the transactivation domains. Illustration created with Illustrator for Biological Sequences.

4.2. Materials and Methods

4.2.1. Public Data Sourcing and Transcriptome Assembly

Most data were retrieved from the NCBI Sequence Read Archive (SRA) using the SRA Toolkit (Table 4.1). Raw SRA files were converted to FASTQ and split into paired end sets. Trinity v2.5.1 (Grabherr et al. 2011; Haas et al. 2013) was used for transcriptome assembly. Reads were quality-trimmed with the --trimmomatic option and normalized (i.e.

reads with > 50x representation were discarded). Transcriptomes for several other annelids and the outgroup were sourced from Kocot et al. (2017), where assemblies are provided. Predicted gene sets for *Capitella teleta* and *Helobdella robusta* genomes were sourced from EnsemblMetazoa (Yates et al. 2020).

4.2.2. Capitella sp. TV RNA-Sequencing and Transcriptome Assembly

Material was sourced from a lab culture started with specimens from Galveston Bay, Texas. Total RNA was extracted from 50 metatrochophore larvae (Stage 9, Seaver et al. 2005) (two replicates of 25 larvae), two adult males, two adult females, and two adult hermaphrodites using TRIzol for extraction and Qiagen RNeasy for cleanup. TRIzol protocol differences: pestle (~25 twists) and 27 gauge syringe (~10 draws) used for homogenization; precipitation of RNA by adding 0.1volume 3M Na Acetate and 2.5volume 100% EtOH, inverting five times, and incubating at -80°C for 110 minutes; air dried RNA pellet for 20 minutes; and final elution in 107µl nuclease-free water. RNeasy protocol differences: all centrifugations were at 4°C; all 15sec centrifugations were for 30sec; the optional full speed centrifugation for 1min was run for 2min followed by leaving the sample in the centrifuge for 10min to dry the column better; a 57µl elution was performed to use 7µl for sample quantification and qualification and the remaining 50µl for library prep; and samples were stored at -80°C until used.

Sequencing libraries were prepared and sequenced by TAMU AgriLife. Illumina TruSeq RNA stranded libraries were prepared. Sequencing was conducted on an Illumina NovaSeq 6000. Illumina software (NCS v1.0.2 and RFV v1.0.2, default settings) were used for sequence cluster identification, quality prefiltering, base calling, and uncertainty assessment. NovaSeq basecall files were demultiplexed and converted to FASTQ using

Illumina bcl2fastq2 v2.19.0. A single transcriptome for the species was assembled using all of the data following the same approach used for the public data.

4.2.3. Open Reading Frame (ORF) Prediction and Basic Local Alignment Tool (BLAST) Searches

ORFs were predicted for all transcriptomes and genomes using TransDecoder v3.01 (Haas & Papanicolaou et al. in prep) with default settings. Nucleotide BLAST databases were built with the predicted ORFs. Reference sequences were obtained from multiple sources (Satou et al. 2003, Simionato et al. 2007, Gyoja 2014, Fortunato et al. 2016, Graham and Presnell 2017). These were used as query sequences in tBLASTn searches to identify putative homologs. BLAST hits with E-Value ≤1E-06 were considered significant, filtered for unique headers, and converted to protein sequences for further analysis.

4.2.4. Gene Identification by Phylogenetic Analysis

Significant ORFs from both BLAST searches were concatenated with HIFα, SIM, NPAS1-4, ARNT(2), and ARNTL(2) sequences from Graham and Presnell (2017). Redundant sequences were removed with Jalview v2.11.0 (Warehouse et al. 2009) and MAFFT v7.266 (Katoh and Standley 2013) accurate local alignment was performed. The multiple sequence alignment was trimmed and again filtered for redundancy in Jalview. A maximum likelihood (ML) phylogeny was constructed with IQ-TREE v1.6.12 (Nguyen et al. 2015) with automatic model selection restricted to only amino acid nuclear models via ModelFinder (Kalyaanamoorthy et al. 2017). Branch support was determined using the SH-like approximate likelihood ratio test (SH-aLRT) (Guindon et al. 2010) and ultrafast bootstrap (UFBoot) (Minh et al. 2013).

These two metrics have different interpretations from standard, nonparametric bootstrapping (Felsenstein 1985). See the above references and IQ-TREE manual for details. *Drosophila melanogaster* Hey gene (Simionato et al. 2007) was used as an outgroup. The tree was viewed using FigTree v1.4.4 (Rambaut 2012) to define HIFα and ARNT clades. Sequences were manually parsed out of the pre-alignment file for subsequent analyses.

4.2.5. Isoform Selection and Gene-Specific Phylogenetic Analyses

To represent each species with a single sequence in each phylogenetic tree, only one isoform can be used. Isoform selection was performed with IsoSel v1.0 (Philippon et al. 2017) with default settings and a species/gene identifier file (for automatic filtering of sequences). Multiple sequence alignments were built using MAFFT accurate local alignment. ML phylogenetic analyses were performed as above. *Crassostrea gigas* was used as an outgroup for both analyses.

4.2.6. Domain Annotation

h*mmsearch* from HMMER v3.1b2 (Eddy 2009) was used to search Hidden Markov Model Profiles (HMMs) for different domains against the final HIFα and ARNT pre-alignment dataframes. HMMs downloaded from the Pfam database include helix-loop-helix (HLH) (PF00010), per-arnt-sim (PAS) (PF00989), HIF N-terminal transactivation domain (NTAD) (HIF-1; PF11413), and HIF-1 alpha C terminal transactivation domain (CTAD) (PF08778).

Given the limited annelid representation in the HMMs, searches were conducted iteratively, updating the HMM with significant hits (E-value ≤1e-5) until the domain was found in all sequences or the search stopped returning significant hits. Significant domains were extracted from sequences using hmmer3_extract_domains.pl (https://github.com/amgraham07/HIF_eukaryote) (Graham and Presnell 2017).

4.3. Results

4.3.1. BLAST Searches and Gene Identification by Phylogenetic Analysis

The final database had 72 annelid species and the outgroup *Crassostrea gigas*. A total of 69 HIFα and 51 ARNT sequences (available in a supplementary file) were used to query the annelid database. Before any reduction, the starting dataframe had 1965 sequences from BLAST searches and 180 annotated bHLHs. After trimming and removal of redundant sequences, the final alignment for analysis consisted of 542 sequences. HIFα and ARNT were recovered for 43 (58.9%) and 40 (54.8%) of the species in the database, respectively (Table 4.1). The black branches in Fig. 4.2 are annelid bHLH sequences which did not cluster with any of the known bHLH genes used and remain to be characterized.

Table 4.1: Species used in this study, NCBI SRA accession information for RNA Seq data, and whether HIFα and ARNT were found for them.

SRA/Accession	Major Group	Family	Species	Environment	HIFα	ARNT
SRR5353284	Sedentaria	Acanthodrilidae	Maoridrilus_wilkini	terrestrial		
SRR1221445	Sedentaria	Acrocirridae	Macrochaeta_clavicornis	marine		
SRR3665377	Sedentaria	Alvinellidae	Paralvinella_grasslei	marine	y	y
SRR5590961	Sedentaria	Ampharetidae	Hypania_invalida	freshwater	y	y
SRR651044	Basal	Amphinomidae	Hermodice_carunculata	marine	y	y
SRR1257732	Basal	Amphinomidae	Paramphinome_jeffreysii	marine	y	
SRR5590965	Sedentaria	Apistobranchidae	Apistobranchus_tullbergi	marine		
SRR2005653	Sedentaria	Arenicolidae	Arenicola_marina	marine	y	y
THIS STUDY	Sedentaria	Capitellidae	Capitella_sp_TV	marine	y	y
EnsemblMetazoa	Sedentaria	Capitellidae	Capitella_teleta	marine	y	y
SRR1646443	Basal	Chaetopteridae	Chaetopterus_sp	marine	y	y
SRR1219647	Basal	Chaetopteridae	Chaetopterus_variopedatus	marine		
SRR1257898	Basal	Chaetopteridae	Phyllochaetopterus_sp	marine	y	y
SRR1224605	Basal	Chaetopteridae	Spiochaetopterus_sp	marine	y	y
SRR5590966	Sedentaria	Cirratulidae	Cirratulus_cirratus	marine	y	y
SRR2018886	Sedentaria	Dinophilidae	Dinophilus_taeniatus	marine		y
SRR2014693	Sedentaria	Dinophilidae	Trilobodrilus_axi	marine		y
SRR2014574	Sedentaria	Dinophilidae(?)	Apharyngtus_punicus	marine		
SRR2131612	Sedentaria	Diurodrilidae	Diurodrilus_subterraneus	marine		y
SRR2014681	Errantia	Dorvilleidae	Protodorvillea_kefersteini	marine		
SRR2017645	Sedentaria	Echiura	Bonellia_viridis	marine	y	y
SRR2040479	Errantia	Eunicidae	Eunice_pennata	marine	y	y
SRR1232833	Errantia	Eunicidae	Marphysa_bellii	marine	y	
SRR2017643	Sedentaria	Fauveliopsidae	Fauveliopsis_sp	marine	y	y
SRR3574613	Sedentaria	Flabelligeridae	Flabelligera_mundata	marine		
EnsemblMetazoa	Sedentaria	Glossiphoniidae	Helobdella_robusta	freshwater		y
Kocot et al. (2017)	Errantia	Glyceridae	Glycera_dibranchiata	marine		
SRR1237870	Errantia	Glyceridae	Glycera_tridactyla	marine		
SRR4162952	Sedentaria	Haemadipsidae	Haemadipsa_cavatuses	terrestrial	y	y
SRR4162958	Sedentaria	Haemopidae	Whitmania_pigra	freshwater	y	y
SRR6371134	Sedentaria	Hirudinidae	Hirudo_nipponica	freshwater	y	y
SRR4162957	Sedentaria	Hirudinidae	Poecilobdella_javanica	freshwater	y	y
SRR923752	Sedentaria	Lumbricidae	Lumbricus_terrestris	terrestrial	y	y
SRR1257639	Basal	Magelonidae	Magelona_berkeleyi	marine		
SRR1222290	Basal	Magelonidae	Magelona_johnstoni	marine	y	y
Kocot et al. (2017)	Sedentaria	Maldanidae	Clymenella_torquata	marine		
Kocot et al. (2017)		MOLLUSCA	Crassostrea_gigas		y	y

