EVOLUTIONARY DIVERGENCE IN NATURAL SYSTEMS: EFFECTS, OF SELECTION, GEOGRAPHY, AND GENOME EVOLUTION

by

KAREN EILEEN BOBIER

(Under the Direction of John P Wares and Byron J Freeman)

ABSTRACT

Numerous mechanisms have played a part in generating the patterns of biodiversity we observe today. Understanding these drivers by examining genetic diversity within and between species is a goal of many evolutionary biologists. In this dissertation I contribute an addition to our understanding of evolutionary divergence and generation of biodiversity, by looking at two biogeographic examples including a geographic isolation of freshwater biota and a deeper examination of a known genetic cline along a latitudinal environmental gradient for a coastally distributed species. Further I examined how a gene family has evolved across vertebrates with interesting patterns of duplication among Actinopterygian fish. In addition to addressing questions of evolution I have also developed several genetic resources that can be used for further studies including a complete mitochondrial genome sequence for the yellowfin shiner, *Notropis lutipinnis*, a de novo transcriptome assembly for the barnacle *Notochthamalus scabrous*, as well as the first genome wide analyses of population divergence for four non-model freshwater fishes. The studies presented here expand our understanding of divers of divergence that generate diversity such as geographic isolation, environmental selection pressures, and genome evolution.

INDEX WORDS: Genetics, Biogeography, Evolution, Divergence, Vertebrates, Barnacles

EVOLUTIONARY DIVERGENCE IN NATURAL SYSTEMS: EFFECTS, OF SELECTION, GEOGRAPHY, AND GENOME EVOLUTION

by

KAREN EILEEN BOBIER

Bachelor of Science, University of California Santa Cruz, 2013

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2020

EVOLUTIONARY DIVERGENCE IN NATURAL SYSTEMS: EFFECTS, OF SELECTION, GEOGRAPHY, AND GENOME EVOLUTION

by

KAREN EILEEN BOBIER

Major Professor:	John P. Wares Byron J. Freeman
Committee:	Allen J. Moore Travis C. Glenn Mary G. Goll

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
December 2020

DEDICATION

To family, friends, and pets who have always been there for me.

ACKNOWLEDGEMENTS

There are many people who have helped me on this journey, and I would like to thank them for all of the support I have received from them. First, I need to thank Jeff who has been an amazing partner supporting my academic ambitions for years from joining me in moving across the country to Georgia to dealing with me being mildly crazy while wrapping up my dissertation. My family has always been supportive of my studies even though they wanted me to be an engineer. I especially want to thank my sister Carrie who, in addition to being amazingly supportive and helped with most of my college applications, also set an example for me by going to college first and then getting a PhD. Without her forging a path and showing me a trail to follow, I doubt I would have made it to this point. My best friend Holly has been an amazing pseudo-lab mate and neighbor. She has been a great help with troubleshooting project ideas or code bugs and is also amazingly fun to hang out with whether we're crafting, drinking, hiking, or helping me catching minnows in the middle of nowhere Georgia. I would also like to thank the fellow graduate students in my lab, Kelly, Paige, and Margot, who have made our lab such a fun environment and I have really missed seeing you all regularly this year (pandemics man). Thank you to Christine, former Wares Lab grad, who laid the foundation for and helped with my chapter on barnacle biogeography. Thank you to everyone in the Glenn Lab who has helped me make 3RAD libraries and showed me how to analyze my data. All of the members of the Goll Lab have also been awesome and showed me the ways of a zebrafish genetics lab. Thanks to Adam Bewick for guiding me on how to approach my Dnmt evolution project. Thanks to Marcus Zokan, Brett

Albanese and everyone else at Georgia DNR who helped me catch fish for my stream capture project.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Main Text

Divergence is an inherent component of evolution and can occur at many levels including individual nucleotides in DNA sequences, allele frequencies in populations, locations of genes on chromosomes, regulation of gene expression, and speciation. These levels of divergence occur due to many processes including mutations of nucleotides, genome rearrangements like inversions and translocations, selection like local adaptation, or sexual selection, geographic isolation, population structure, genetic drift and many more. Aquatic habitats are often home to highly diverse biota, are generally understudied, and provide critically important ecosystem services. For these reasons I have focused my dissertation on studying mechanisms of divergence and generators of diversity in aquatic systems. In this dissertation I present studies examining a variety of processes, in a variety of systems.

The second chapter of my dissertation is a brief phylogenetic examination of mitochondrial genomes of a group of cypriniform fishes. Only recently have we been able to easily obtain whole mitochondrial sequences for numerous closely related non-model taxa. This increase in our potential for collecting genetic data is critically important for understanding the diversity of cypriniform fishes. Many studies are using a single mitochondrial gene as a barcode for phylogenetics and species delineations, however recent studies have shown that analyses of many gene regions of the

mitochondrion, applying appropriate mutation models, can enhance our ability to distinguish species with molecular tools (Dupuis et al., 2012; Tang et al., 2014; Wares, 2014). In this chapter I assembled the mitochondrial genome of the yellowfin shiner, *Notropis lutipinnis*, and examined phylogenetic relationships between this species and 29 other closely related cypriniform fishes. While this chapter illustrates the power of increasingly complex data for examining species relationships, future studies could improve our understanding of species relationships by including several nuclear loci.

The third chapter of my dissertation focuses on examining genome wide divergence following a vicariant geologic event, to assess the effects of the geologic process of stream capture as a mechanism of dispersal and vicariance in freshwater systems. I used the described capture of the headwaters of the Chattahoochee River by the Savannah River as a case study of freshwater vicariance. A comparison of genomic data of populations separated by this stream capture will provide us with a better understanding of how these geologic processes have shaped present day biodiversity. These findings will help inform studies of freshwater biogeography and increase our understanding of mechanisms generating diversity in freshwater systems.

The fourth chapter of my dissertation is a study of a latitudinal genomic cline in *Notochthamalus scabrosus*, a small barnacle distributed along the coast of Chile. Similarly to my second chapter, this chapter focuses on a biogeographic transition; however the biogeographic forces affecting an intertidal organism distributed along a continuous coast line are not as obvious as a land barrier between rivers (Haye et al.,

2014; Wares, Gaines, & Cunningham, 2001). There are however differences in environmental conditions, such as temperature, salinity, and level of upwelling, at the Northern and Southern extent of the species range. Additionally, there is a known biogeographic break in the middle of the species range caused by an onshore current that splits into a north flowing and south flowing current along the coast (Ewers-Saucedo et al., 2016). I am assessing loci across the genome to identify genes under selection that may be diverging due to difference in environment at the northern and southern extents of the species range.

The fifth chapter of my dissertation focuses on genome evolution by examining the evolution of a gene family across vertebrates. In addition to biogeographic effects, divergence can also be driven by gene duplication and loss, chromosomal rearrangements, and genome duplications. All of these molecular mechanisms have been involved in the evolution of the DNA Methyltransferase (Dnmt) gene family among vertebrates. Dnmts are interesting because of their role in DNA methylation, and potential changes in gene expression, and chromatin structure, that may be associated with methylation patterns. Dnmts are known to have a highly conserved region and the gene family has not changed much across mammals. Prior to this study the gene family had been examined in zebrafish but not many other non-tetrapod vertebrates. However, there were known duplications of Dnmt3s in some fishes, assumed to be the result of the teleost whole genome duplication. By Utilizing 232 annotated vertebrate genomes, I found that Dnmt1 is maintained in single copy, and is conserved at sequence level. When Dnmt3 is duplicated multiple copies can be retained in the genome. There are interesting

patterns of Dnmt duplication and retention in fishes, including a gain of a Calponin Homology domain, before the teleost whole genome duplication, and corresponding to an initial tandem duplication of Dnmt3.

References

Dupuis, J. R., Roe, A. D., & Sperling, F. A. (2012). Multi-locus species delimitation in closely related animals and fungi: One marker is not enough. *Molecular Ecology*, *21*(18), 4422–4436.

Ewers-Saucedo, C., Pringle, J. M., Sepúlveda, H. H., Byers, J. E., Navarrete, S. A., & Wares, J. P. (2016). The oceanic concordance of phylogeography and biogeography: A case study in Notochthamalus. *Ecology and Evolution*, *6*(13), 4403–4420.

Haye, P. A., Segovia, N. I., Muñoz-Herrera, N. C., Gálvez, F. E., Martínez, A., Meynard, A., . . . Faugeron, S. (2014). Phylogeographic Structure in Benthic Marine Invertebrates of the Southeast Pacific Coast of Chile with Differing Dispersal Potential. *PLoS ONE, 9*(2), e88613. doi:10.1371/journal.pone.0088613

Tang, M., Tan, M., Meng, G., Yang, S., Su, X. U., Liu, S., Song, W., Li, Y., Wu, Q., & Zhang, A. (2014). Multiplex sequencing of pooled mitochondrial genomes—A crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Research*, *42*(22), e166–e166.

Wares, J. P. (2014). Mitochondrial cytochrome b sequence data are not an improvement for species identification in Scleractinian corals. *PeerJ*, *2*, e564.

Wares, J. P., Gaines, S., & Cunningham, C. W. (2001). A comparative study of asymmetric migration events across a marine biogeographic boundary. *Evolution, 55*(2), 295-306. doi:10.1111/j.0014-3820.2001.tb01294.x

CHAPTER 2

THE COMPLETE MITOCHONDRIAL GENOME OF THE YELLOWFIN SHINER, NOTROPIS LUTIPINNIS[1]

[1] Bobier, K. E. (2020). The complete mitochondrial genome of the yellowfin shiner, *Notropis lutipinnis*. *Mitochondrial DNA Part B*, *5*(3), 3203-3205. This is an Accepted Manuscript of an article published by Taylor & Francis in *Journal of Mitochondrial DNA Part B* on *August 1, 2020*, available online: https://doi.org/10.1080/23802359.2020.1809541. Reprinted here with permission of publisher.

Abstract

The complete mitochondrial genome of the yellowfin shiner (*Notropis lutipinnis*) 16706 bp and contained 13 protein coding genes, 2 rRNAs, 22tRNAs, and one control region. The overall base composition was A (28.8%), T (27.0%), C (26.7%), G (17.5%). Phylogenetics analyses of *N. lutipinnis* and 29 closely related species found similar discrepancies between genetic relationships and taxonomic delineations, highlighting the need for further studies of phylogenetic and biogeographic relationships among the closely related taxa of the subfamily Pogonichthyinae.

Main-Text

Notropis lutipinnis (Jordan & Brayton, 1878), commonly known as the yellowfin shiner, is a cypriniform fish in the family *Leuciscidae* found in freshwater streams across the Southeastern United States from Alabama to North Carolina, in the Mobile, Tennessee, Apalachicola, Altamaha, Savannah, Edisto, and Santee River Drainages[1, 2, 3]. *N. lutipinnis* has a natural distribution in both Gulf and Atlantic coast drainage basins and is found in headwater streams, creeks, and small rivers. When present, *N. lutipinnis* is often among the most abundant fish species at a site, where it has unique ecological interactions with other species such as complex spawning aggregations and nest parasitism of the bluehead chub (*Nocomis leptocephalus*).

Genomic DNA was isolated from a caudal fin clip of a specimen collected from Turnpike Creek (35.15141, -84.26057), a tributary to the Flint River; the voucher is in the Georgia

Museum of Natural History Tissue Collection (GMNHTC# 11921). DNA was isolated using the Puregene DNA extraction kit and sequenced by Illumina MiSeq (250bp paired-end reads) at the Georgia Genomics and Bioinformatics Core (formerly the Georgia Genomics Facility). Raw reads were trimmed and quality checked with Trim Galore[4]. The Geneious Read Mapper algorithm[5] in Geneious 8.1.9 was used to map reads to the mitochondrial genome of *Notropis chrosomus* (AP012108.1 ref). We found 99642 reads mapped to the *N. chrosomus* reference. The mapped reads were then used to create a *de novo* assembly with mean coverage of 1308.4 ± 204.9. No large structural variants were observed between *N. lutipinnis* and *N. chrosomus* mitochondrial genomes. Annotations were completed with Mitofish Annotator[6]. The complete, annotated mitochondrial sequence is accessible through GenBank under accession number MT333789.

The complete mitochondrial genome of *Notropis lutipinnis* (16706 bp) consists of 13 protein coding genes, two rRNA genes, 22 tRNA genes, and the control region (D-loop), as expected for a vertebrate mitochondrial genome. The tRNA genes varied in length from 68bp (*tRNA-Cys*) to 76bp (*tRNA-Leu*). The overall base composition was A (28.8%) > T (27.0%) > C (26.7%) > G (17.5%). The percentage of GC (44.2%) was lower than AT. The start codon for all protein coding genes was ATG, with the exception of Cytochrome c Oxidase subunit I (COXI), which was GTG[7]. Six protein coding genes use the stop codon TAA, two use the incomplete stop codon TA- and the remaining five use the incomplete stop codon T--. Presumably, these are cleaved at the base immediately following the partial stop codon during RNA processing, to keep the start codon of the subsequent gene intact, then converted to TAA stop codons upon poly-adenylation[8, 9].

The heavy strand acts as the coding strand for the majority of protein coding genes and tRNAs, however one protein coding gene (ND6) and 8 of 22 tRNAs (*tRNA-Pro*, *tRNA-Glu*, *tRNA-Ser*, *tRNA-Tyr*, *tRNA-Cys*, *tRNA-Asn*, *tRNA-Ala*, *tRNA-Gln*) use the light strand as the coding strand.

Phylogenetic analyses were completed using the complete mitochondrial genome sequence of *N. lutipinnis*, 28 other species of the subfamily *Pogonichthyinae*, and *Phenacobius mirabilis* (another *Pogonichthyinae*) as an outgroup[10]. Sequences were aligned with MUSCLE[11] with the maximum number of iterations set to 8. Iteration 1 of the alignment used the kmer4_6 distance measure, while iteration 2 used pctid_kimura, all iterations used the UPGMB distance measure, pseudo tree rooting, the CLUSTALW sequence weighting scheme, half penalty for terminal gaps, spm objective score, anchor spacing 32, open gap score of -1, minimum length of 24, margin of 5, minimum column anchor score of 90, hydrophobicity multiplier of 1.2 and, window size of 5. The phylogenetic tree was built with RaxML[12] implemented in Geneious using the GTR+G+I model of nucleotide substitution, new rapid hill climbing as the tree search algorithm, and 1000 inferences of the original tree on distinct randomized maximum parsimony trees (Figure 2.1).

We found that mitochondrial markers indicate that the genus *Notropis* is polyphyletic, whose species are interspersed with other genera as part of a larger monophyletic clade within the subfamily Pogonichthyinae of the family Leuciscidae. Species of *Notropis* phylogenetically cluster with *Algansea*, *Cyprinella*, *Hybopsis*, *Hybognathus*, *Luxilus*,

Opsopoeodus, *Lythrurus*, *Pimephales*, *Pteronotropis*, and *Tampichthes*. Similar results have been found in broader phylogenetic studies[10]. The combination of the phylogenetic relationships found here and in previous studies suggest detailed genetic studies of this group should be conducted to clarify taxonomic and phylogenetic relationships.