Table 4.1 Continued

SRA/Accession	Major Group	Family	Species	Environment	HIFα	ARNT
SRR1237872		Myzostomatidae	Myzostoma_cirriferum	marine	y	y
SRR1232795	Errantia	Nephtyidae	Nephtys_caeca	marine		
Kocot et al. (2017)	Errantia	Nereididae	Alitta_succinea	marine		y
SRR1742987	Errantia	Nereididae	Platynereis_dumerilii	marine	y	y
SRR2014581	Sedentaria	Nerillidae	Mesonerilla_fagei	marine		
SRR2017631	Sedentaria	Opheliidae	Thoracophelia_mucronata	marine	y	y
SRR1222216	Sedentaria	Orbiniidae	Phylo_foetida	marine		
SRR1221444	Sedentaria	Orbiniidae	Scoloplos_armiger	marine		
SRR1222288	Basal	Oweniidae	Owenia_fusiformis	marine	y	y
SRR2027869	Sedentaria	Parergodrilidae	Stygocapitella_subterranea	marine		
SRR2057036	Sedentaria	Pectinariidae	Pectinaria_gouldii	marine		y
Kocot et al. (2017)	Basal	Phascolosomatidae	Phascolosoma_agassizii	marine	y	
SRR1231565	Basal	Phascolosomatidae	Phascolosoma_granulatum	marine		
SRR2016923	Errantia	Phyllodocidae	Phyllodoce_medipapillata	marine	y	y
SRR2014676	Errantia	Polygordiidae	Polygordius_lacteus	marine	y	y
SRR1237766	Errantia	Polynoidae	Harmothoe_extenuata	marine	y	y
SRR2014684	Errantia	Protodrilidae	Protodrilus_adhaerens	marine	y	
SRR2016233	Errantia	Protodriloidae	Protodriloides_chaetifer	marine	y	y
SRR2017810	Sedentaria	Sabellariidae	Neosabellaria_cementarium	marine	y	y
SRR1232634	Sedentaria	Sabellariidae	Sabellaria_alveolata	marine		
SRR1231830	Sedentaria	Sabellidae	Megalomma_vesiculosum	marine	y	
SRR2005708	Sedentaria	Sabellidae	Sabella_pavonina	marine	y	
SRR2014689	Errantia	Saccocirridae	Saccocirrus_burchelli	marine	y	
SRR5590970	Sedentaria	Scalibregmatidae	Scalibregma_inflatum	marine	y	y
SRR516531	Sedentaria	Serpulidae	Pomatoceros_lamarckii	marine		y
SRR3556248	Sedentaria	Siboglinidae	Lamellibrachia_luymesi	marine	y	
SRR3574382	Sedentaria	Siboglinidae	Osedax_rubiplumus	marine	y	y
SRR3560108	Sedentaria	Siboglinidae	Sclerolinum_brattstromi	marine		
SRR3560206	Sedentaria	Siboglinidae	Siboglinum_fiordicum	marine		
SRR3571603	Sedentaria	Siboglinidae	Spirobrachia_sp	marine	y	
Kocot et al. (2017)	Sedentaria	Spionidae	Boccardia_proboscidea	marine	y	
SRR1222145	Sedentaria	Spionidae	Scolelepis_squamata	marine		
SRR2017800	Sedentaria	Sternaspidae	Sternaspsis_affinis	marine	y	y
SRR1224604	Errantia	Syllidae	Syllis_sp	marine		
SRR5590962	Sedentaria	Terebellidae	Lanice_conchilega	marine	y	y
SRR1237767	Errantia	Tomopteridae	Tomopteris_helgolandica	marine		

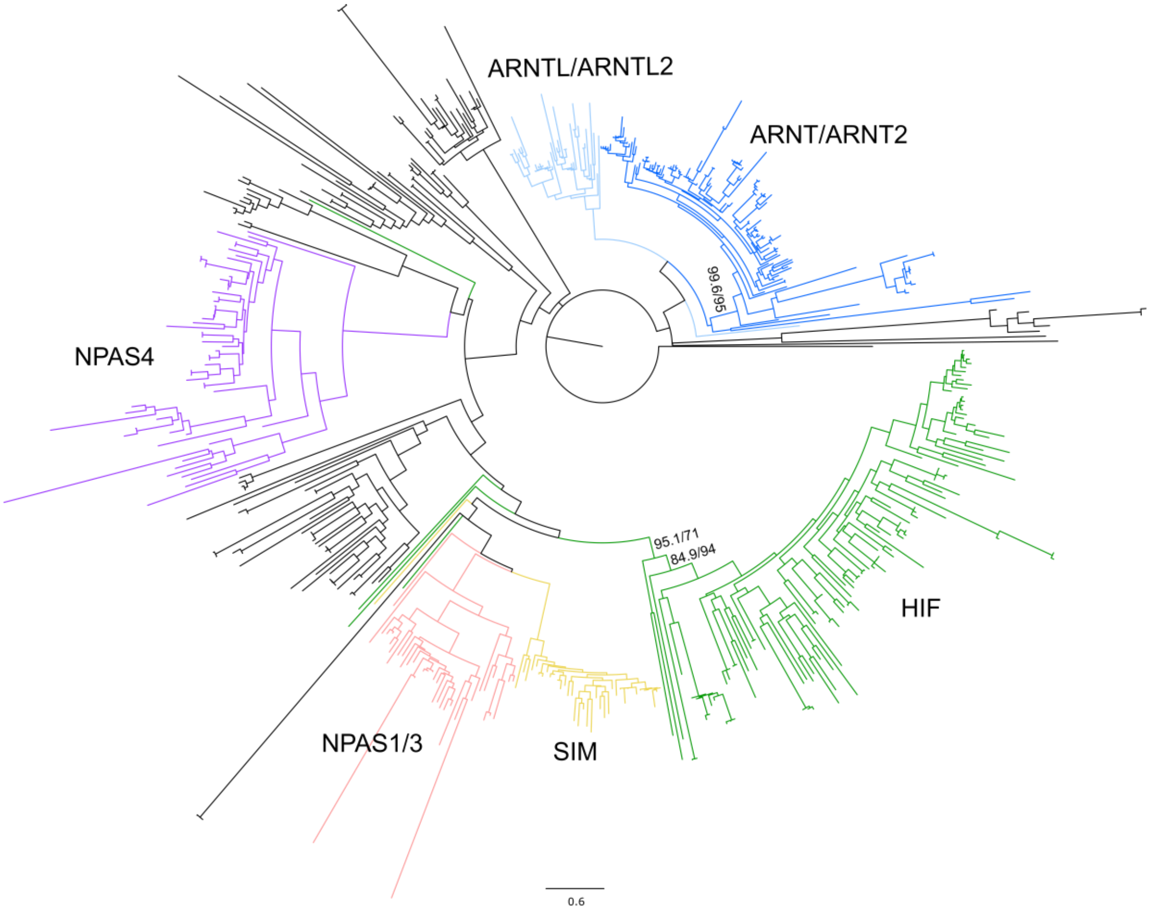


Figure 4.2: Maximum likelihood phylogenetic tree of annotated bHLH genes (Graham and Presnell 2017) and extracted annelid and *Crassostrea gigas* sequences. Only support values for branches relevant to HIFα and ARNT identification are shown and they are in the format of SH-aLRT/UFBoot. The black branches are annelid sequences that did not have clear membership to one of the known clades (NPAS1/3, NPAS4, SIM, ARNTL(2), ARNT2).

4.3.2. Gene-Specific Phylogenies

In the HIFα phylogenetic tree (Fig. 4.3), Clitellata is recovered with moderate support and Hirudinea is well supported (Figure 4.3). Families are well supported in cases of two or more species (Capitellidae, Chaetopteridae, Amphinomidae, Eunicidae, and Sabellidae) (Figure 4.3). A relationship between Capitellidae and Echiura is also well supported (Figure 4.3).

In the ARNT phylogenetic tree (Figure 4.4), families with two or more species (Nereididae, Chaetopteridae, Capitellidae, and Dinophilidae) are well supported (Figure 4.4). Most of Clitellata and Hirudinea form a well-supported clade (Figure 4.4), but two clitellates form a long branch close to the base of the tree (Figure 4.4). The relationship of Echiura with Capitellidae is poorly supported (Figure 4.4). Two notable clades that were mostly recovered are Sedentaria and Errantia/Basal groups. They are poorly supported and polyphyletic, but do reflect patterns in the known annelid phylogeny (Weigert and Bleidorn 2016) (Figure 4.4).

4.3.3. Domain Annotation

All of the known domains of HIFα and ARNT were recovered within Annelida (Graham and Presnell 2017) (Figures 4.3 and 4.4). Ten of the HIFα sequences lacked the HLH domain (Figure 4.3) as did twelve of the ARNT sequences (Figure 4.4). The HLH HMM was updated until HLH was no longer found. PAS domains were recovered in every species but in eight of the HIFα sequences and 16 of the ARNT sequences, only a single PAS domain was present. The PAS HMM was updated until there was no improvement in the resulting hits with no indication of a second PAS domain in the relevant sequences. CTAD was only recovered in three HIFα sequences and NTAD was found in 18 of the sequences (Figure 4.3). Both of the transactivation domain HMMs were updated until they were no longer found in the remaining sequences.

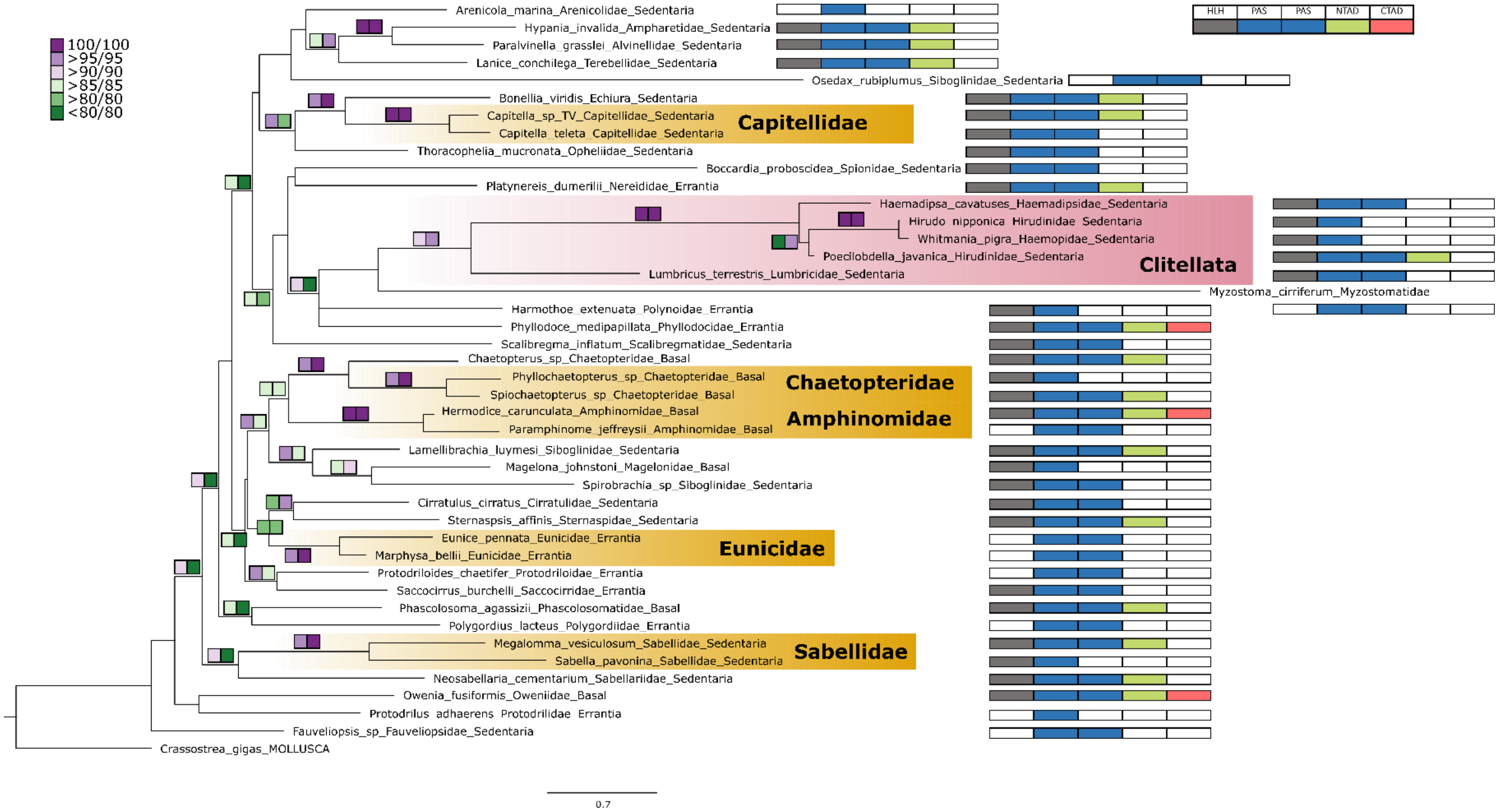

100/100
>95/95
>90/90
>85/85
>80/80
<80/80
HLH
PAS
PAS
NTAD
CTAD
Arenicola_marina_Arenicolidae_Sedentaria
Hypania_invalida_Ampharetidae_Sedentaria
Paralvinella_grasslei_Alvinellidae_Sedentaria
Lanice_conchilega_Terebellidae_Sedentaria
Osedax_rubiplumus_Siboglinidae_Sedentaria
Bonellia_viridis_Echiura_Sedentaria
Capitella_sp_TV_Capitellidae_Sedentaria
Capitella_teleta_Capitellidae_Sedentaria
Capitellidae
Thoracophelia_mucronata_Opheliidae_Sedentaria
Boccardia_proboscidea_Spionidae_Sedentaria
Platynereis_dumerilii_Nereididae_Errantia
Haemadipsa_cavatuses_Haemadipsidae_Sedentaria
Hirudo_nipponica_Hirudinidae_Sedentaria
Whitmania_pigra_Haemopidae_Sedentaria
Poecilobdella_javanica_Hirudinidae_Sedentaria
Lumbricus_terrestris_Lumbricidae_Sedentaria
Clitellata
Myzostoma_cirriferum_Myzostomatidae
Harmothoe_extenuata_Polynoidae_Errantia
Phyllodoce_medipapillata_Phyllodocidae_Errantia
Scalibregma_inflatum_Scalibregmatidae_Sedentaria
Chaetopterus_sp_Chaetopteridae_Basal
Phyllochaetopterus_sp_Chaetopteridae_Basal
Spiochaetopterus_sp_Chaetopteridae_Basal
Chaetopteridae
Hermodice_carunculata_Amphinomidae_Basal
Paramphinome_jeffreysii_Amphinomidae_Basal
Amphinomidae
Lamellibrachia_luymesi_Siboglinidae_Sedentaria
Magelona_johnstoni_Magelonidae_Basal
Spirobrachia_sp_Siboglinidae_Sedentaria
Cirratulus_cirratus_Cirratulidae_Sedentaria
Sternaspsis_affinis_Sternaspidae_Sedentaria
Eunice_pennata_Eunicidae_Errantia
Marphysa_bellii_Eunicidae_Errantia
Eunicidae
Protodriloides_chaetifer_Protodriloidae_Errantia
Saccocirrus_burchelli_Saccocirridae_Errantia
Phascolosoma_agassizii_Phascolosomatidae_Basal
Polygordius_lacteus_Polygordiidae_Errantia
Megalomma_vesiculosum_Sabellidae_Sedentaria
Sabella_pavonina_Sabellidae_Sedentaria
Sabellidae
Neosabellaria_cementarium_Sabellariidae_Sedentaria
Owenia_fusiformis_Oweniidae_Basal
Protodrilus_adhaerens_Protodrilidae_Errantia
Fauveliopsis_sp_Fauveliopsidae_Sedentaria
Crassostrea_gigas_MOLLUSCA
0.7

Figure 4.3: Maximum likelihood phylogenetic tree of annelid HIFα genes. Branch support values are in the format of SH-aLRT/UFBoot. Presence of protein domains is indicated following the key in the top right corner.