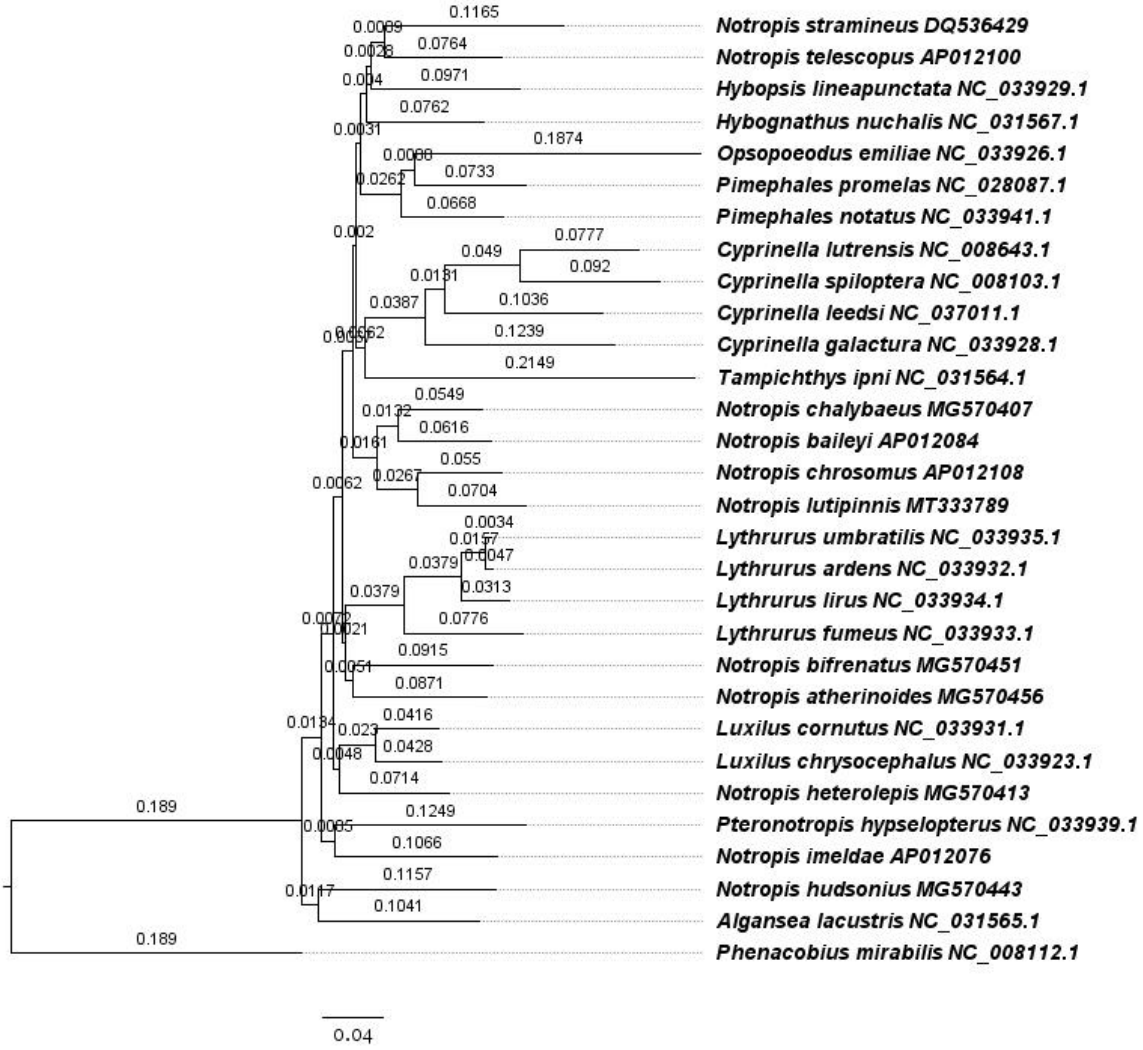

Figure 2.1 - Phylogenetic tree of shiner mitochondrial genomes built with RAxML; branch labels are substitutions per site.

The availability of this complete mitochondrial genome will support future ecological and biogeographic studies within this widespread species – including questions about

how fishes end up on opposite sides of the continental divide[13] and across the diverse lineage of *Notropis*.

References

1. Wood, Robert M., and Richard L. Mayden. "Systematics, evolution, and biogeography of *Notropis chlorocephalus* and *N. lutipinnis*." Copeia (1992): 68-81.
2. Scott, C. H., et al. "An awkward introduction: phylogeography of *Notropis lutipinnis* in its 'native' range and the Little Tennessee River." Ecology of Freshwater Fish 18.4 (2009): 538-549.
3. Cashner, Mollie F., Kyle R. Piller, and Henry L. Bart. "Phylogenetic relationships of the North American cyprinid subgenus *Hydrophlox*." Molecular phylogenetics and evolution 59.3 (2011): 725-735.
4. Krueger, Felix. "Trim Galore: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries." URL http://www. bioinformatics. babraham. ac. uk/projects/trim_galore/.(Date of access: 28/04/2016) (2012).
5. Kearse, Matthew, Shane Sturrock, and Peter Meintjes. "The Geneious 6.0. 3 read mapper." Biomatters Ltd.: Auckland, New Zealand (2012).
6. Iwasaki, Wataru, et al. "MitoFish and MitoAnnotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline." Molecular biology and evolution 30.11 (2013): 2531-2540.

7. Delarbre, Christiane, et al. "The main features of the craniate mitochondrial DNA between the ND1 and the COI genes were established in the common ancestor with the lancelet." Molecular biology and evolution 14.8 (1997): 807-813.
8. Clayton, David A. "Transcription and replication of mitochondrial DNA." Human reproduction 15.suppl_2 (2000): 11-17.
9. Ojala, Deanna, Julio Montoya, and Giuseppe Attardi. "tRNA punctuation model of RNA processing in human mitochondria." Nature 290.5806 (1981): 470-474.
10. Schönhuth, Susana, et al. "Phylogenetic relationships and classification of the Holarctic family Leuciscidae (Cypriniformes: Cyprinoidei)." Molecular phylogenetics and evolution 127 (2018): 781-799.
11. Edgar, Robert C. "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic acids research 32.5 (2004): 1792-1797.
12. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014 May 1;30(9):1312-3.
13. Johnson, Douglas Wilson. "River capture in the Tallulah district, Georgia." Science 25.637 (1907): 428-432.

Funding Details

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institute of Health under award number T32GM007103. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Disclosure Statement

The authors report no conflicts of interest.

Data Availability Statement

Data that support the findings of this study are openly available in Genbank with reference accession number MT333789 at https://www.ncbi.nlm.nih.gov/nuccore/MT333789.

Geolocation Information

The specimen was collected under the Georgia Department of Natural Resources Permit Number 29-WBH-12-129 issued to Byron J. Freeman. on August 6, 2012 by Mary C Freeman, and Rachel A Katz at Turnpike Creek, Flint River Drainage, Georgia, 33.152829 -84.261203, Site: Turnpike Creek DS of Perkins Rd crossing. The specimen has been added to the Georgia Museum of Natural History's Tissue Collection as GMNH 11921.

Acknowledgements

I would like to thank my co-advisers John Wares and Bud Freeman for their mentorship and guidance throughout my doctoral studies; my friends and family for their support; and the UGA Genetics Department and Integrated Life Sciences Program.

CHAPTER 3

GENETIC DIVERGENCE OF FRESHWATER FISHES ASSOCIATED WITH A GEOLOGIC STREAM CAPTURE[2]

Introduction

Around 0.01% of the world's water is freshwater, but this limited habitat is home to one third of all vertebrate species and 6% of all described species (Dudgeon et al., 2006). Freshwater biodiversity provides crucial ecosystem services to humans. These services include: food, clean drinking water, and flood protection (Aylward et al., 2005; Gamfeldt, Hillebrand, & Jonsson, 2008). The biodiversity of freshwater ecosystems is under threat worldwide, and many communities risk losing the services of local ecosystems. Forty-two percent of streams nationwide are considered to be in poor biological condition. This condition is due largely to anthropogenic activities affecting sedimentation, water temperature, and nutrient abundances in waterways (Regions, 2006). The factors affecting steam condition are numerous, and vary within and between watersheds (Regions, 2006). With these freshwater ecosystems under threat and the apparent importance of biodiversity to humans it becomes important to study this biodiversity and attempt to understand what mechanisms play a role in generating biodiversity.

[2] Bobier, K. E., Freeman, B., and Wares, J.

The Southeast is home to the highest aquatic biodiversity in the continental United States (Burr & Mayden, 1992), but drivers of this diversity are not completely understood. This diversity is not restricted to a specific taxon, but is reflected across fishes, amphibians, snails, mollusks, and turtles (Burr & Mayden, 1992), itself a conundrum for how these distributions formed in fully aquatic species. This diversity includes many endemic species but also species that occur in many drainages throughout the southeast (Burr & Mayden, 1992). Some of these organisms have apparent mechanisms by which dispersal between drainages could occur. For example, amphibians and turtles may simply be able to walk between drainages. However, for other organisms, such as fishes and mussels, dispersal mechanisms are not as obvious. This region of elevated aquatic diversity is centered around the southern Appalachian Mountains (Boschung & Mayden, 2004; Crandall & Buhay, 2007; Duellman & Sweet, 1999; Elkins et al., 2019; Etnier & Starnes, 1993; Kozak, Mendyk, & Wiens, 2009; Kozak & Wiens, 2010; Lundberg, Kottelat, Smith, Stiassny, & Gill, 2000; Parmalee & Bogan, 1998; Petrnka, 1998). With aquatic species widespread across numerous drainages, elucidation of the biogeography of the region is critical for understanding dispersal, distribution, and past vicariance events. In this study we evaluate the effects of the geologic process of "stream capture" in generating the diversity we observe today and how it has shaped the biogeography of aquatic species in the Southeastern United States. Stream capture is the process by which a stream's path is diverted – typically through headwater erosion - to flow into another river, and has occurred repeatedly throughout the Southeastern US on geologic scales that are appropriate to this question. The river networks of the Southeast have been

geologically dynamic, due to processes of erosion and uplift shifting paths of streams and rivers.

Stream Capture

Stream capture is the process by which a portion of one river system is transferred to another; these events are often facilitated by erosion or uplift resulting in connections between adjacent drainages that divert flow of water from its original course to a new one (Figure 3.1). Stream capture events thrust populations upstream of the capture point into a different drainage, generating dispersal between drainages. When this happens, the populations in distinct drainages diverge (Albert & Crampton, 2010; Planes & Fauvelot, 2002).

Vicariance and Novel Habitat

After a stream capture event organisms that were transferred will have their habitat connected to a novel habitat. This new habitat could include new predators and pathogens, it is also likely that near the point of capture there would temporarily be increased sediment loads in the water. Furthermore, organisms left behind in the original drainage will also experience a drastic change in their habitat. If they were just downstream of the capture, then water flows entering that reach of stream likely decreased or completely disappeared, in this case these organisms would need to adapt to the change or move elsewhere in the stream system. This forced dispersal will likely result in exposure to habitat that differs from the organism's original habitat. By exposing

organisms to new environmental factors stream capture events could facilitate selection based divergence and lead to local adaptation.

Repeated dispersal, isolation, and subsequent divergence due to these stream capture events seem to have played a major role in creating the diversity of aquatic species found in the southeast. For example, genetic relationships of the salamanders *Desmognathus marmoratus* and *D. quadramaculatus*, were examined across four major river drainages and found polyphyletic relationships between the two species, with deeper genetic divergences between rivers than between the two described species in each drainage (M. T. Jones, Voss, Ptacek, Weisrock, & Tonkyn, 2006). This suggests that what has been considered two species may actually be two species in each major drainage, whose divergence patterns match hypothesized stream capture dispersal events. Furthermore, these stream capture events may have contributed to wide distributions of some fish species observed today.

The geologic history and physiographic diversity of the southeast set the stage for a diverse array of aquatic species. For example, the minnow genus *Notropis* has 31 described species in Georgia alone (Straight, Albanese, & Freeman, updated 2009 March 25). Given this diversity, misidentifications are common in the field. It is also likely that there are still more cryptic undescribed species in this region yet to be identified. For example, the Halloween darter, *Percina crypta*, has recently been described as a separate species from the black-banded darter, *Percina nigrofasciata*, which can only be visually differentiated by slight variation in coloration (Freeman, Freeman, Burkhead, & Straight,

2008). A further complication in accurate field identifications of this diverse group of species is the possibility of more recent anthropogenic dispersal, such as bait bucket transfer. Genetic analyses will enable distinguishing cryptic species. These analyses will also identify, and differentiate, historic dispersal between drainages by stream capture and more recent anthropogenic dispersal. A closer examination of multiple species' morphological and genetic divergence across drainages could elucidate the biogeography of all southeastern freshwater species.

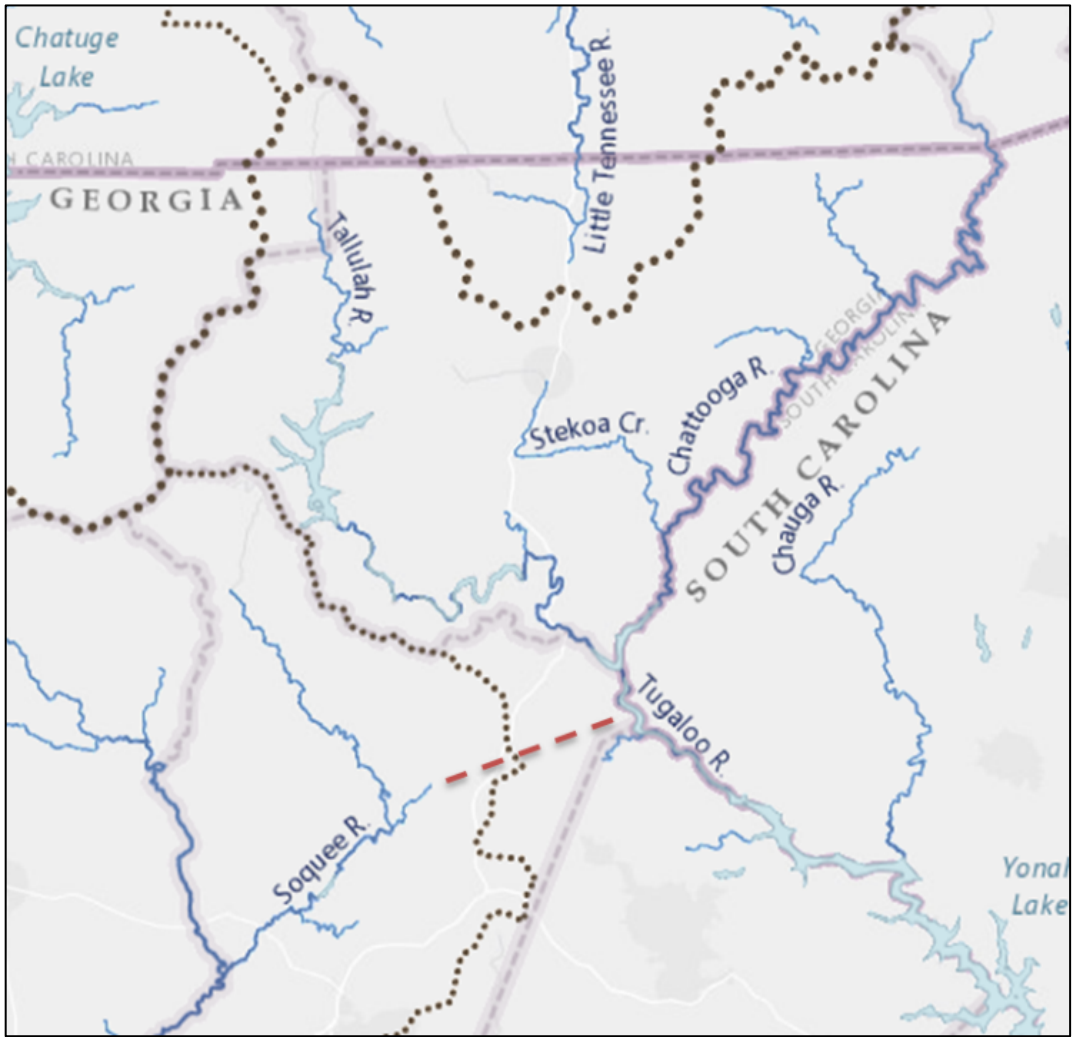

Figure 3.1 - Map showing the region of the stream capture from the Chattahoochee to Savannah River. The red dashed line indicates approximate former flow path based on (Hayes & Campbell, 1900; Johnson, 1907a, 1907b; Voss et al., 1995). Map generated with ESRI ArcGIS Online.