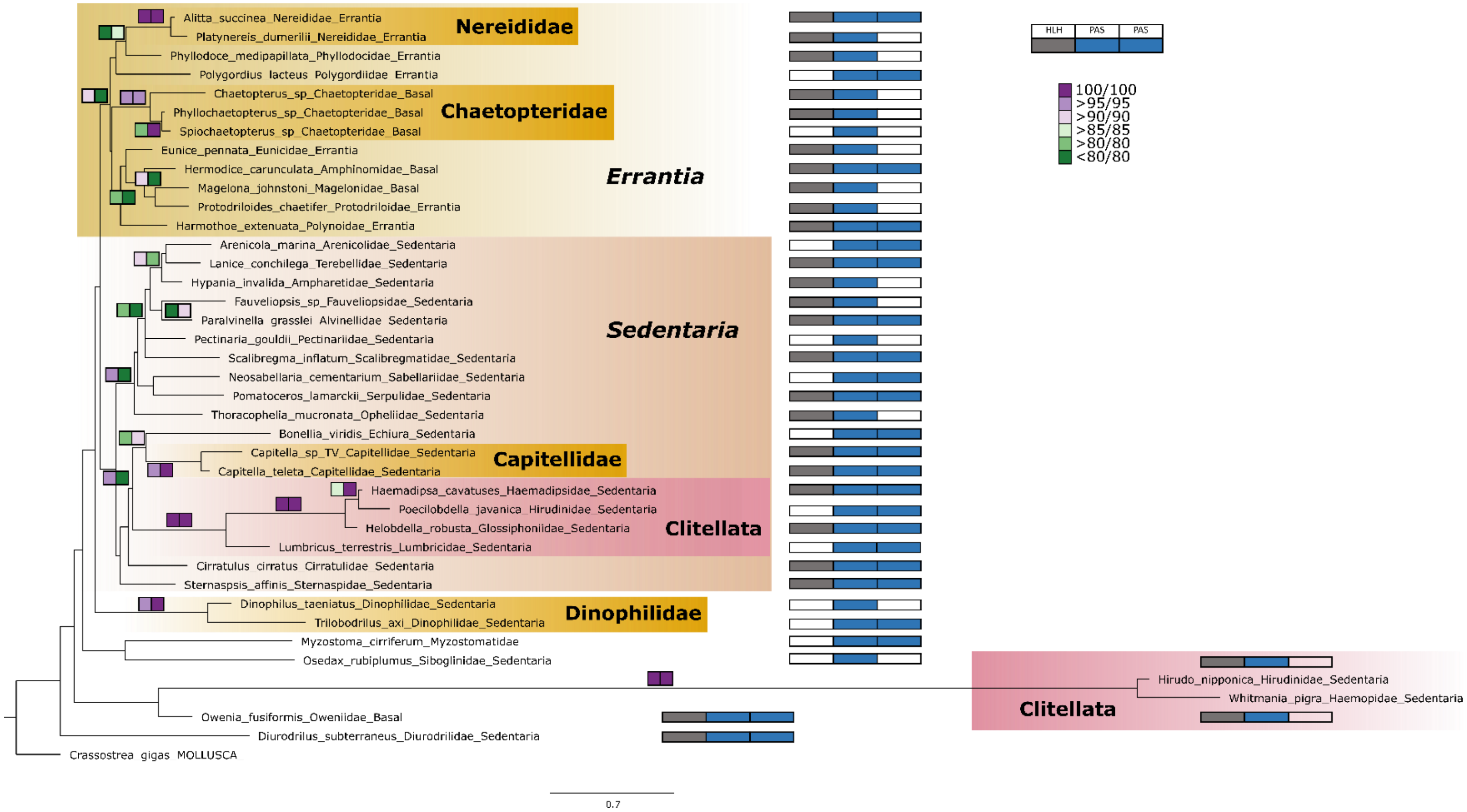

Alitta_succinea_Nereididae_Errantia
Platynereis_dumerilii_Nereididae_Errantia
Nereididae
Phyllodoce_medipapillata_Phyllodocidae_Errantia
Polygordius_lacteus_Polygordiidae_Errantia
Chaetopterus_sp_Chaetopteridae_Basal
Phyllochaetopterus_sp_Chaetopteridae_Basal
Spiochaetopterus_sp_Chaetopteridae_Basal
Chaetopteridae
Eunice_pennata_Eunicidae_Errantia
Hermodice_carunculata_Amphinomidae_Basal
Magelona_johnstoni_Magelonidae_Basal
Protodriloides_chaetifer_Protodriloidae_Errantia
Harmothoe_extenuata_Polynoidae_Errantia
Errantia
Arenicola_marina_Arenicolidae_Sedentaria
Lanice_conchilega_Terebellidae_Sedentaria
Hypania_invalida_Ampharetidae_Sedentaria
Fauveliopsis_sp_Fauveliopsidae_Sedentaria
Paralvinella_grasslei_Alvinellidae_Sedentaria
Pectinaria_gouldi_Pectinariidae_Sedentaria
Scalibregma_inflatum_Scalibregmatidae_Sedentaria
Neosabellaria_cementarium_Sabellariidae_Sedentaria
Pomatoceros_lamarckii_Serpulidae_Sedentaria
Thoracophelia_mucronata_Opheliidae_Sedentaria
Sedentaria
Bonellia_viridis_Echiura_Sedentaria
Capitella_sp_TV_Capitellidae_Sedentaria
Capitella_teleta_Capitellidae_Sedentaria
Capitellidae
Haemadipsa_cavatuses_Haemadipsidae_Sedentaria
Poecilobdella_javanica_Hirudinidae_Sedentaria
Helobdella_robusta_Glossiphoniidae_Sedentaria
Lumbricus_terrestris_Lumbricidae_Sedentaria
Clitellata
Cirratulus_cirratus_Cirratulidae_Sedentaria
Sternaspsis_affinis_Sternaspidae_Sedentaria
Dinophilus_taeniatus_Dinophilidae_Sedentaria
Trilobodrilus_axi_Dinophilidae_Sedentaria
Dinophilidae
Myzostoma_cirriferum_Myzostomatidae
Osedax_rubiplumus_Siboglinidae_Sedentaria
Hirudo_nipponica_Hirudinidae_Sedentaria
Whitmania_pigra_Haemopidae_Sedentaria
Clitellata
Owenia_fusiformis_Oweniidae_Basal
Diurodrilus_subterraneus_Diurodrilidae_Sedentaria
Crassostrea_gigas_MOLLUSCA
0.7
HLH
PAS
PAS
100/100
>95/95
>90/90
>85/85
>80/80
<80/80

Figure 4.4: Maximum likelihood phylogenetic tree of annelid ARNT genes. Branch support values are in the format of SH-aLRT/UFBoot. Presence of protein domains is indicated following the key in the top right corner.

4.4. Discussion

Our findings suggest that domain structure of HIFα and ARNT varies greatly across Annelida. Identification of these genes in annelids was only possible through phylogenetic analysis, as BLAST searches using reference sequences as queries were not specific enough and detected related other bHLH+PAS genes (ARNT2, ARNTL(2), SIM 1/2, and NPAS 1/3/4) beside the target genes.

It is likely that BLAST searches primarily captured the highly conserved PAS domains. For example, the first *hmmsearch* of the Pfam PAS HMM (with no annelids in the profile) returned 42 significant hits for HIFα. On the other hand, HLH domains have been found to be diverse (Graham and Presnell 2017) and the annelids are not an exception. The first *hmmsearch* of the Pfam HLH HMM (which only has *Helobdella robusta* in the profile) returned three significant hits for HIFα. The lack of the HLH domain in some species is likely an artifact resulting from removal of low-quality data. Examination of the multiple sequence alignment indicates total lack of data in this region for the respective species. The alternative of true loss of the HLH domain is not probable as this domain is responsible for DNA binding activity.