Tallulah Gorge Study System

The Tallulah River, an example of stream capture, now flows into the Tugaloo then the Savannah River towards the Atlantic Ocean (Figure 3.1). However, examination of the region's geology and physiographic features shows the Tallulah formerly flowed into the Chattahoochee River drainage towards the Gulf of Mexico (DuBose, 2017; Hayes & Campbell, 1900; Johnson, 1907b; M. T. Jones et al., 2006; Voss et al., 1995). This stream capture formed the Tallulah Gorge, which is estimated to have occurred during the Pleistocene.

The minnow genus *Notropis* has thirty-one described species in Georgia alone (Straight et al., updated 2009 March 25). A closer examination of multiple species' morphological and genetic divergence across drainages could elucidate the biogeography of all Southeastern freshwater species. The Tallulah River stream capture provides an opportunity to examine genetic divergence after a known geologic event. We can use the time of the Tallulah Stream Capture as a set point in the past and assess the divergence of populations in the Chattahoochee and Savannah Rivers since that point. Here we examine how these stream capture events shaped current species distributions by examining genetic divergence on four species that occur in both the Chattahoochee and Savannah River drainages.

Methods

In order to assess divergence of populations in the Chattahoochee and Savannah River Drainages, I have collected tissue samples from four individuals from each of four species (*Notropis lutipinnis*, *Nocomis leptocephalus*, *Luxilus zonistius*, and *Percina nigrofasciata*) from both river drainages (Table S3.1). Specimens are stored in 95% ethanol and will be added to the Georgia Museum of Natural History Genomics Tissue Collection.

DNA was isolated from fin-clips or lateral muscle tissue using the Qiagen Puregene DNA Isolation kit. Reduced-representation genotype-by-sequencing protocols were used to generate single nucleotide polymorphism (SNP) data sets representing all 32 fish. Here, we use the Adapterama III protocol (Bayona-Vásquez et al., 2019) with a combination of XbaI, EcoRI, NheI restriction enzymes, and molecular ID tags (a random eight nucleotide sequence in the outer barcode) developed by (Bayona-Vásquez et al., 2019) for bioinformatic separation of fragments prior to analysis. Size selection for libraries was conducted using Pippin Prep with a range of 550bp ± 12%. Genomic libraries were sequenced on an Illumina NovaSeq platform.

Stacks Analyses

Using STACKS (Catchen, Amores, Hohenlohe, Cresko, & Postlethwait, 2011) process_radtags, sequences were demultiplexed by the i7 outer barcodes, then demultiplexed by internal Adapterama barcodes, checked for restriction sites, and quality filtered. To remove PCR duplicates the clone_filter option was used. Parameters for

assembly of sequence fragments were tested to determine optimal values for maximizing appropriate data by running denovo_map.pl through ten iterations with M (the parameter that sets number of mismatches allowed between stacks within individuals) varying from 1 to 10, followed by 10 iterations of n (the parameter that sets the number of mismatches allowed between stacks between individuals) from 1 to 10 after M was set. Values of M and n were selected to maximize the number of loci recovered while minimizing splitting of actual loci (Paris, Stevens, & Catchen, 2017). For all species M was set to 3 and n was set to 4 for final analyses. Outputs from denovo_map.pl were mapped against reference genomes using stacks-integrate-alignments. *Perca flavescens* (GenBank GCA_004354835.1) was used as a reference for mapping *P. nigrofasciata* sequences. *Pimephales promelas* (GenBank GCA_000700825.1) was used as a reference for mapping *N. lutipinnis*, *N. leptocephalus*, and *L. zonistius*. Finally, the populations option was used to calculate F_{ST} and generate genpop and vcf files for subsequent analyses.

R Analyses

In the R package adegenet (Jombart & Collins, 2015), we initially evaluated variation among SNPs in each species using discriminant analysis of principle components (DAPC), to assess variance in the dataset and determine the optimal number of PCs using the alpha-score which assess PCs based on reassignment of data. This DAPC was used to indicate initial population assignment based on genotype.

The R package poppr (Kamvar, Tabima, & Grünwald, 2014) was used to calculate a Nei's distance matrix, and used to plotted with the program SplitsTree v5 (Huson & Bryant, 2006).

Site frequency spectra for each taxon were calculated using easySFS (https://github.com/isaacovercast/easySFS) a wrapper for DaDi (Gutenkunst, Hernandez, Williamson, & Bustamante, 2009) and fastsimcoal2 (Excoffier, Dupanloup, Huerta-Sánchez, Sousa, & Foll, 2013). To accommodate missing data that is inherent in RADseq datasets, we projected the data down to a fewer number of individuals to maximize the number of segregating sites included in the SFS (Gutenkunst et al., 2009). For each species we projected the data to three diploids per population from the four diploids sequenced.

Results

After demultiplexing and initial filtering 91,444,485 reads were retained across all 16 genomic libraries. Mean paired reads per sample ranged from 1827 to 87455 after decloning (Table 3.1). We recovered between 12428 and 19708 genotyped loci per species using the STACKS pipeline with coverage ranging from 16.3-57.3X (Table 3.2, S3.2). Summary statistics found differentiation between the Chattahoochee and Savannah Rivers for all four species (Table 3.3).

The ideal number of PCs to use based on alpha scores was 1, 1, 3, and 1 for *N. lutipinnis*, *N. leptocephalus*, *L. zonistius*, and *P. nigrofasciata* respectively. We retained 5 PCs to obtain the following results. Population reassignment based on PCs was successful, with the exception of one *N. lutipinnis* individual with lower coverage (Figure 3.2). Similarly, to population reassignment, network trees, based on Nei's distance show separation of individuals from the Savannah and Chattahoochee populations for each species. The one

exception to this was the single *N. lutipinnis* sample with low coverage, excluding this individual resulted in clear differentiation between the two populations.

To estimate divergence time relative to effective population size we used proportion of reciprocally monophyletic SNPs to total variant SNPs and compared them to a curve of expected reciprocal monophyly given time (McKenzie & Eaton, 2020; Rosenberg, 2003). Figure 3.3 illustrates these proportions for each species as synonymous with the probability of reciprocal monophyly. Though the sample size for each population is small, the overall difference in proportion of reciprocal monophyly and the indication of time diverged suggests distinct processes between the darter *P. nigrofasciata* and the 3 cypriniform fishes.

Table 3.1 – Reads per sample after decloning

	N. lutipinnis	***N. leptocephalus***	***L. zonistius***	***P. nigrofasciata***
Average	10641.6875	35201.5	53169.4375	46899.25
Min	1827	6958	26492	2757
max	18245	62420	68941	87455

Table 3.2 – Number of loci recovered and number of loci gained for each increase of n and M parameters in STACKS.

Value of n	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
# loci	9355	11038	12156	13035	13814	14267	14645	14925	15088	15244
# gained		1683	1118	879	779	453	378	280	163	156
Value of M	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
# loci	5307	5950	6413	6803	6946	7049	7099	7060	7021	6954
# gained		643	463	390	143	103	50	-39	-39	-67

Table 3.3 – Loci identified for each species with STACKS.

Species	Number of Genotyped Loci	Mean per Sample Coverage	Mean Sites per Locus
Notropis lutipinnis	12428	16.3	275.2
Nocomis leptocephalus	14358	41.7	273.7
Luxilus zonisitus	19708	57.3	270.2
Percina nigrofasciata	16673	54.8	269.5

Table 3.4 – Population differentiation between the Savannah (Sav) and Chattahoochee (Cht) Rivers. π (nucleotide diversity) and F_{IS} are reported for both variant sites and all sites. The proportion of variant loci that are fixed between populations.

Species	*N. lutipinnis*	*N. leptocephalus*	*L. zonisitus*	*P. nigrofasciata*
F_{ST}	0.1996	0.1015	0.1080	0.5062
F_{ST}'	0.4109	0.0454	0.0736	0.7539
φ_{ST}	0.2394	0.0472	0.0738	0.6199
Sav π (variant)	0.1947	0.2130	0.2751	0.0737
Cht π (variant)	0.2800	0.2991	0.1826	0.1932
Sav π (all sites)	0.0026	0.0011	0.0015	0.0009
Cht π (all sites)	0.0037	0.0016	0.0010	0.0024
Sav F_{IS} (variant)	0.0279	0.0107	0.0751	0.0044
Cht F_{IS} (variant)	0.06687	0.05681	0.01728	0.03931
Sav F_{IS} (all sites)	0.00037	0.00006	0.0004	0.00005
Cht F_{IS} (all sites)	0.00089	0.0003	0.00009	0.00049
% fixed diff	0.02396	0.00213	0.00299	0.34662
Nei's D	0.20632	0.07694	0.07876	0.75445

Discussion

With only four individuals per population, we were able to distinguish populations using population reassignment based DAPC. Certainly, the clearest divergence across this biogeographic event is represented by *P. nigrofasciata*. For a small sample size –

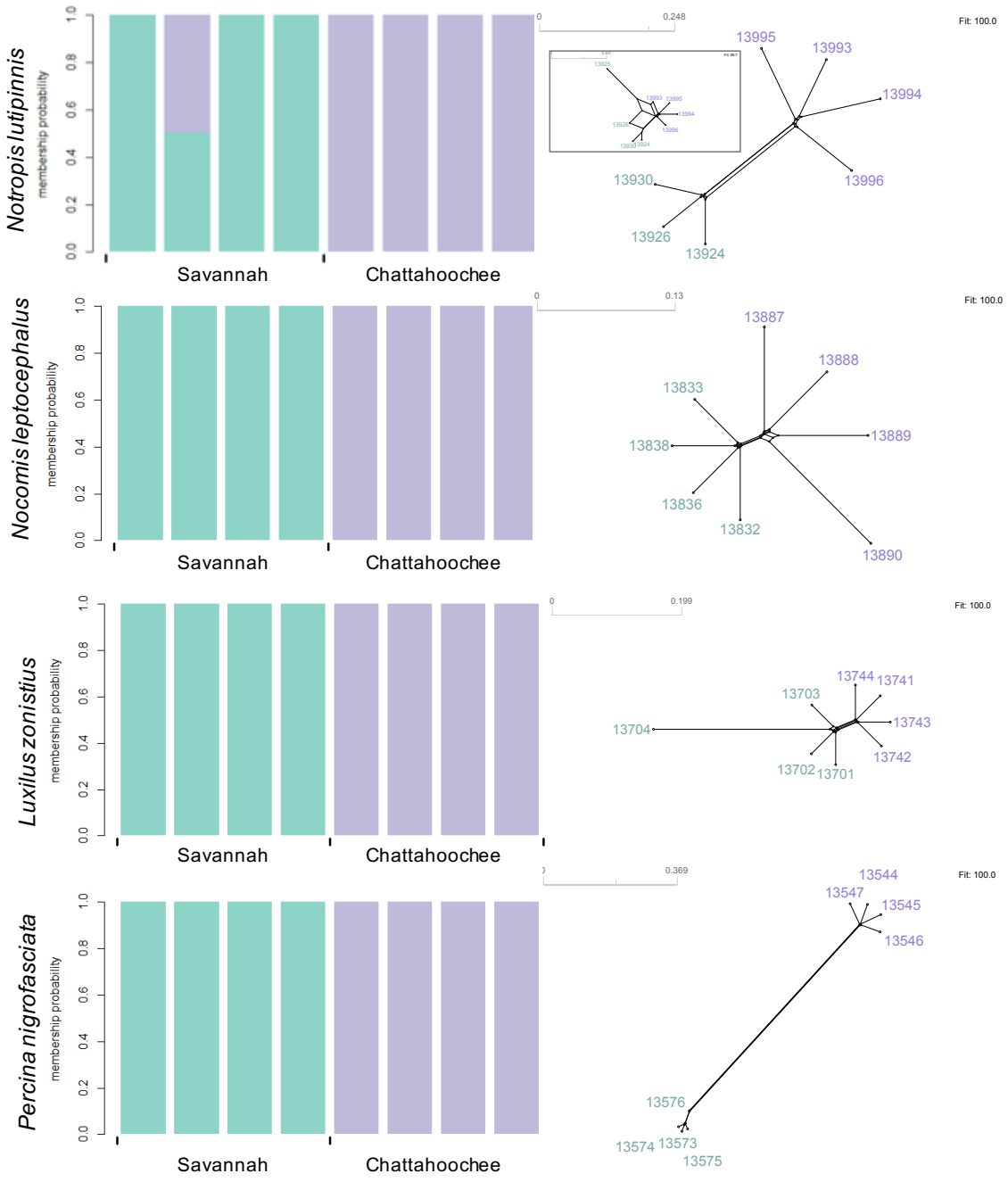

Figure 3.2 – Left - Population reassignment based on use of 5 PCs from DAPC analysis, plotted with the compoplot function in the adegenet package in R which calculates group assignments based on geometric criteria in the discriminate space. Right – Nei's 1972 genetic distance tree Plotted with SplitsTree V4.10, based on distance matrix for all individuals calculated in the R package StaMPP v1.6.1.