There is no clear relationship between transactivation domain presence and phylogeny (Figure 4.3). NTAD is present throughout Annelida but lost in several lineages (Figure 4.3). CTAD has a very limited occurrence and the three species it was found in are not closely related (Figure 4.3). However, *Hermodice carunculata* and *Owenia fusiformis* both have basal positions in the current annelid phylogeny. *Phyllodoce medipapillata* is in the Errantia grouping. This may indicate loss of the CTAD domain within Sedentaria (GIGA 2014, Weigert and Bleidorn 2016, Helm et al. 2018). While these domains are presumed to be important for HIF activity and specificity (Lando et al. 2002, Hu et al. 2007), they may not be entirely necessary for annelids, and invertebrates in general (Graham and Presnell 2017), or they are just so diverged that we could not detect them.

Capitellids and echiurans are frequently cited as the sister group to clitellates (GIGA 2014, Weigert and Bleidorn 2016, Helm et al. 2018). A relationship between capitellids and echiurans is present in both genes, but their status as sister to clitellates is not supported by HIFα (Figure 4.3) and only poorly supported by ARNT (Figure 4.4). Clitellata is derived and moderately supported in both gene trees (Figures 4.3 and 4.4). The long branch for the clade containing *Hirudo nipponica* and *Whitmania pigra* outside of Clitellata in the ARNT phylogeny may be caused by extreme diversification and long-branch attraction (Bergsten 2005), but we did not evaluate this. It is noteworthy that both *H. nipponica* and *W. pigra* are freshwater species, but so are *Helobdella* and *Poecilobdella*. Recent work on clitellates indicates duplication events of Na+/K+ - ATPases with freshwater colonization (Horn et al. 2019). Further work with a better representation of Clitellata would reveal any oxygen-sensing pathway novelties associated with transitions to freshwater and terrestrial environments.

ARNT recovered two large clades, Sedentaria and Errantia/Basal groups (Figure 4.4). While support is low, it is congruent with current annelid phylogenies (GIGA 2014, Weigert and Bleidorn 2016, Helm et al. 2018). This pattern was not recovered in HIFα (Figure 4.3) and likely due to HIFα being under less evolutionarily constraint. ARNT takes part in various other processes from HIF dimerization (Graham and Presnell 2017).

In conclusion, we have demonstrated high diversity in the transcription factor HIF across Annelida. Its heterodimer ARNT is more conserved and reflects the current hypotheses of annelid phylogeny more closely. Protein domains were recovered in varying degrees and it is not clear if this is due to domain loss, extreme divergence, or loss of low-quality data during transcriptome construction. Through annotation of the domains, HMM profiles were updated to better reflect diversity in Annelida. Use of these updated models for further gene discovery with studies on the activity of this transcription factor will continue to shed light on the evolution of this important cellular pathway.

4.5. References

Bergsten J (2005) A review of long-branch attraction. Cladistics 21(2): 163-193

Berner RA, VandenBrooks JM, and Ward PD (2007) Oxygen and evolution. Science 316: 557-558

Cai X, Huang Y, Zhang X, Wang S, Zou Z, Wang G, Wang Y, and Zhang Z (2014) Cloning, characterization, hypoxia and heat shock response of hypoxia inducible factor-1 (HIF-1) from the small abalone Haliotis diversicolor. Gene 534: 256-264. doi http://dx.doi.org/10.1016/j.gene.2013.10.048

Díaz RJ and Rosenberg R (1995) Marine benthic hypoxia: a review of its ecological effects and the behavioural responses of benthic macrofauna. Oceanography and Marine Biology: an Annual Review 33: 245-303

Díaz RJ and Rosenberg R (2008) Spreading dead zones and consequences for marine ecosystems. Science 321: 926-929. doi 10.1126/science.1156401

Dudley R (1998) Atmospheric oxygen, giant Paleozoic insects and the evolution of aerial locomotor performance. Journal of Experimental Biology 201: 1043-1050

Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. Genome Informatics 2009: 205-211

Epstein ACR, Gleadle JM, McNeill LA, Hewitson KS, O'Rourke J, Mole DR, Mukherji M, Metzen E, Wilson MI, Dhanda A, Tian Y-M, Masson N, Hamilton DL, Jaakkola P, Barstead R, Hodgkin J, Maxwell PH, Pugh CW, Schofield CJ, and Ratcliffe PJ (2001) C. elegans EGL-9 and Mammalian Homologs Define a Family of Dioxygenases that Regulate HIF by Prolyl Hydroxylation. Cell 107: 43-54. doi http://dx.doi.org/10.1016/S0092-8674(01)00507-4

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39(4): 783-791

Fortunato SAV, Vervoort M, Adamski M, and Adamska M (2016) Conservation and divergence of bHLH genes in the calcisponge Sycon ciliatum. EvoDevo 7: 23. doi 10.1186/s13227-016-0060-8

Friedrich J, Janssen F, Aleynik D, Bange HW, Boltacheva N, Çagatay MN, Dale AW, Etiope G, Erdem Z, Geraga M, Gilli A, Gomoiu MT, Hall POJ, Hansson D, He Y, Holtappels M, Kirf MK, Kononets M, Konovalov S, Lichtschlag A, Livingstone DM, Marinaro G, Mazlumyan S, Naeher S, North RP, Papatheodorou G, Pfannkuche O, Prien R, Rehder G, Schubert CJ, Soltwedel T, Sommer S, Stahl H, Stanev EV, Teaca A, Tengberg A, Waldmann C, Wehrli B, and Wenzhöfer F (2014) Investigating hypoxia in aquatic environments: diverse approaches to addressing a complex phenomenon. Biogeosciences 11: 1215-1259. doi 10.5194/bg-11-1215-2014

Giannetto A, Maisano M, Cappello T, Oliva S, Parrino V, Natalotto A, De Marco G, Barberi C, Romeo O, Mauceri A, and Fasulo S (2015) Hypoxia-Inducible Factor α and Hif-prolyl Hydroxylase Characterization and Gene Expression in Short-Time

Air-Exposed Mytilus galloprovincialis. Marine Biotechnology 17: 768-781. doi 10.1007/s10126-015-9655-7

GIGA (2014) The Global Invertebrate Genomics Alliance (GIGA): Developing community resources to study diverse invertebrate genomes. Journal of Heredity 105(1): 1-18. doi 10.1093/jhered/est084

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, and Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29(7): 644-52. doi 10.1038/nbt.1883

Graham A and Presnell J (2017) Hypoxia inducible factor (HIF) transcription factor family expansion, diversification, divergence and selection in eukaryotes. PLOS ONE 12(6): e0179545. https://doi.org/10.1371/journal.pone.0179545

Graham JB, Aguilar NM, Dudley R, and Gans C (1995) Implications of the late Palaeozoic oxygen pulse for physiology and evolution. Nature 375: 117-120

Guindon S, Dufayard J, Lefort J, Anisimova M, Hordijk W, and Gascuel O (2010) New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology 59(3): 307-321. doi https://doi.org/10.1093/sysbio/syq010

Gyoja F (2014) A genome-wide survey of bHLH transcription factors in the Placozoan Trichoplax adhaerens reveals the ancient repertoire of this gene family in metazoan. Gene 542: 29-37. doi http://dx.doi.org/10.1016/j.gene.2014.03.024

Haas B and Papanicolaou A (n.d.) Transdecoder (find coding regions within transcripts). Retrieved from http://transdecoder.github.io

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, and Regev A (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols 8(8): 1494-512. doi 10.1038/nprot.2013.084

Heinrich EC, Farzin M, Klok CJ, and Harrison JF (2011) The effect of developmental stage on the sensitivity of cell and body size to hypoxia in Drosophila melanogaster. Journal of Experimental Biology 214: 1419-1427

Helm C, Beckers P, Bartolomaeus T, Drukewitz SH, Kourtesis I, Weigert A, Purschke G, Worsaae K, Struck TH, and Bleidorn C (2018) Convergent evolution of the ladder-like ventral nerve cord in Annelida. Frontiers in Zoology 15: 36. doi 10.1186/s12983-018-0280-y

Horn KM, Williams BW, Erséus C, Halanych KM, Santos SR, Châtelliers M, and Anderson FE (2019) Na+/K+-ATPase gene duplications in clitellate annelids are associated with freshwater colonization. Journal of Evolutionary Biology 32: 580-591. doi https://doi.org/10.1111/jeb.13439

Hourdez S and Lallier F (2007) Adaptations to hypoxia in hydrothermal-vent and cold-seep invertebrates. Reviews in Environmental Science and Bio/Technology 6: 143-159. doi 10.1007/s11157-006-9110-3

Hu C-J, Sataur A, Wang L, Chen H, and Simon MC (2007) The N-terminal transactivation domain confers target gene specificity of hypoxia-inducible factors HIF-1α and HIF-2α. Molecular biology of the cell 18: 4528-4542. doi https://doi.org/10.1091/mbc.E06-05-0419

Jiang H, Guo R, and Powell-Coffman JA (2001) The Caenorhabditis elegans hif-1 gene encodes a bHLH-PAS protein that is required for adaptation to hypoxia. Proceedings of the National Academy of Sciences 98: 7916-7921

Kaelin Jr WG (2005) The von Hippel-Lindau protein, HIF hydroxylation, and oxygen sensing. Biochemical and Biophysical Research Communications 338: 627-638. doi 10.1016/j.bbrc.2005.08.165

Kaelin Jr WG and Ratcliffe PJ (2008) Oxygen Sensing by Metazoans: The Central Role of the HIF Hydroxylase Pathway. Molecular Cell 30: 393-402. doi http://dx.doi.org/10.1016/j.molcel.2008.04.009

Kalyaanamoorthy S, Minh B, Wong T, Haeseler A, and Jermiin LS (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods 14: 587-589. doi https://doi.org/10.1038/nmeth.4285

Katoh K and Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution 30(4): 772-780. doi https://doi.org/10.1093/molbev/mst010

Kawabe S and Yokoyama Y (2012) Role of Hypoxia-Inducible Factor α in Response to Hypoxia and Heat Shock in the Pacific Oyster Crassostrea gigas. Marine Biotechnology 14: 106-119. doi 10.1007/s10126-011-9394-3

Kocot KM, Struck TH, Merkel J, Waits DS, Todt C, Brannock PM, Weese DA, Cannon JT, Moroz LL, Lieb B, and Halanych KM (2017) Phylogenomics of Lophotrochozoa with Consideration of Systematic Error. Systematic Biology 66(2): 256-282. https://doi.org/10.1093/sysbio/syw079

Kristensen E (1983) Ventilation and oxygen uptake by three species of Nereis (Annelida: Polychaeta). I. Effects of hypoxia. Marine Ecology Progress Series 12: 289-297

Lando D, Peet DJ, Whelan DA, Gorman JJ, and Whitelaw ML (2002) Asparagine hydroxylation of the HIF transactivation domain: a hypoxic switch. Science 295: 858-861. doi https://doi.org/10.1126/science

Ledent V, Paquet O, and Vervoort M (2002) Phylogenetic analysis of the human basic helix-loop-helix proteins. Genome Biology 3(6): research0030.1-0030.18

Leonarz C, Coleman M, Boleininger A, Schierwater B, Ratcliffe P, and Schofield C (2011) The hypoxia-inducible transcription factor pathway regulates oxygen sensing in the simplest animal, Trichoplax adherens. EMBO Rep 12: 63-70

Levin LA (2003) Oxygen minimum zone benthos: adaptation and community response to hypoxia. Oceanography and Marine Biology: an Annual Review 41: 1-45

Melzner F, Thomsen J, Koeve W, Oschlies A, Gutowska M, Bange H, Hansen H, and Körtzinger A (2013) Future ocean acidification will be amplified by hypoxia in coastal habitats. Marine Biology 160: 1875-1888. doi 10.1007/s00227-012-1954-1

Minh BQ, Nguyen M, and Haeseler A (2013) Ultrafast Approximation for Phylogenetic Bootstrap. Molecular Biology and Evolution 30(5): 1188-1195. doi https://doi.org/10.1093/molbev/mst024

Nambu JR, Chen W, Hu S, and Crews ST (1996) The Drosophila melanogaster similar bHLH-PAS gene encodes a protein related to human hypoxia-inducible factor 1α and Drosophila single-minded. Gene 172: 249-254. doi http://dx.doi.org/10.1016/0378-1119(96)00060-1

Nguyen L, Schmidt HA, Haeseler A, and Minh BQ (2015) IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution. 32(1): 268-274 doi https://doi.org/10.1093/molbev/msu300

Philippon H, Souvane A, Brochier-Armanet C, and Perrière G (2017) IsoSel: Protein Isoform Selector for phylogenetic reconstructions. PLOS ONE 12(3): e0174250. https://doi.org/10.1371/journal.pone.0174250

Piontkivska H, Chung JS, Ivanina AV, Sokolov EP, Techa S, and Sokolova IM (2011) Molecular characterization and mRNA expression of two key enzymes of hypoxia-sensing pathways in eastern oysters Crassostrea virginica (Gmelin): Hypoxia-inducible factor α (HIF-α) and HIF-prolyl hydroxylase (PHD). Comparative Biochemistry and Physiology Part D: Genomics and Proteomics 6: 103-114. doi http://dx.doi.org/10.1016/j.cbd.2010.10.003

Pörtner HO, Langenbuch M, and Michaelidis B (2005) Synergistic effects of temperature extremes, hypoxia, and increases in CO2 on marine animals: From Earth history to global change. Journal of Geophysical Research: Oceans 110: C09S10. doi 10.1029/2004JC002561

Qu F, Nunnally C, and Rowe GT (2015) Polychaete annelid biomass size spectra: The effects of hypoxia stress. Journal of Marine Biology 2015: 983521. doi 10.1155/2015/983521

Rabalais NN, Turner RE, and Wiseman WJ (2002) Gulf of Mexico Hypoxia, a.k.a. "The Dead Zone". Annual Review of Ecology and Systematics 33: 235-263

Rambaut A (2012) FigTree. Retrieved from http://tree.bio.ed.ac.uk/software/figtree/

Rosa R and Seibel BA (2008) Synergistic effects of climate-related variables suggest future physiological impairment in a top oceanic predator. Proceedings of the National Academy of Sciences 105: 20776-20780. doi 10.1073/pnas.0806886105

Rosa R, Trübenbach K, Repolho T, Pimentel M, Faleiro F, Boavida-Portugal J, Baptista M, Lopes VM, Dionísio G, Leal MC, Calado R, and Pörtner HO (2013) Lower hypoxia thresholds of cuttlefish early life stages living in a warm acidified ocean. Proceedings of the Royal Society of London B: Biological Sciences 280: 20131695. doi https://doi.org/10.1098/rspb.2013.1695

Satou Y, Imai KS, Levine M, Kohara Y, Rokhsar D, and Satoh N (2003) A genomewide survey of developmentally relevant genes in Ciona intestinalis. Development Genes and Evolution 213: 213-221. doi 10.1007/s00427-003-0319-7

Seaver EC, Thamm K, and Hill SD (2005) Growth patterns during segmentation in the two polychaete annelids, Capitella sp. I and Hydroides elegans: comparisons at distinct life history stages. Evolution and Development 7(4): 312-326. doi https://doi.org/10.1111/j.1525-142X.2005.05037.x

Simionato E, Ledent V, Richards G, Thomas-Chollier M, Kerner P, Coornaert D, Dengen BM, and Vervoort M (2007) Origin and diversification of the basic helix-loop-helix gene family in metazonas: insights from comparative genomics. BMC Evolutionary Biology 7: 33. doi 10.1186/1471-2148-7-33

Sun S, Xuan F, Fu H, Ge X, Zhu J, Qiao H, Jin S, and Zhang W (2016) Molecular characterization and mRNA expression of hypoxia inducible factor-1 and cognate inhibiting factor in Macrobrachium nipponense in response to hypoxia. Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology 196-197: 48-56. doi http://dx.doi.org/10.1016/j.cbpb.2016.02.002

Vaquer-Sunyer R and Duarte CM (2008) Thresholds of hypoxia for marine biodiversity. Proceedings of the National Academy of Sciences 105: 15452-15457. doi 10.1073/pnas.0803833105

Wang GL, Jiang BH, Rue EA, and Semenza GL (1995) Hypoxia-inducible factor 1 is a basic-helix-loop-helix-PAS heterodimer regulated by cellular O2 tension. Proceedings of the National Academy of Sciences 92: 5510-5514

Waterhouse AM, Procter JB, Martin DMA, Clamp M, and Barton GJ (2009) Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191. doi:10.1093/bioinformatics/btp033

Weigert A and Bleidorn C (2016) Current status of annelid phylogeny. Organisms Diversity & Evolution 16: 345-362. doi https://doi.org/10.1007/s13127-016-0265-7

Yan J, Ma Z, Xu X, and Guo A (2014) Evolution, functional divergence and conserved exon-intron structure of bHLH/PAS gene family. Molecular Genetics and Genomics 289: 26-36. doi 10.1007/s00438-013-0786-0

Yates AD et al. (2020) Ensembl. Nucleic Acids Research 48(D1): D682-D688/ https://doi.org/10.1093/nar/gkz966

5. CONCLUSIONS

While the results of each chapter of this dissertation represent a significant contribution, they also contain interconnecting relevance. Gene evolution and function impacts the ecology of organisms by restricting and/or expanding their ecological niche. The interaction of these aspects of their biology contribute to patterns of speciation and, ultimately, a deeper understanding of evolutionary history.