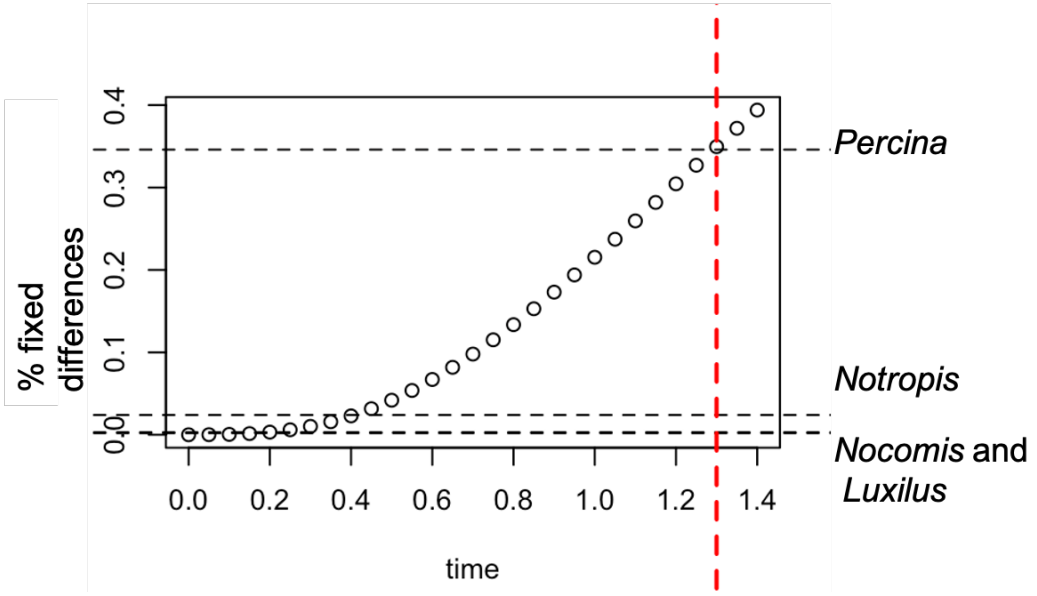

Figure 3.3 - Plot of expected reciprocal monophyly per coalescent time unit (black circles) from equation 14 of Rosenberg (2003). Dashed horizontal lines represent estimates of proportion of reciprocal monophyly for each of the 4 taxon pairs; vertical red dashed line is the inferred coalescent time of population divergence for *P. nigrofasciata*; the other 3 species exhibit lower apparent divergences.

generating some uncertainty in these estimators of divergence – F_{ST}' is greater than 0.75 and ~a third of the loci represent fixed differences. Assuming the transformation to reciprocal monophyly as a Bayesian posterior probability of single-locus monophyly estimates, this would indicate the divergence of *Percina* to be on the order of 1.2-1.4*Ne generations. Estimation of *Ne* itself can be problematic and requires assumed knowledge of the mutation rate μ (Hare et al., 2011; Turner, Wares, & Gold, 2002), but if Ne is large as one could expect for a small freshwater fish (e.g. coalescent Ne for the Rio Grande silvery minnow is over a million; (Alò & TURNER, 2005), this divergence may be associated with the Pleistocene separation caused by the Tallulah stream capture (DuBose, 2017; M. T. Jones et al., 2006).

Remembering that these data were intended to be a preliminary study into overall genomic diversity and divergence across this suite of fishes, the three cypriniform species appear to exhibit far shallower divergences and possibly polyphyly across drainages. Divergence estimates for *N. leptocephalus* and *L. zonistius* based on F-statistics and fixed differences are an order of magnitude lower than in *P. nigrofasciata*. This may suggest that although the stream capture is a true vicariance event for many aquatic taxa (Jones et al 2006), anthropogenic movement (perhaps as bait fish) has influenced the distribution of genomic diversity in *Nocomis* and *Luxilus*. The position of *N. lutipinnis* in this distribution is perhaps uncertain given the lower efficacy of data capture for samples. The *N. lutipinnis* individual GMNHTC 13925 was only genotyped at 398 loci, and data for this individual should be ignored rather than interpreted as admixture. Similarly, the *L. zonistius* individual 13704 that appears to be an outgroup to all other individuals in this dataset was genotyped at a large number of loci but had half the coverage of other individuals of this species. These small sample sizes and high variance in data quality across individuals may be complicating our interpretations.

Previous assessments of these species have also generated mixed results for understanding their divergence across major drainages. Scott et al (2009), examining mitochondrial and nuclear divergence in *N. lutipinnis*, found statistically robust divergence between the Savannah and Chattahoochee drainages, but a lack of reciprocal monophyly even for the mitochondrial data. As mitochondrial genomes should exhibit reciprocal monophyly much faster than nuclear loci (Hudson & Coyne, 2002), we now

think that it is unlikely that these results are generated by what is believed to be a Pleistocene vicariance event (DuBose, 2017; M. T. Jones et al., 2006).

The potential for expanding our understanding of these populations with large numbers of nuclear loci, however, is undeniable. The single nuclear locus evaluated previously in *N. lutipinnis* (the *transferrin* gene) (Scott, Cashner, Grossman, & Wares, 2009) itself was highly indicative of a divergence between these Gulf and Atlantic drainages; conflict between mitochondrial and nuclear genomic divergences are not uncommon (DeSalle & Giddings, 1986; Morales, Pavlova, Joseph, & Sunnucks, 2015; Toews & Brelsford, 2012; Wiens, Kuczynski, & Stephens, 2010). However, we clearly see there is more genetic divergence between the *P. nigrofasciata* populations compared to the three cypriniform species. This could suggest our observations of *P. nigrofasciata* are reflective of the vicariance event caused by the capture of the Tallulah system by the Savannah River and that while the other species may have also been moved though this geologic event they have also potentially had subsequent dispersal reducing the genetic divergence between populations. If there has been ongoing dispersal of the three cypriniform species between these two drainages, a potential mechanism of dispersal is though anthropogenic transfer for fishes being used for bait.

Though recent or ongoing dispersal appears to be maintaining low genetic differentiation between the Savannah and Chattahoochee population of the three cypriniform species, differences in life history characters compared to *P. nigofasciata* should also be accounted for. While the three cypriniform species each form large spawning

aggregations freely releasing large quantities of eggs into the water column, *P. nigrofasciata* will generally have mating pairs, that spawn a smaller number of eggs onto a substrate. These differences in spawning behavior could have a large impact on the effective population size for each species and therefore on the effects of drift changing our expectation for time to observed reciprocal monophyly. These differences in life history traits could also affect the probability of eggs being transferred between river drainages by animals. For example, a small percent of Asian carp eggs passed through the digestive system of mallards can emerge and be viable (Lovas-Kiss et al., 2020). Though eggs of these small bodied fish are likely more fragile than those of the Asian carp, it does suggest some dispersal by animals is possible.

These initial findings provide us with enough data to fuel future investigations such as finding out how much modern dispersal between drainages is ongoing, by estimating the number of migrants per generation from genetic data or investigating if other animals could act as dispersal vectors, and examining the extent of admixture across each species range. Since *P. nigrofasciata* is clearly diverged between the Savannah and Chattahoochee, further studies could examine the divergence of *Percina* species within the Chattahoochee river. A species that is cryptic with *P. nigrofasciata*, *P. crypta* has already been described, though some biologists suspect there is a third distinct lineage of *Percina* within this drainage (B. Freeman, *pers. Comm.*).

Examples of Stream Capture in the Southeast

Here we have used a method to examine genome wide divergence and ask about effects of geologic stream captures on patterns of divergence. Many other stream captures are suspected to have moved populations of freshwater organisms between river drainages and have potential for exciting genomic investigations.

Examples of stream captures in this region include the Chestatee River, which is suggested to have undergone a stream capture event from the Etowah River system to the Chattahoochee; unfortunately, there does not appear to be an estimation of when this event occurred (Hayes & Campbell, 1900). Interestingly, the Coosa Shiner, *Notropis xaenocephalus*, whose range is otherwise restricted to the Coosa and Tallapoosa systems, was historically reported from the Chestatee. The presence of Coosa Shiners in the Chestatee has several possible explanations, including a historic stream capture event, bait bucket transfer, or misidentification. To further complicate the situation, recent surveys have identified Tennessee Shiners, *Notropis leuciodus*, in the Chestatee, which was not previously recorded from the Chattahoochee drainage (Freeman personal communication). Another stream capture event from the Toccoa River, a tributary of the Tennessee drainage, to the Chestatee could have occurred and could explain the presence of the *N. leuciodus* in the Chestatee.

References

Albert, J., & Crampton, W. (2010). The geography and ecology of diversification in Neotropical freshwaters. *Nature Education Knowledge, 1*, 13-19.

Alò, D., & Turner, T. F. (2005). Effects of Habitat Fragmentation on Effective Population Size in the Endangered Rio Grande Silvery Minnow. *Conservation biology, 19*(4), 1138-1148. doi:10.1111/j.1523-1739.2005.00081.x

Aylward, B., Bandyopadhyay, J., Belausteguigotia, J.-C., Borkey, P., Cassar, A., Meadors, L., . . . Tognetti, S. (2005). Freshwater ecosystem services. *Ecosystems and human well-being: policy responses, 3*, 213-256.

Bayona-Vásquez, N. J., Glenn, T. C., Kieran, T. J., Pierson, T. W., Hoffberg, S. L., Scott, P. A., . . . Troendle, N. (2019). Adapterama III: Quadruple-indexed, double/triple-enzyme RADseq libraries (2RAD/3RAD). *bioRxiv*, 205799.

Boschung, H. T., & Mayden, R. L. (2004). *Fishes of Alabama*: Smithsonian Books.

Burr, B. M., & Mayden, R. L. (1992). Phylogenetics and North American freshwater fishes. *Systematics, historical ecology, and North American freshwater fishes. Stanford University Press, Stanford, California*, 18-75.

Byers, J., E., & Pringle, J., M. (2006). Going against the flow: retention, range limits and invasions in advective environments. *Marine Ecology Progress Series, 313*, 27-41. Retrieved from https://www.int-res.com/abstracts/meps/v313/p27-41/

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics, 1*(3), 171-182.

Crandall, K. A., & Buhay, J. E. (2007). Global diversity of crayfish (Astacidae, Cambaridae, and Parastacidae—Decapoda) in freshwater. In *Freshwater animal diversity assessment* (pp. 295-301): Springer.

DeSalle, R., & Giddings, L. V. (1986). Discordance of nuclear and mitochondrial DNA phylogenies in Hawaiian *Drosophila*. *Proceedings of the National Academy of Sciences, 83*(18), 6902-6906.

DuBose, D. (2017). Geochemical Signatures of Stream Capture in the Retreating Blue Ridge Escarpment, Southern Appalachian Mountains.

Dudgeon, D., Arthington, A. H., Gessner, M. O., Kawabata, Z.-I., Knowler, D. J., Lévêque, C., . . . Stiassny, M. L. (2006). Freshwater biodiversity: importance, threats, status and conservation challenges. *Biological Reviews, 81*(02), 163-182.

Duellman, W. E., & Sweet, S. S. (1999). Distribution patterns of amphibians in the Nearctic region of North America. *Patterns of distribution of amphibians. Johns Hopkins Univ. Press, Baltimore, MD*, 31-109.

Elkins, D., Sweat, S. C., Kuhajda, B. R., George, A. L., Hill, K. S., & Wenger, S. J. (2019). Illuminating hotspots of imperiled aquatic biodiversity in the southeastern US. *Global Ecology and Conservation, 19*, e00654.

Etnier, D. A., & Starnes, W. C. (1993). *The fishes of Tennessee*: University of Tennessee Press.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genet, 9*(10), e1003905.

Freeman, M. C., Freeman, B. J., Burkhead, N. M., & Straight, C. A. (2008). A new species of *Percina* (Perciformes: Percidae) from the Apalachicola River drainage, southeastern United States. *Zootaxa, 1963*, 25-42.

Gamfeldt, L., Hillebrand, H., & Jonsson, P. R. (2008). Multiple functions increase the importance of biodiversity for overall ecosystem functioning. *Ecology, 89*(5), 1223-1231.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet, 5*(10), e1000695.

Hayes, C., & Campbell, M. (1900). The relation of biology to physiography. *Science*, 131-133.

Hudson, R. R., & Coyne, J. A. (2002). Mathematical consequences of the genealogical species concept. *Evolution, 56*(8), 1557-1565.

Huson, D. H., & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution, 23*(2), 254-267.

Johnson, D. W. (1907a). *Drainage modifications in the Tallulah district.* Paper presented at the Proc. Boston Soc. Nat. Hist.

Johnson, D. W. (1907b). River Capture in the Tallulah District, Georgia. *Science, 25*(637), 428-432. Retrieved from http://www.jstor.org/stable/1633459

Jombart, T., & Collins, C. (2015). A tutorial for discriminant analysis of principal components (DAPC) using adegenet 2.0. 0. *London: Imperial College London, MRC Centre for Outbreak Analysis and Modelling*.

Jones, M. T., Voss, S. R., Ptacek, M. B., Weisrock, D. W., & Tonkyn, D. W. (2006). River drainages and phylogeography: an evolutionary significant lineage of

shovel-nosed salamander (*Desmognathus marmoratus*) in the southern Appalachians. *Molecular Phylogenetics and Evolution, 38*(1), 280-287.

Kamvar, Z. N., Tabima, J. F., & Grünwald, N. J. (2014). Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *Peerj, 2*, e281.

Kozak, K. H., Mendyk, R. W., & Wiens, J. J. (2009). Can parallel diversification occur in sympatry? Repeated patterns of body-size evolution in coexisting clades of North American salamanders. *Evolution: International Journal of Organic Evolution, 63*(7), 1769-1784.

Kozak, K. H., & Wiens, J. J. (2010). Niche conservatism drives elevational diversity patterns in Appalachian salamanders. *The American Naturalist, 176*(1), 40-54.

Lovas-Kiss, Á., Vincze, O., Löki, V., Pallér-Kapusi, F., Halasi-Kovács, B., Kovács, G., . . . Lukács, B. A. (2020). Experimental evidence of dispersal of invasive cyprinid eggs inside migratory waterfowl. *Proceedings of the National Academy of Sciences, 117*(27), 15397-15399.

Lundberg, J. G., Kottelat, M., Smith, G. R., Stiassny, M. L., & Gill, A. C. (2000). So many fishes, so little time: an overview of recent ichthyological discovery in continental waters. *Annals of the Missouri Botanical Garden*, 26-62.

McKenzie, P.F. & Eaton, D.A. (2020). The Multispecies Coalescent in Space and Time. *BioRxiv*.

Morales, H. E., Pavlova, A., Joseph, L., & Sunnucks, P. (2015). Positive and purifying selection in mitochondrial genomes of a bird with mitonuclear discordance. *Molecular ecology, 24*(11), 2820-2837.

Paris, J. R., Stevens, J. R., & Catchen, J. M. (2017). Lost in parameter space: a road map for stacks. *Methods in Ecology and Evolution, 8*(10), 1360-1373.

Parmalee, P. W., & Bogan, A. E. (1998). *Freshwater mussels of Tennessee*: University of Tennessee Press.

Petrnka, J. W. (1998). *Salamanders of the united States and Canada*: Washington: Smithsonian Institution Press.

Planes, S., & Fauvelot, C. (2002). Isolation by distance and vicariance drive genetic structure of a coral reef fish in the Pacific Ocean. *Evolution, 56*(2), 378-399.

Regions, W. M. (2006). Wadeable Streams Assessment.

Rosenberg, N.A. (2003). The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. *Evolution*, 57, 1465-1477.

Scott, C., Cashner, M., Grossman, G., & Wares, J. (2009). An awkward introduction: phylogeography of *Notropis lutipinnis* in its 'native'range and the Little Tennessee River. *Ecology of Freshwater Fish, 18*(4), 538-549.