I have demonstrated that the Gulf of Mexico (GoM) contains many unknown species of *Capitella* with the discovery of at least five new species and the first report of *Capitella nonatoi*, originally described from Brazil. This is significant in multiple ways. First, it simply adds to our understanding of the evolutionary history of *Capitella*. It will take substantial effort to ultimately determine the total number of species contained in this genus and their phylogeny. For example, I found three sympatric species at Apollo Beach Preserve, Tampa Bay, Florida. One of the species was also found throughout the GoM while the range of the other two appears very restricted, since they were recovered only from that field site. This indicates that a very targeted sampling effort, inclusive of different sediment depths, is necessary to recover all *Capitella*. Beyond estuarine/shallow-water habitats, there are still deep-sea and specialized habitats (whale-bone and wood falls and squid egg masses) that likely harbor additional diversity.

Second, an expanded understanding of *Capitella* phylogeny could lead to the eventual establishment of new spiralian evo-devo models. *Capitella teleta* is very well studied and has helped to fill gaps in our understanding of protostome and bilaterian evolution; sequencing annelid genomes revealed a strong likeness to non-protostomes and further uncovered characteristics of the metazoan last common ancestor (Simakov et al.

2012). While *Capitella* cryptic species are presumed closely related, and thus assigned to the same genus, they can be notably diverse in their chromosome diploid number, ranging 12-26 with only 19 species evaluated (Grassle et al. 1987). This suggest many genome rearrangements during the evolution of *Capitella*. Novel insights that can come from evo-devo studies comparing species of *Capitella* include the role of genome rearrangement events in speciation, regeneration rates of posterior segments, early segmentation patterns during embryonic development, and spiralian development (e.g. cell fate specification, development of eye spots, and nervous system development).

Third, interesting questions about population dynamics are raised with the discovery of sympatric species. For example, discovery of *Capitella nonatoi* in the GoM and its dominance in the May samplings contrasts with *Capitella* sp. TV that was found at other times. This may represent an instance of seasonal dynamics as was suggested in the originally detected cryptic *Capitella* species in Massachusetts (Grassle and Grassle 1976). There may also be resource partitioning for *Capitella* in the GoM as cryptic species have been found inhabiting different sediment depths in the Mediterranean (Gamenick et al. 1998). Differences in larval development, maternal investments (i.e. presence/absence of yolk reserve), and timing of reproductive events (Méndez et al. 2000) between species would allow larvae to functionally inhabit different niches and reduce competition (Levin and Huggett 1990).

Finally, resolved species boundaries will facilitate more focused use of *Capitella* as pollution indicators. It is well established that *Capitella* have a lot of potential use as environmental sentinels. There has even been evidence that they can degrade polycyclic aromatic hydrocarbons (Li et al. 2004). This indicates that they may play a role in bioremediation of organically-polluted sites, contributing to their known r-selected nature. However, their capacity in this role may vary as cryptic species have shown physiological differences, tolerating different sulphide levels (Gamenick et al. 1998). Species delimitation studies such as this one will make research on these topics more targeted and repeatable.

Regarding contribution to their ecology, I showed that, among the models evaluated, there is not a single strategy that works well for capitellids, despite them filling a similar ecological niche (burrowing, deposit feeders). However, Hurdle and GAM-Tweedie models performed well overall, possibly resulting from their handling of zeros. Bay segment was important overall, indicating that unexplained processes (water circulation patterns, anthropogenic activity, etc.) within these regions of the bay may have a strong influence over capitellid occurrences and distributions. For *C. capitata* complex specifically, depth was most important.

These results are significant because they 1) provide the first comparison of this kind for modeling capitellid species abundance and 2) generate hypotheses for future work on *Capitella* ecology. For example, the importance of depth in such a shallow system was surprising. These model results suggest the hypothesis that *C. capitata* complex is most abundant in shallower, near-shore waters, and this could be tested.

Furthermore, finding five new species in the GoM, three in Tampa Bay, Florida, raises the question as to whether the results for *C. capitata* complex are reliable. It is likely that the referenced three species, among others, are present within these samples. However, the specimens were preserved for morphology and not DNA integrity, making DNA barcoding difficult. Using Ancient DNA extraction methods in the future could turn this dataset into a useful resource for species delimitation in the GoM and, ultimately, refine our understanding of their ecology.

I did not directly study *Capitella* physiology but evaluated the phylogeny of two genes that are crucial in cellular oxygen-sensing, hypoxia inducible factor alpha (HIFα) and aryl receptor nuclear transferase (ARNT). Under hypoxia these two genes dimerize, forming a transcription factor that initiates cellular response. Therefore, understanding their evolution across Annelida might be informative to understand evolutionary pathways in annelids.

I found that these genes do not mirror the known annelid phylogeny and reflect complex evolutionary pathways that remain to be understood. Annotation of domains revealed that annelids are quite diverse, particularly in the helix-loop-helix region. Capitellidae was well supported in both phylogenies and their sister relationship to Echiura, which is commonly recovered in annelid phylogenies, was supported. Additionally, capitellids and echiurans were sister to most of the clitellates in the ARNT phylogeny, which is also recovered in annelid phylogenies.

These findings are significant because they 1) represent the first analysis of HIFα and ARNT evolution across Annelida, 2) resulted in updated Hidden Markov Models (HMMs) of HIFα and ARNT domains to better reflect diversity within Annelida, and 3) contribute omics data of another *Capitella* species. The original HMM profiles had little or no annelid representation. Use of the updated profiles can lead to better discovery of HIFα and ARNT across Annelida and further resolution of their phylogenies.

Capitella are found in low-oxygen environments such as hydrothermal vents of the Mediterranean and development of larvae and juveniles has shown resilience to hypoxia

(Gamenick et al. 1998, Pechenik et al. 2016). Having many species from diverse habitats, being easily amenable to culture, and reproducing multiple times a year make *Capitella* a good candidate system for studying functional response to hypoxia across life stages. This and other studies on the functional response to varying oxygen levels across Annelida, will shed light on the complex evolutionary history of the HIF transcription factor.

In conclusion, this dissertation furthers understanding of three areas of *Capitella* biology. Knowledge of oxygen sensing gene evolution can give perspective on species distributions and their environmental drivers. Both gene evolution and ecology play a role in speciation and can shed light on observed patterns of phylogeography. Continued work on these aspects will lead to better utilization of *Capitella* as environmental sentinels, new insights into spiralian evolution, and understanding of annelid evolution as a whole.

5.1. References

Gamenick I, Vismann B, Grieshaber MK, and Giere O (1998) Ecophysiological differentiation of Capitella capitata (Polychaeta). Sibling species from different sulfidic habitats. Marine Ecology Progress Series 175: 155-166

Grassle JP and Grassle JF (1976) Sibling species in the marine pollution indicator Capitella (Polychaeta). Science 192(4239): 567-569

Grassle JP, Gelfman CE, and Mills SW (1987) Karyotypes of Capitella sibling species, and of several species in the related genera Capitellides and Capitomastus (Polychaeta). Bulletin of the Biological Society of Washington 7: 77-88

Levin LA and Huggett DV (1990) Implications of Alternative Reproductive Modes for Seasonality and Demography in an Estuarine Polychaete. Ecology 71(6): 2191-2208

Li B, Bisgaard HC, and Forbes VE (2004) Identification and expression of two novel cytochrome P450 genes, belonging to CYP4 and a new CYP331 family, in the polychaete Capitella capitata sp. I. Biochemical and Biophysical Research Communications 325: 510-517

Méndez N, Linke-Gamenick I, and Forbes V (2000) Variability in reproductive mode and larval development within the Capitella capitata species-complex. Invertebrate Reproduction and Development 38: 131-142

Pechenik JA, Chaparro OR, Pilnick A, Karp M, Acquafredda M, and Burns R (2016) Effects of Embryonic Exposure to Salinity Stress or Hypoxia on Post-metamorphic Growth and Survival of the Polychaete Capitella teleta. The Biological Bulletin 231: 103-112

Simakov O, Marletaz F, Cho SJ, Edsinger-Gonzales E, Havlak P, Hellsten U, Kuo DH, Larsson T, Lv J, Arendt D, Savage R, Osoegawa K, de Jong P, Grimwood J, Chapman JA, Shapiro H, Aerts A, Otillar RP, Terry AY, Boore JL, et al. (2013) Insights into bilaterian evolution from three spiralian genomes. Nature 493(7433): 526-531

APPENDIX A

COMPARATIVE SPECIES ABUNDANCE MODELING OF CAPITELLIDAE (ANNELIDA) IN TAMPA BAY, FLORIDA

Table A.1: Model calibration metrics between observed values and model predicted values. Cells highlighted in blue indicate that those values are not present in the graphs of Figures 7 and 8. **Apparent (not corrected for optimism) statistics reported.