Straight, C. A., Albanese, B., & Freeman, B. J. (updated 2009 March 25). Fishes of Georgia Website, Georgia Museum of Natural History. Retrieved from http://fishesofgeorgia.uga.edu

Toews, D. P., & Brelsford, A. (2012). The biogeography of mitochondrial and nuclear discordance in animals. *Molecular ecology, 21*(16), 3907-3930.

Turner, T. F., Wares, J. P., & Gold, J. R. (2002). Genetic Effective Size Is Three Orders of Magnitude Smaller Than Adult Census Size in an Abundant, Estuarine-Dependent Marine Fish (Sciaenops ocellatus). *Genetics, 162*(3), 1329. Retrieved from http://www.genetics.org/content/162/3/1329.abstract

Voss, S. R., Smith, D. G., Beachy, C. K., & Heckel, D. G. (1995). Allozyme variation in neighboring isolated populations of the plethodontid salamander *Leurognathus marmoratus*. *Journal of Herpetology, 29*(3), 493-497.

Wiens, J. J., Kuczynski, C. A., & Stephens, P. R. (2010). Discordant mitochondrial and nuclear gene phylogenies in emydid turtles: implications for speciation and conservation. *Biological Journal of the Linnean Society, 99*(2), 445-461.

Supplements

Table S3.1 – collection location information

Species	GNM HTC	Drainage	Stream	Field ID	Latitude	Longitude
N. lutipinnis	13993	Chattahoochee	Soque_River	KEB18011	34.711156	-83.5753419
N. lutipinnis	13994	Chattahoochee	Soque_River	KEB18011	34.711156	-83.5753419
N. lutipinnis	13995	Chattahoochee	Soque_River	KEB18011	34.711156	-83.5753419
N. lutipinnis	13996	Chattahoochee	Soque_River	KEB18011	34.711156	-83.5753419
N. lutipinnis	13924	Savannah	Chatooga	KEB18001	34.973854	-83.11572
N. lutipinnis	13925	Savannah	Chatooga	KEB18001	34.973854	-83.11572
N. lutipinnis	13926	Savannah	Chatooga	KEB18001	34.973854	-83.11572
N. lutipinnis	13930	Savannah	Chatooga	KEB18001	34.973854	-83.11572
N. leptocephalus	13887	Chattahoochee	Soque_River	KEB18011	34.711156	-83.5753419
N. leptocephalus	13888	Chattahoochee	Soque_River	KEB18011	34.711156	-83.5753419
N. leptocephalus	13889	Chattahoochee	Soque_River	KEB18011	34.711156	-83.5753419
N. leptocephalus	13890	Chattahoochee	Soque_River	KEB18011	34.711156	-83.5753419
N. leptocephalus	13832	Savannah	Panther_Creek	KEB18007	34.666944	-83.364167
N. leptocephalus	13833	Savannah	Panther_Creek	KEB18007	34.666944	-83.364167
N. leptocephalus	13836	Savannah	Panther_Creek	KEB18007	34.666944	-83.364167
N. leptocephalus	13838	Savannah	Panther_Creek	KEB18007	34.666944	-83.364167
L. zonistius	13741	Chattahoochee	Soque_River	KEB18011	34.711156	-83.5753419
L. zonistius	13742	Chattahoochee	Soque_River	KEB18011	34.711156	-83.5753419
L. zonistius	13743	Chattahoochee	Soque_River	KEB18011	34.711156	-83.5753419
L. zonistius	13744	Chattahoochee	Soque_River	KEB18011	34.711156	-83.5753419
L. zonistius	13701	Savannah	Stonewall_Creek	KEB18009	34.802	-83.431
L. zonistius	13702	Savannah	Stonewall_Creek	KEB18009	34.802	-83.431
L. zonistius	13703	Savannah	Stonewall_Creek	KEB18009	34.802	-83.431
L. zonistius	13704	Savannah	Stonewall_Creek	KEB18009	34.802	-83.431
P. nigrofasciata	13573	Savannah	Panther_Creek	KEB18007	34.666944	-83.364167
P. nigrofasciata	13574	Savannah	Panther_Creek	KEB18007	34.666944	-83.364167
P. nigrofasciata	13575	Savannah	Panther_Creek	KEB18007	34.666944	-83.364167
P. nigrofasciata	13576	Savannah	Panther_Creek	KEB18007	34.666944	-83.364167
P. nigrofasciata	13544	Chattahoochee	Soque_River	KEB18011	34.711156	-83.5753419
P. nigrofasciata	13545	Chattahoochee	Soque_River	KEB18011	34.711156	-83.5753419
P. nigrofasciata	13546	Chattahoochee	Soque_River	KEB18011	34.711156	-83.5753419
P. nigrofasciata	13547	Chattahoochee	Soque_River	KEB18011	34.711156	-83.5753419

Table S3.2 – Sequencing reads per sample, number of reads filtered out for lacking a radtag, number filtered for low quality, number of retained reads, mean coverage after final run of denovo map, and number of loci after final run of denovo map.

Species	GMNHTC	Total Reads	NoRadTag	LowQuality	Retained	mean_cov	num_loci
L. zonistius	13701	5499562	16930	30427	5452205	58.97	10565
L. zonistius	13702	4427478	12963	24788	4389727	52.26	9719
L. zonistius	13703	4608988	13443	25650	4569895	53.34	9704
L. zonistius	13704	2524770	8098	13996	2502676	29.06	10310
L. zonistius	13741	5368850	14857	30243	5323750	54.09	11136
L. zonistius	13742	5315918	14730	29791	5271397	60.02	10124
L. zonistius	13743	4627992	11288	25808	4590896	55	9627
L. zonistius	13744	2105204	6487	11667	2087050	29.84	8433
N. lutipinnis	13924	1094818	4026	6239	1084553	19.23	7443
N. lutipinnis	13925	49790	1579	295	47916	6.86	398
N. lutipinnis	13926	1223884	4535	6846	1212503	20.71	7602
N. lutipinnis	13930	607952	2188	3386	602378	15.55	5146
N. lutipinnis	13993	241346	1964	1315	238067	7.66	4172
N. lutipinnis	13994	1051800	3584	5794	1042422	18.01	7497
N. lutipinnis	13995	787728	2624	4317	780787	14.98	6824
N. lutipinnis	13996	486734	1595	2681	482458	11.01	6068
N. leptocephalus	13832	1370358	5059	7494	1357805	21.9	7182
N. leptocephalus	13833	453050	3498	2561	446991	10.45	5153
N. leptocephalus	13836	5185892	10253	29315	5146324	53.94	9908
N. leptocephalus	13838	891368	2534	4909	883925	18.69	5484
N. leptocephalus	13887	5155084	9264	28816	5117004	52.83	10074
N. leptocephalus	13888	5075298	13610	28178	5033510	51.69	10145
N. leptocephalus	13889	1896338	4492	10376	1881470	28.99	7554
N. leptocephalus	13890	3109872	6423	17118	3086331	35.98	9260
P. nigrofasciata	13544	4433206	12826	24657	4395723	74.81	9201
P. nigrofasciata	13545	3585826	14105	19859	3551862	42.56	8700
P. nigrofasciata	13546	3211158	10273	18077	3182808	57.38	8326
P. nigrofasciata	13547	3311448	8422	18605	3284421	8.04	8825
P. nigrofasciata	13573	6898772	19039	38348	6841385	54.59	9810
P. nigrofasciata	13574	3046174	12145	17262	3016767	46.63	8189
P. nigrofasciata	13575	4411268	15481	24610	4371177	43.55	8463
P. nigrofasciata	13576	171126	1877	947	168302	44.74	2401

CHAPTER 4

SIGNATURES OF SELECTION ALONG AN ENVIRONMETNAL CLINE[3]

Introduction

Coastal benthic marine species have limits to their distributions. Pappalardo et al 2015 (Pappalardo et al., 2015) show that few coastal marine species have ranges greater than 3000km. These range limitations exist despite wide dispersal potential of pelagic larvae. Mechanisms that limit species ranges and limit adaptation beyond their distributional range have long intrigued biologists (Antonovics, 1976). Many broadly distributed marine species exhibit significant population structure suggesting limited gene flow across major regions, often leading to cryptic diversity (Álvarez-Noriega et al., 2020) . Recent work by Nunez (Nunez et al., 2020) suggests that barnacles in particular, with the massive population of larval nauplii produced each year, maintain high levels of polymorphic diversity specifically to handle large environmental gradients. The species examined here, *Notochthamalus scabrosus*, appears to exhibit significant divergence across its 3000km range yet we have not before directly examined selection and how it unites and separates these populations.

The distribution of species can be governed by various factors, including predation, competition, physical barriers to dispersal, and environmental clines (Sexton et al., 2009). The presence of biogeographic breaks, at which many species with different ecological and evolutionary histories end or begin their distributional range, argues for

[3] Bobier, K. E.*, Ewers-Saucedo, C.*, and Wares, J. *co-first authors.

the importance of abiotic drivers for species distributions. In the face of global climate change, especially the study of environmental limits to species distributions becomes a pressing issue, as environmental clines may be shifting or disappearing.

Environmentally driven species boundaries may be particularly common in the marine realm, as other possible drivers are negligible. In contrast to continental ecosystems, the world's oceans have few physical barriers to dispersal. While predation and competition shape the micro-distribution of species, e.g. the zonation of barnacles along the intertidal zone, they are unlikely to shift species boundaries. Many marine organisms have high dispersal potential and large effective population size, which should allow for efficient/effective selection. Thus environmental clines are likely to shape the ranges of many marine species (Deutsch et al., 2020). That said, two factors may complicate the issue. Firstly, the large dispersal potential of marine species with a pelagic larval stage may swamp out the signal of local adaptation. Secondly, currents are important drivers of dispersal, potentially impacting species distributions more than environmental factors (Byers & Pringle, 2006; Pringle & Wares, 2007).

Nonetheless, the environment, especially temperature, appears to determine the distribution of many marine organisms (Pappalardo et al., 2015; Thiel, 2007). It is apparent that this includes many sister lineages that were previously assigned to one broadly distributed species. These closely related, sometimes not reproductively isolated lineages represent ideal study systems to dissect climate adaptations (Zakas & Wares, 2012), i.e. understand how lineages have adapted to different environments.

One such system is the South American barnacle *Notochthamalus scabrosus* (Darwin, 1854). These barnacles have a broad latitudinal range across more than 5000 km of Pacific coastline from Peru to Tierra del Fuego (Darwin, 1854) and the Falkland Islands (Häussermann & Försterra, 2009). As in all chthamalid barnacles, their habitat is restricted to the upper littoral habitats with long intervals of air exposure between high tides. These periods of exposure and their shallow intertidal distribution exposes them to large temperature amplitudes (Connell, 1961; Nunez et al., 2020).

Recent studies identified the presence of two genetically distinct lineages: a northern lineage, present from Peru to central Chile (15-42°S), and a southern lineage, present from central Chile to the Tierra del Fuego (30-55°S) (Ewers-Saucedo et al., 2016; Laughlin et al., 2012; Zakas et al., 2009, 2014). These distributions match the two biogeographic regions of the area: the Peruvian province to the north and the Magellanic province to the south. While the break is also marked by divergent currents and oceanic forcing to the north and south, respectively, Ewers-Saucedo et al. (2016) inferred that the currents alone cannot maintain the lineage distributions, and that local differential fitness of the two lineages has to be evoked to maintain the lineage boundaries. It remains to validate these modeling results and to identify the targets of selection.

In the present study, we attempted to identify signatures of selection in the exomes of *N. scabrosus* barnacles from the northern and southern lineage. We hypothesize that certain genes are under divergent selection between the northern and southern clade. These genes

may modulate temperature responses, thermal niches and desiccation stress. In particular, metabolic genes are frequent targets of natural selection, and may be responsible for environmental adaptation in this system as well (Marden, 2013; Skibinski & Ward, 2004). A particularly interesting candidate gene is mannose phosphate isomerase (MPI), which is under balancing selection in another intertidal barnacle, *Semibalanus balanoides* (Nunez et al., 2020). Different variants of this gene appear to be advantageous in distinct thermal environments, respectively (Schmidt et al., 2000; Schmidt & Rand, 1999, 2001).

Methods

De Novo Transcriptome Assembly

The specimen used for preparing the transcriptome was collected from Coquimbo, Chile in April 2018. The individual was immediately placed in RNAlater, frozen after overnight refrigeration, and shipped to Athens, Georgia. RNA was isolated from the specimen using the Qiagen RNeasy Mini Kit. Prior to sequencing, RNA quality was assessed using an Agilent BioAnalyzer 6000 Nano Chip at the Georgia Genomics and Bioinformatics Core. The RNA library was prepared and sequenced by NovoGene using 250-300 bp insert cDNA library and sequenced on an Illumina platform with 60million 150bp paired end reads.

The Oyster River Protocol (ORP version 2.0.0) (MacManes, 2018) pipeline, which combines assemblies from Trinity, SPAdes, and Shannon assemblers, was used to *de novo* assemble the transcriptome of *Notochthamalus scabrocus*. ORP was run using 480gb of memory on 28 computing threads on the Georgia Advanced Computing

Resource Center's Sapelo2 cluster. We then used Transdecoder version 2.1.0 (https://github.com/TransDecoder/TransDecoder) to identify the longest open reading frames (ORFs) greater than 100 amino acids long for each transcript, perform Pfam and BlastP searches to enable homology-based coding region identification, and finalize coding region predictions.

Population Genomics Analyses

Specimens for genomic DNA sequencing were collected from the northern extent of the species range in Arica, Chile (18.5°S), and at the southern extent of the species range in Ushuaia, Argentina (54.1°S). These are the same individuals as used in Ewers-Saucedo *et al* 2016 and Wares 2013. Short read 100bp paired end reads were obtained for seven individuals from Ushuaia, Argentina and five individuals from Arica, Chile.

Raw reads (fastq files) were aligned to the transcriptome using Burrows-Wheeler Aligner (BWA version 0.7.15 (sam files)). Aligned reads were converted to bam format and sorted using SortSam implemented in picard version 2.16.0. Duplicate reads were then removed using picard MarkDuplicates (http://broadinstitute.github.io/picard) and indexed with SAMtools index version 1.9. We then used GATK version 4.0.3.0 HaplotypeCaller to phase sequences and call variants for each individual (gVCF file). GATK GenotypeGVCFs was then used to create variant calling files for individuals and populations. All relevant scripts can be found at https://github.com/karenbobier/Notochthamalus/wiki.

Following assembly, sequence variants were evaluated across both populations for net nucleotide divergence (d_A) and within-population site frequency spectra was evaluated using Fu and Li's D using DnaSP version 6.11.01. We hypothesized that strongly divergent gene regions would be more likely to reflect neutral or directional sweep dynamics, and gene regions with nearly no divergence may reflect neutral or balancing selection dynamics.

Identification of Metabolic Genes

To identify specific metabolic genes in our dataset we created a blast database from our transcriptome and used tblastn with an e-value cut off of 1E-5 to search for 40 allozyme and metabolic loci from (Marden, 2013; Nunez et al., 2020; Skibinski & Ward, 2004).

Examination of Transcripts with High F_{ST}

To further investigate what molecular functions may be experiencing diversifying selection between the Northern and Southern populations we ran transcripts with an F_{ST}=1 through Interproscan v5.44-79.0 (Jones et al., 2014), to identify conserved protein domains. We then filtered to variant sites that were biallelic, genotyped in at least two individuals per population and had fixed genotypes between populations, and then calculated Ka/Ks between populations using ParaAT v 1.0 (Zhang et al., 2012) and KaKs_Calculator v1.2 (Zhang et al., 2006).

Results

Transcriptome Quality

Sequencing of RNA from the individual from Coquimbo, Chile resulted in 95554928 paired reads that were used for the de novo transcriptome assembly. ORP produced a transcriptome of 269455 contigs with a mean length of 456.76 bases and ranged in size from 172 to 13861 bases, 33914 of these contigs contained an open reading frame (ORF). The assembly has a Transrate score of 0.19522, and Transrate Optimal score of 0.38279. The Transcriptome assembly has a n90 score of 251, a n50 score of 487, and a n10 score of 1930. Of 303 eukaryote BUSCO (Based on evolutionarily-informed expectations of gene content of near-universal single-copy orthologs) groups searched we identified 238 complete (180 single copy, and 58 duplicated), 60 fragmented, and 5 missing BUSCO groups.

Transdecoder analyses resulted in 77997 mRNA sequences, of which 61002 had coding sequences with ORFs greater than 100 amino acids in length. These were used as the reference for mapping gDNA reads to for population analyses.

We recovered 3,269,066-68,185,994 reads per sample with an average of 19,026,744 reads per sample. After mapping reads for each sample to the transcriptome, we found 695,095 variable SNPs. Arica, Chile population samples had an average of 51.83% missing data at variable sites, while samples from Ushuaia, Argentina had 41.41% missing genotype data.

PCA (Figure 4.3), DAPC (Figure 4.4), and distance based phylogenetic trees (Figure 4.5) all found differentiation between samples from the Chile and Argentina populations. Interestingly, individuals that appear the most distant from the main cluster on the PCA are those with less missing data. Conversely, in the compoplot, which shows probability of assignment to a population based on genotype, individuals that have a proportion of assignment to both populations are the samples with the most missing data. These both demonstrate the importance of complete datasets for distinguishing individuals and populations with genetic data, however when working with non-model organisms a perfect dataset is not always possible.

Of the 40 metabolic loci we were searching for we were able to identify a Blast hit for each in our transcriptome, and 37 contained at least 1 conserved protein domain. Of those, fifteen loci had adequate genotype information across individuals sampled to calculate summary statistics. These loci had between 1 and 148 segregating sites, though when the data excluded all sites with missing data this ranged from 0 to 13 segregating sites. All of these had an F_{ST} less than 0.01, suggesting little population differentiation at these loci. Both the Alcohol dehydrogenase (*Adh*) and Phosphoglycerate kinase (*Pgk*) loci had negative values of Tajima's D (Table 4.2). *Adh* and *Pgm* had Kxy (average number of nucleotide substitutions per site between populations) higher than the transcriptome wide average of 0.08794.

To further assess which loci may be diversifying between these populations, we further examined transcripts with F_{ST}=1. For these transcripts, we searched for conserved protein

domains using Interproscan, and identified 224 conserved domains in these 534 transcripts. We then filtered to variant sites that were biallelic (1841 sites), genotyped in at least two individuals per population and had fixed genotypes between populations. This resulted in 70 variable SNPs across 49 transcripts. One of these lacked an ORF, 15 of the remaining 48 transcripts did not have fixed variants withing the ORF and Ka/Ks was not calculated for these loci. This left us with 33 transcripts with an F_{ST} of 1 between populations that we could calculate Ka/Ks for. The number of fixed variant sites for these transcripts range from 1 to 4. Nine of these loci had a Ka/Ks >1 indicating they may be under positive selection, though they only contained 1-2 variable sites and did not have significant p-value for this statistic. One locus had Ka/K≈1. The remaining 23 transcripts had a Ka/Ks between 0 and 0.34.

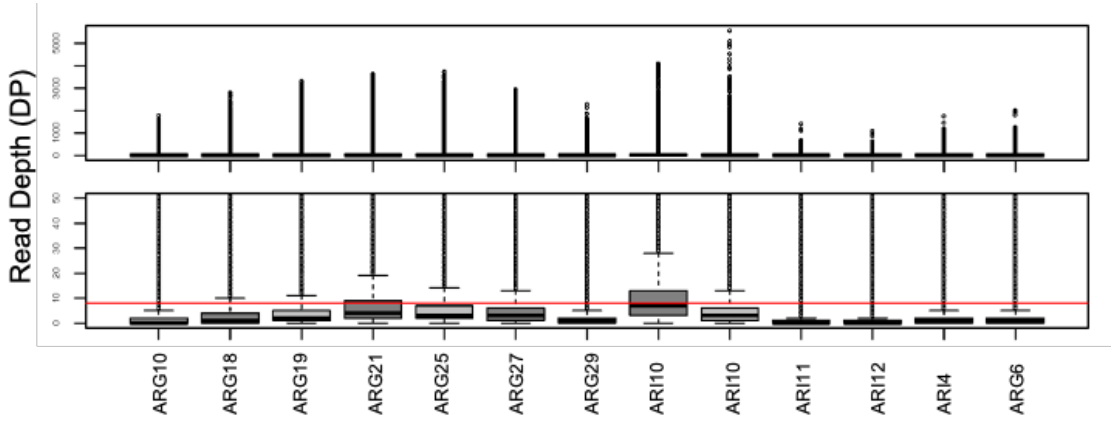

Figure 4.1 - Read depth per sample for all sites before filtering, bottom panel is zoomed into smaller values of coverage so the boxplots can be seen.

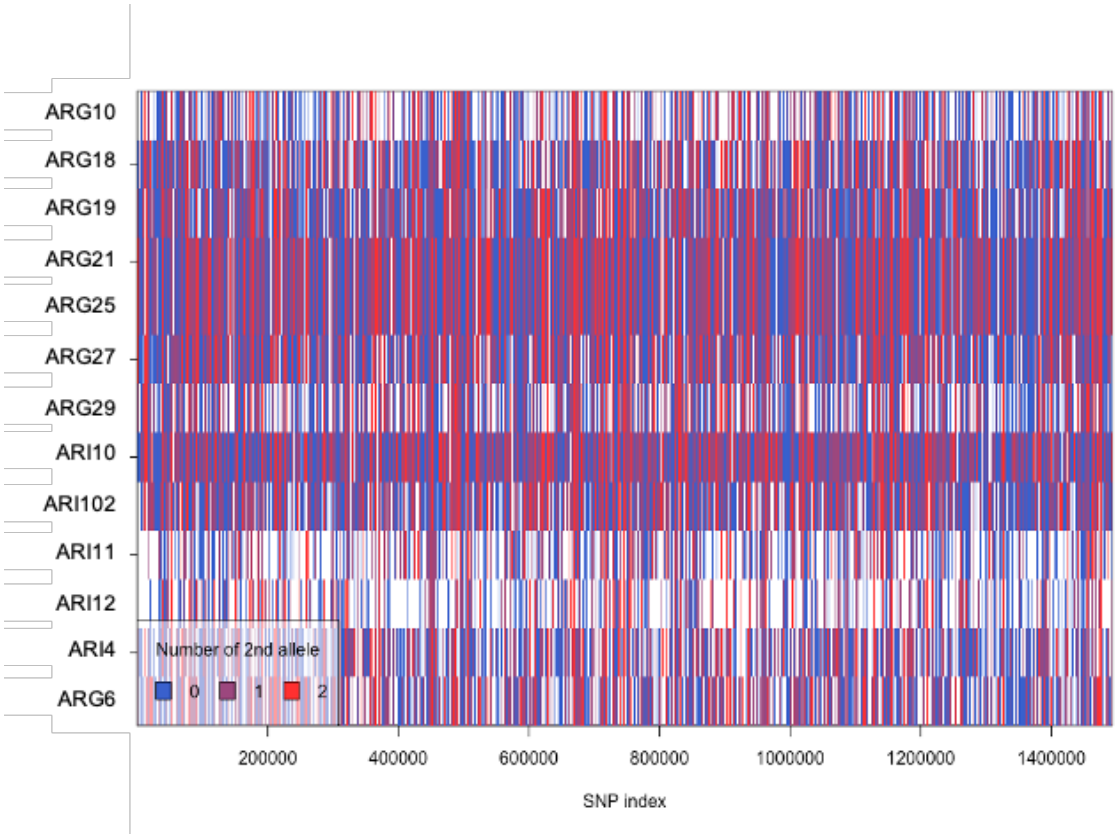

Figure 4.2 - Data Matrix for 695,095 variable SNPs, for 7 samples from Argentina with 41.45% missing genotypes, and 6 samples from Arica with 51.83% missing genotypes. White indicates missing data.

Table 4.1 - Number of variant SNPs missing genotypes per sample, % missing, and mean read depth (RD) of 695,095 variable sites.

Sample	ARG10	ARG18	ARG19	ARG21	ARG25	ARG27	ARG29	ARI10	ARI102	ARI11	ARI12	ARI4	ARI6
Missing	468838	314714	262068	157528	187323	231593	394968	100221	262216	498909	523343	392228	384745
% Missing	67.45	45.28	37.7	22.66	26.95	33.32	56.82	14.42	37.72	71.78	75.29	56.43	55.35
Mean RD	3.275	6.834	10.36	17.17	14.08	11.14	4.715	22.9	10.36	1.83	1.575	4.065	4.41

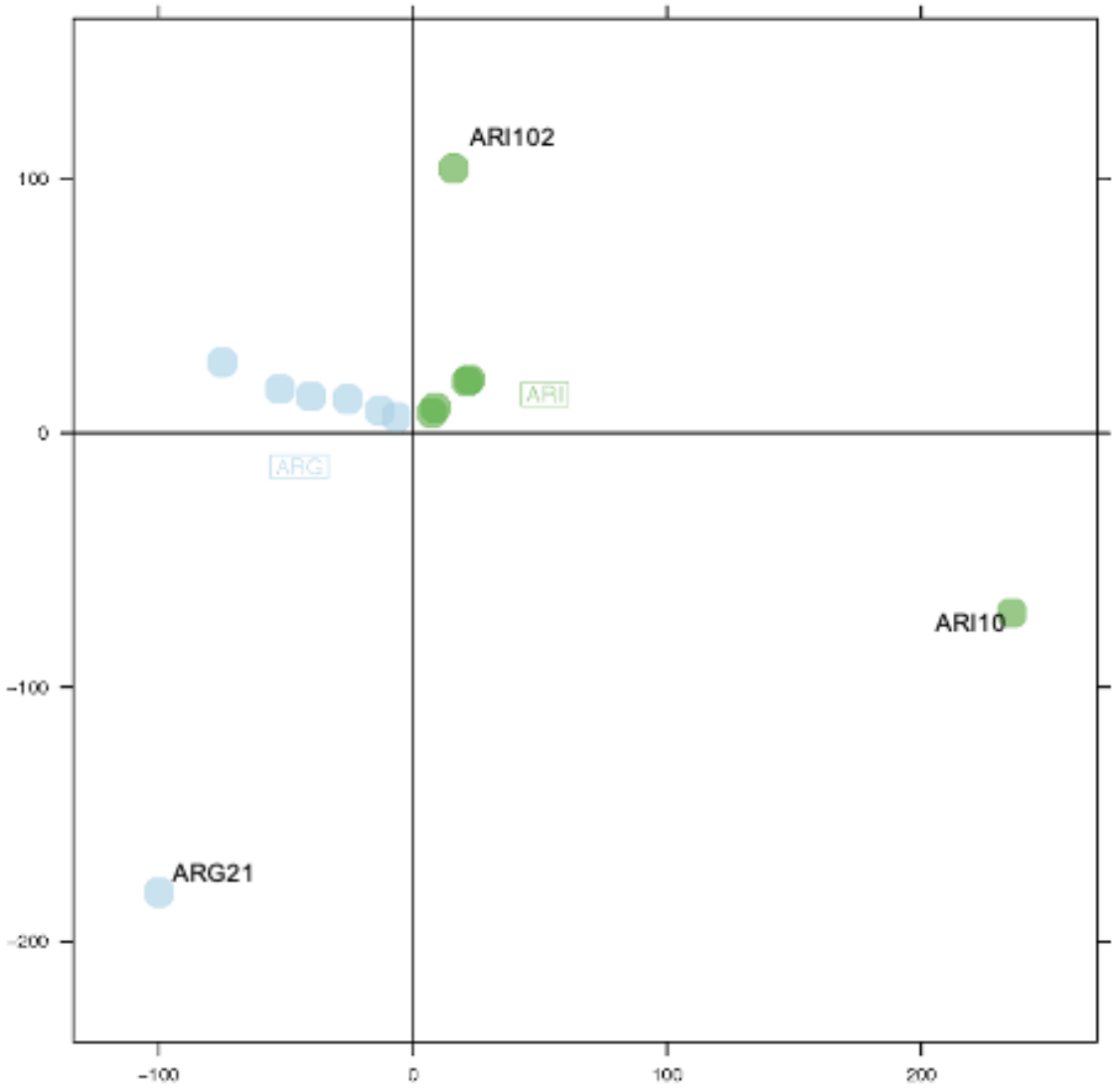

Figure 4.3 - PCA of all SNPs

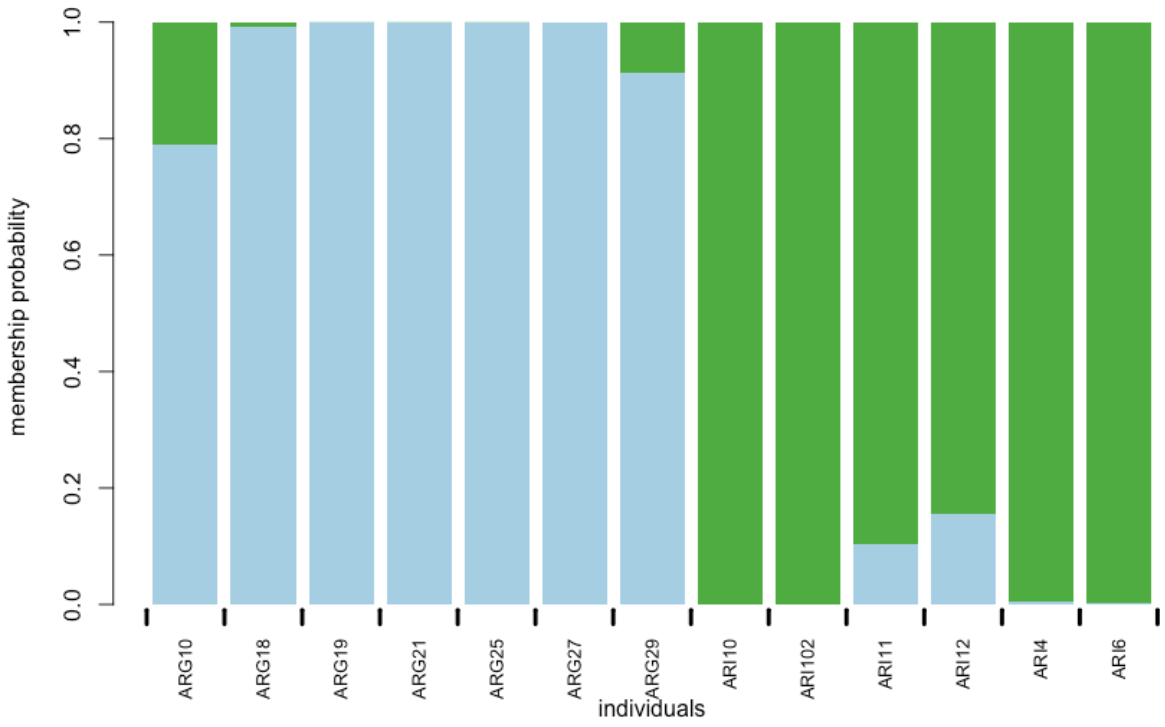

Figure 4.4 - Compoplot. Based on DAPC, bars indicate probability of reassignment to each group based on genotype. Assignment to the Argentina population is indicated by blue and assignment to the Arica population is indicated by green.