	Pearson Correlation	Spearman Rank Correlation	Slope	Intercept	Root Mean Square Error	Average Error
Capitella capitata						
Poisson	-0.06	0.37	1.03	-0.05	7.72E+18	6.25E+17
Negative Binomial	0.18	0.40	0.40	0.78	-815.58	-61.30
Tweedie	0.11	0.40	0.76	0.35	-133.54	32.04
Zero-Inflated Poisson	-0.11	0.26	-7.54E-08	1.54	-3.64E+83	-1.04E+82
Hurdle	0.31	0.40	0.95	0.04	-40.01	0.41
Boosted Regression Tree	0.15	0.39	1.44	-0.57	7.76	2.21
*Capitella aciculata***						
Poisson	1.00	0.31	1.01	0.00	0.05	0.01
Negative Binomial	0.05	0.14	0.03	0.12	5.09	0.52
Tweedie	0.21	0.14	0.79	0.05	2.66	0.21
Zero-Inflated Poisson	0.00	-0.01	-5.12E-08	0.13	7.50E+04	2.21E+03
Hurdle	0.00	0.08	-5.00E-08	0.13	1.26E+05	6.08E+03
Boosted Regression Tree	0.09	0.08	238.71	-30.81	2.71	0.26
Capitella jonesi						
Poisson	-0.22	0.09	2.04	-0.15	1.96	0.25
Negative Binomial	-0.05	0.14	0.47	0.07	1.55	0.35
Tweedie	-0.21	0.10	0.97	0.01	1.55	0.26
Zero-Inflated Poisson	-0.02	0.02	-1.04E-09	0.14	3.23E+06**	8.82E+282
Hurdle	0.06	0.17	0.51	0.06	-2.52E+09	2.49E+07
Boosted Regression Tree	0.00	0.10	4.30	-0.36	1.70	0.24

Table A.1 Continued

	Pearson Correlation	Spearman Rank Correlation	Slope	Intercept	Root Mean Square Error	Average Error
Heteromastus filiformis						
Poisson	0.02	0.32	1.11	-0.06	2.50**	2.99E+177
Negative Binomial	0.15	0.34	0.21	0.36	7.62E+05	-2.25E+04
Tweedie	0.03	0.33	0.81	0.10	-1.20E+06	2.14E+04
Zero-Inflated Poisson	0.40	0.30	0.98	-0.02	-4.01E+108	-1.36E+107
Hurdle	0.35	0.34	1.01	-0.01	-5.15E+05	-2.38E+04
Boosted Regression Tree	0.30	0.30	1.25	-0.10	3.99	0.81
Mediomastus ambiseta						
Poisson	0.13	0.24	1.20	-0.31	8.06**	1.46E+167
Negative Binomial	0.30	0.28	0.98	0.26	10.80	2.27
Tweedie	0.36	0.27	1.30	-0.19	10.98	2.27
Zero-Inflated Poisson	0.33	0.25	0.65	0.45	10.64**	-7.60E+301
Hurdle	0.03	0.11	0.12	1.25	10.71	3.03
Boosted Regression Tree	0.37	0.22	1.28	-0.32	11.49	2.27
Mediomastus californiensis						
Poisson	-0.03	0.23	1.32	-0.24	4.96E+27	1.79E+24
Negative Binomial	0.10	0.25	0.54	0.35	3.65	1.41
Tweedie	0.04	0.24	0.94	0.07	4.80	1.34
Zero-Inflated Poisson	0.00	0.19	0.00	0.76	20.21**	4.37E+241
Hurdle	0.07	0.15	0.72	0.15	-15.98	2.11
Boosted Regression Tree	0.12	0.20	1.64	-0.41	5.43	1.30

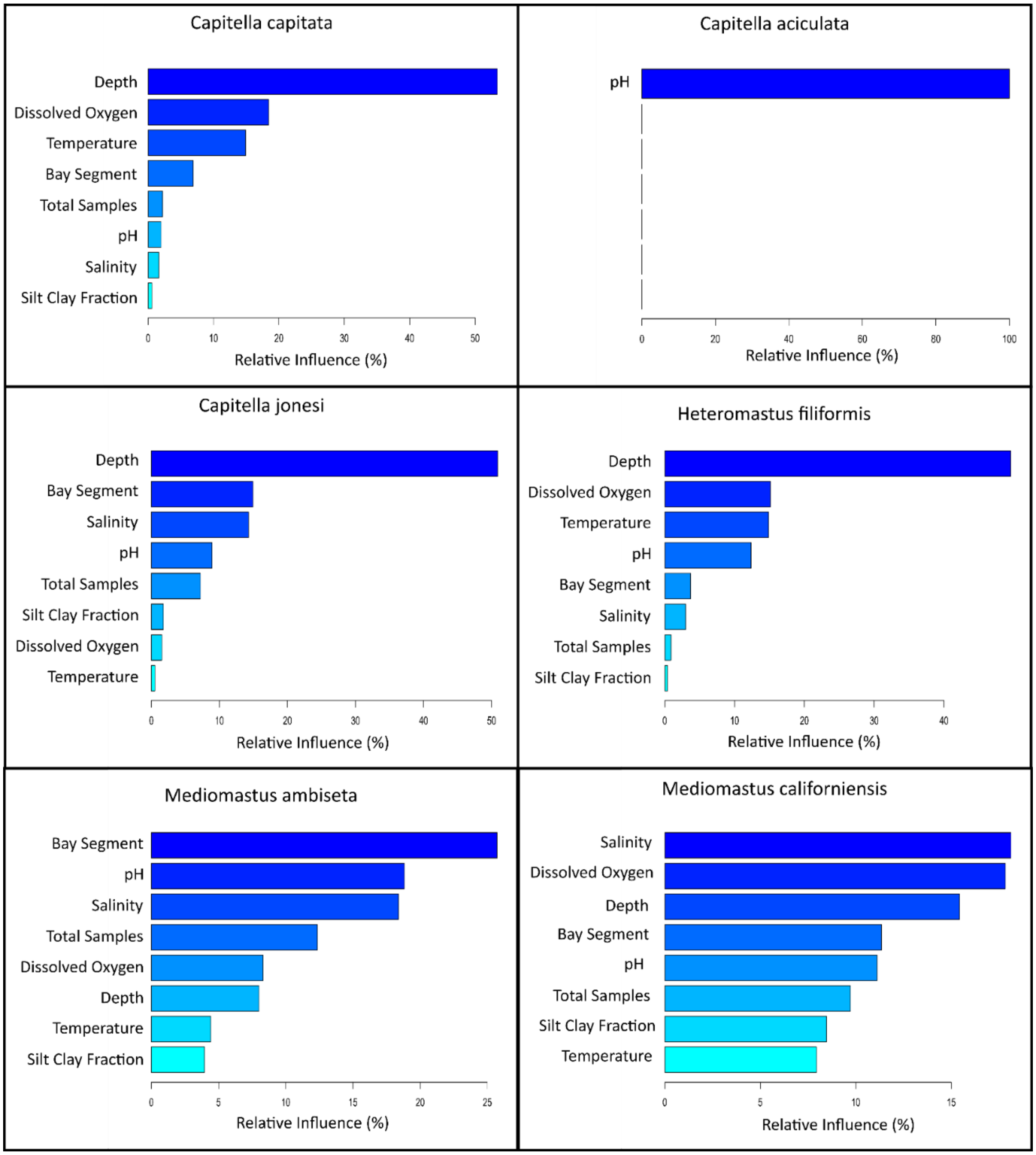


Figure A.1: Plots of the relative influence of each term as determined with the Boosted Regression Trees. pH is the only term labeled for *Capitella aciculata* because every other term had zero influence.