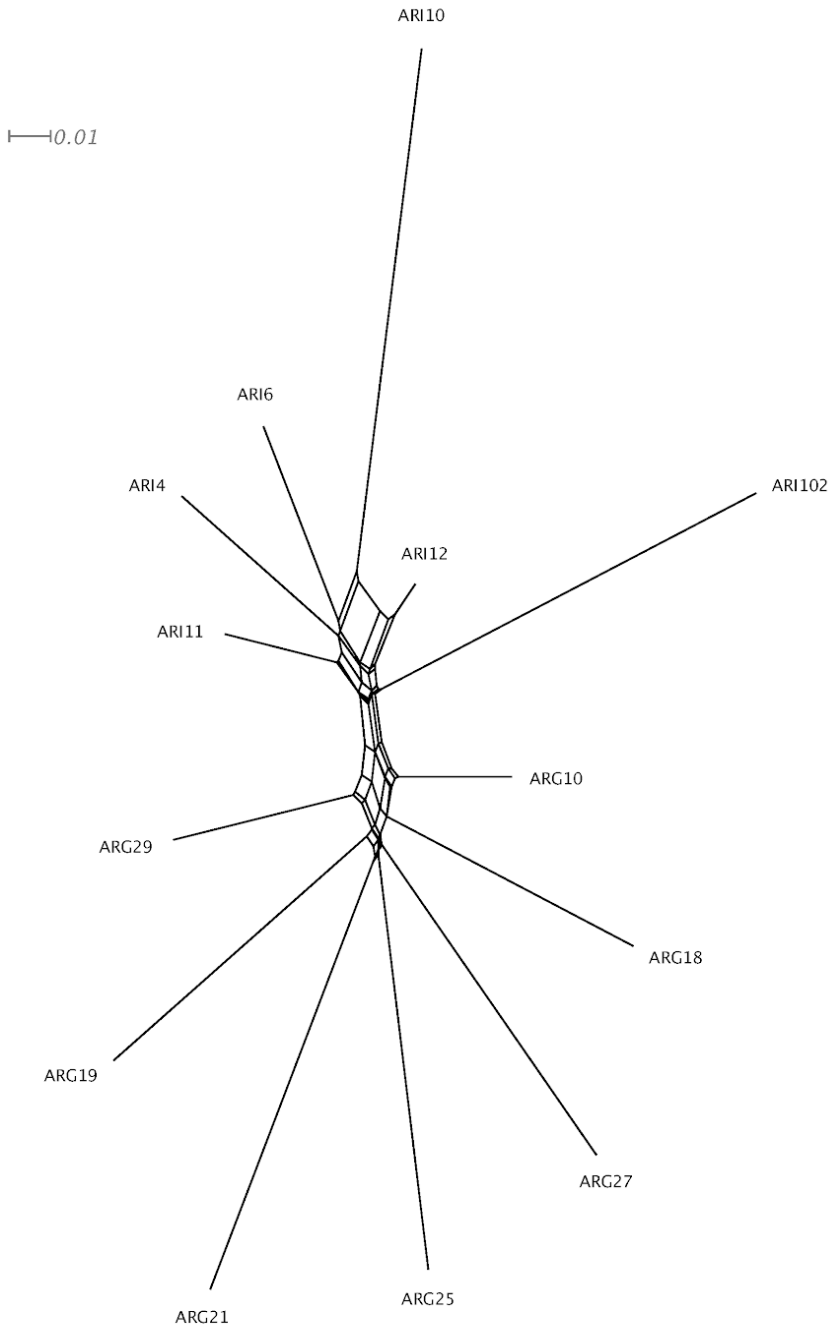

Figure 4.5 - Nei's 1972 genetic distance tree Plotted with SplitsTree V4.10, based on distance matrix for all individuals calculated in the R package StaMPP v1.6.1.

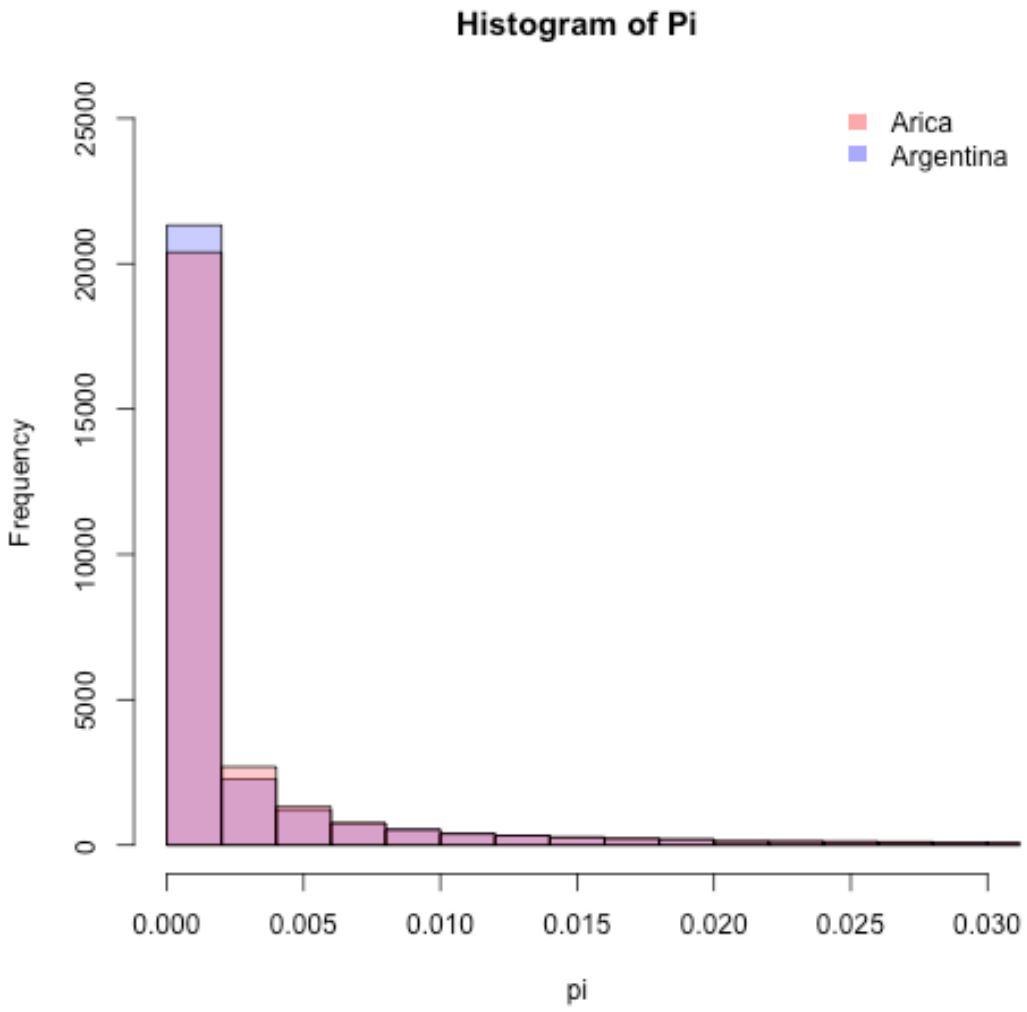

Figure 4.6 - Distribution of pi calculated with DnaSP for each population

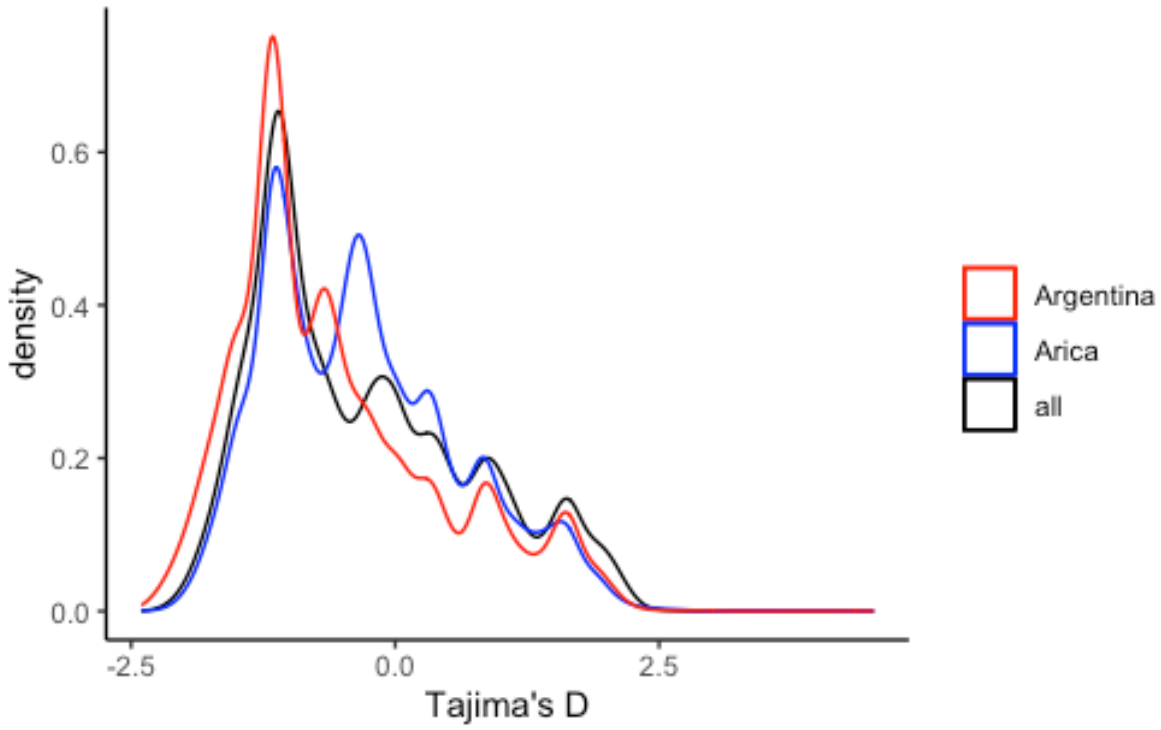

Figure 4.7 - Tajima's D calculated for each locus using DnaSP

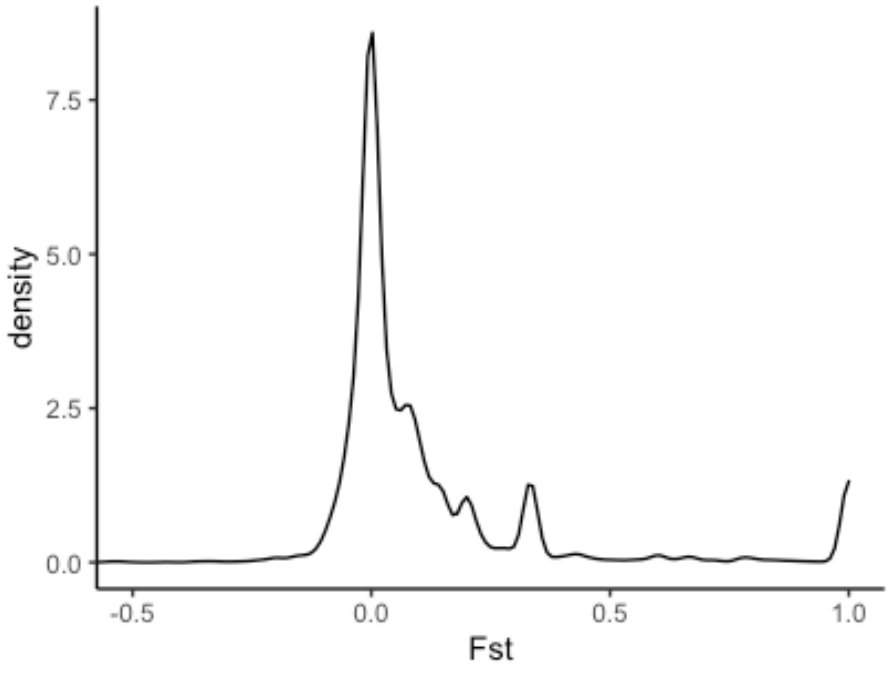

Figure 4.8 - Distribution of F_{ST} (Hudson 1992) between populations calculated with DnaSP

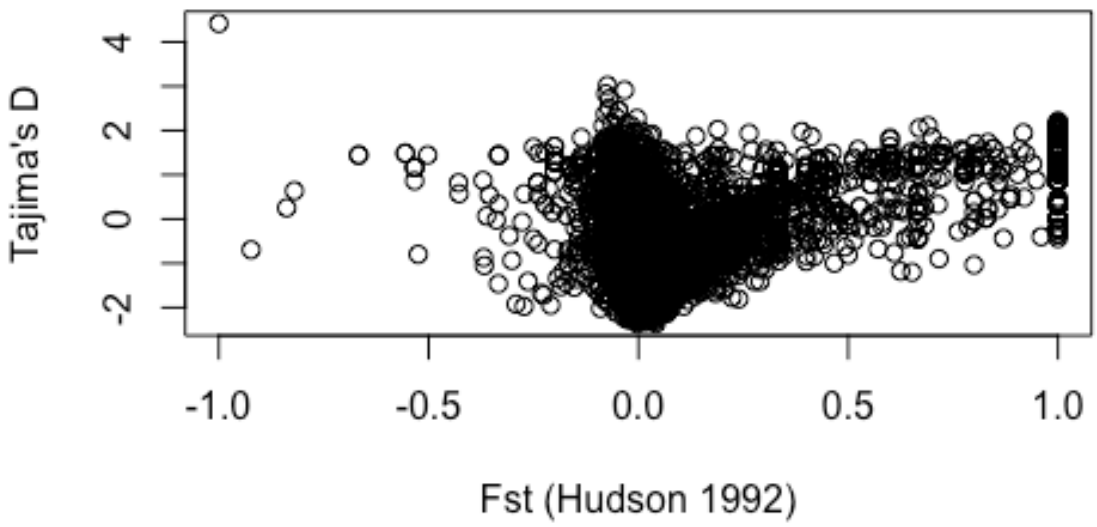

Figure 4.9 - F_{ST} (Hudson 1992) and Tajima's D per locus calculated in DnaSP.

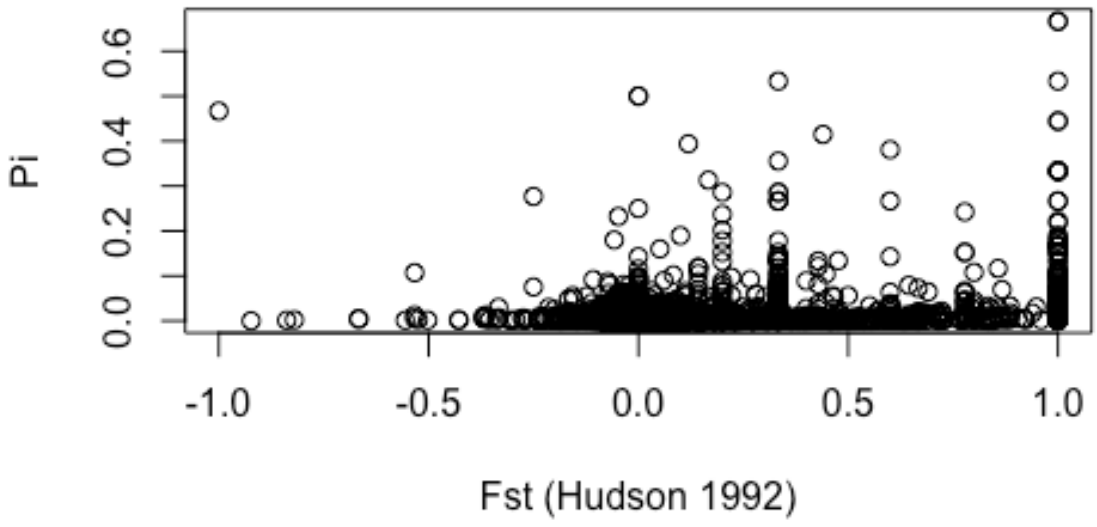

Figure 4.10 - F_{ST} (Hudson 1992) and Pi per locus calculated in DnaSP.

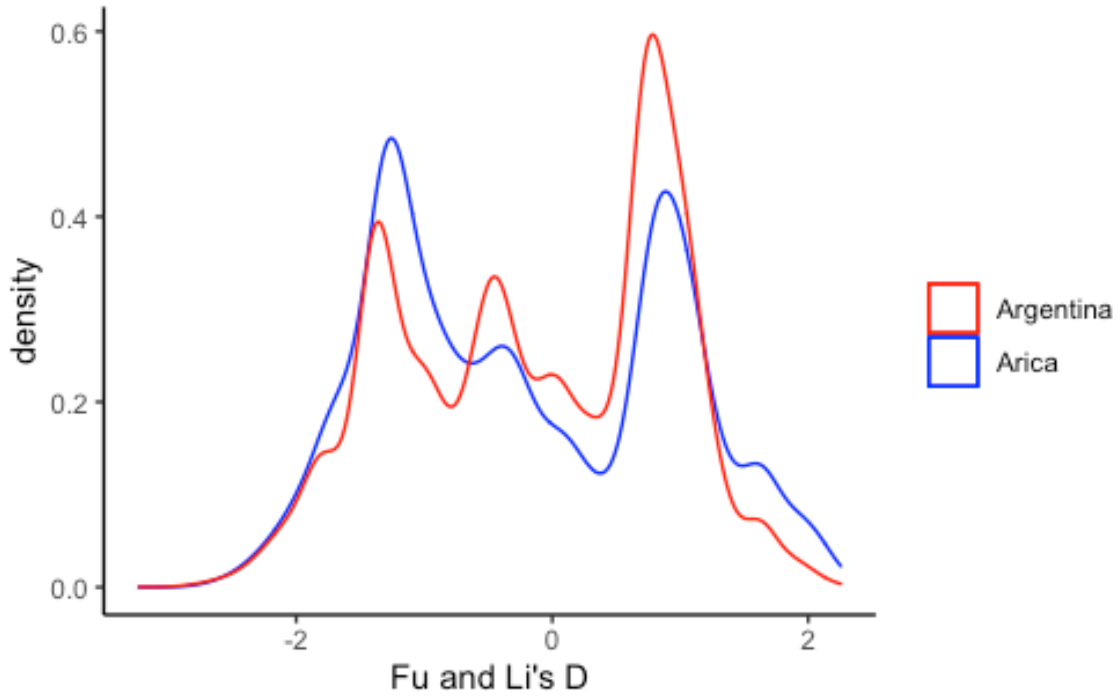

Figure 4.11 - Distribution of Fu and Li's D for each population, calculated in DnaSP.

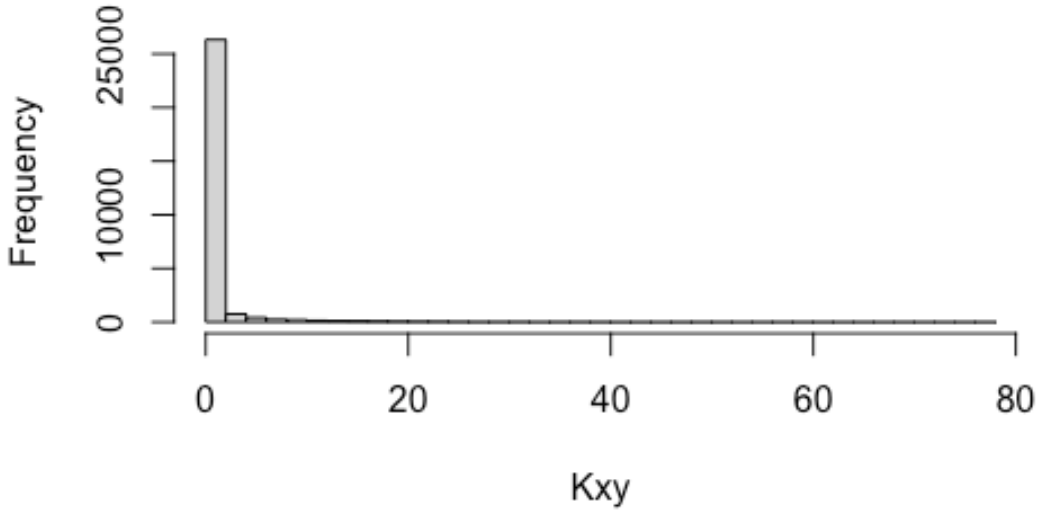

Figure 4.12 - Histogram of Kxy, the average number of differences between populations.

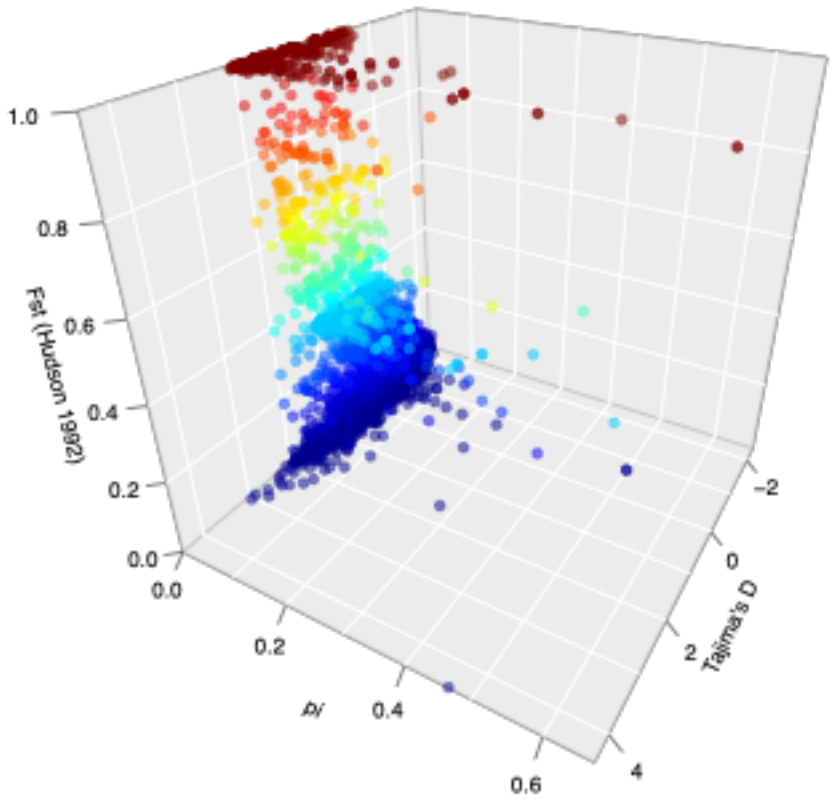

Figure 4.13 - F_{ST} (Hudson 1992), Tajima's D, and pi calculated in DnaSP for each transcript.

Discussion

A key goal of this exploration is to identify how exome diversity exhibits divergence between populations of a species (*Notochthamalus scabrosus* Darwin 1854) that spans an extraordinary latitudinal and environmental range. We know that these populations are genomically distinct (Ewers-Saucedo et al 2016), but the functional consequences of population divergence are better evaluated looking at transcribed diversity. In particular,

Table 4.2 - Summary statistics for metabolic loci calculated with DnaSP. N.A. indicated insufficient data for calculations.

					Overall		Arica		Argentina	
Gene name	**gene**	**F_{ST}**	**Kxy**	**Pi**	**TajimaD**	**FuLiD***	**TajimaD**	**FuLiD***	**TajimaD**	**FuLiD***
superoxide dismutase	Sod	n.a.	0	0	n.a.	n.a.	n.a.	n.a.	0.01499	0.80424
Phosphoglucomutase	Pgm	-0.0679	1.083	0.00079	0.17508	-0.8969	0.063534	-0.45894	0.40954	0.5052
Glucose dehydrogenase	DHGL	n.a.	0	0	n.a.	n.a.	n.a.	n.a.	-0.7671	-0.6138
Malic enzyme	E1JIZ4	n.a.	0	0	n.a.	n.a.	0.219917	0.972946	-0.6364	-1.3975
Alanine transaminase	ALAT	n.a.		0	n.a.	n.a.	n.a.	n.a.	-1.2898	-1.8817
Alcohol dehydrogenase*	Adh	n.a.	0	0	n.a.	n.a.	-0.98247	-1.28583	-1.1178	-1.0973
Dehydrogenase/ reductase**	Q9I7R3	0.0584	2.03571	0.00242	-1.4312	-1.8759	-0.78355	-0.99700	-0.4173	-0.5937
Glutamate dehydrogenase	Q8T453	n.a.	0	0	n.a.	n.a.	n.a.	n.a.	-0.7793	-0.9152
Mannose-6-phosphate isomerase	Mpi	n.a.	0	0	n.a.	n.a.	-0.05001	0.062207	n.a.	n.a.
Pyruvate dehydrogenase	PDK	n.a.	0	0	n.a.	n.a.	n.a.	n.a.	-0.2006	-0.4457
alkaline phosphatase	ALP	0.08197	2.17857	0.00122	-0.3627	0.81213	-0.82382	-0.71134	-0.5731	-0.1065
Fructose-bisphosphate aldolase	ALF	n.a.	0	0	n.a.	n.a.	-1.11173	-1.24341	0.60313	1.234
Phosphoglycerate kinase	PGK	0.07143	0.5000	0.00051	-1.4709	-1.3092	-0.52473	-0.17637	n.a.	n.a.
Phosphoglyceromutase	Q24450	0.07246	0.32857	0.00022	0.13869	0.62273	-0.18393	-0.28019	-0.809	0.65415
Isocitrate dehydrogenase	B7Z0E0	n.a.	0	0	n.a.	n.a.	n.a.	n.a.	-0.7036	-0.2236
Glycerol-3-phosphate dehydrogenase	GPDA	-0.0821	0.21428	0.00017	-0.2484	0.62273	-1.11173	-1.24341	-1.0372	-0.3536

*best blast hit to Nunez et al Adh sequence
**best blast hit to Drosophila Adh sequence, 23.114 % identity, contains a short chain dehydrogenase domain

Table 4.3 – Ka/Ks values and BLAST results for tblastn against the GenBank nucleotide database for transcripts with an F_{ST} of 1. *p <0.10

Transcript	Ka/Ks	P-Value (Fisher)	Len	Blast hit	Description	% id
NODE_104675_m.93123	0.3236	0.5794	342	XM_037236579	PREDICTED: *Pollicipes pollicipes* slowpoke-binding protein-like (LOC119112398), mRNA	71.2
NODE_109115_m.94562	0.0010	0.0000*	330	XM_037220417	PREDICTED: *Pollicipes pollicipes* transmembrane protein 94-like (LOC119097372), mRNA	79.8
NODE_121523_m.98528	0.0010	0.0774*	309	XM_037214153	PREDICTED: *Pollicipes pollicipes* serine/threonine-protein kinase pim-3-like (LOC119091384), transcript variant X2, mRNA	84.3
NODE_24555_m.13469	0.0010	0.0879*	1026	XM_037219144	PREDICTED: *Pollicipes pollicipes* rho guanine nucleotide exchange factor 10-like (LOC119096306), transcript variant X2, mRNA	63.3
NODE_24640_m.64961	0.9916	0.7499	420			
NODE_32303_m.15928	0.0010	0.0000*	642	XM_037236058	PREDICTED: *Pollicipes pollicipes* transient receptor potential channel pyrexia-like (LOC119112074), mRNA	55.5
NODE_52662_m.76421	0.3319	0.2225	591	XM_037219498	PREDICTED: *Pollicipes pollicipes* polyamine deacetylase HDAC10-like (LOC119096597), mRNA	45.5
NODE_5419_m.53999	50.000	0.3679	690	XM_037226737	PREDICTED: *Pollicipes pollicipes* uncharacterized LOC119103270 (LOC119103270), transcript variant X3, mRNA	82.2
NODE_71642_m.82814	50.000	0.3679	432	HG965802	Bartonella henselae, strain BM1374163 complete genome	35.6
NODE_74730_m.28165	48.008	0.4963	453	XM_037220349	PREDICTED: *Pollicipes pollicipes* protein arginine N-methyltransferase 1-like (LOC119097298), mRNA	85.4
NODE_75527_m.28377	0.0010	0.1369	336	XM_037216502	PREDICTED: *Pollicipes pollicipes* cuticle protein 7-like (LOC119093542), mRNA	83.6
NODE_84728_m.86972	50.000	0.3679	324			
NODE_93858_m.34511	0.0010	0.0000*	378	CP002379	*Pseudarthrobacter phenanthrenivorans* Sphe3 chromosome, complete genome	38.1
R9384835_m.1453	0.0010	0.0436*	798	XM_037224489	PREDICTED: *Pollicipes pollicipes* general transcription factor IIE subunit 1-like (LOC119101197), mRNA	86.5
S1484002_m.102625	0.0010	0.3679	303			
S5236363_m.104765	0.0010	0.3679	348	XM_037225248	PREDICTED: *Pollicipes pollicipes* tubby-related protein 4-like (LOC119101821), mRNA	54.3
S9353633_m.107442	0.0010	0.0468*	438	XM_037235576	PREDICTED: *Pollicipes pollicipes* transcription factor 12-like (LOC119111758), mRNA	76.9
S9444781_m.109043	0.0010	0.0517*	306	XM_037216681	PREDICTED: *Pollicipes pollicipes* histone-lysine N-methyltransferase trithorax-like (LOC119093679), transcript variant X2, mRNA	82.4
S9461781_m.109365	50.000	0.3679	315	XR_005094120	PREDICTED: *Pollicipes pollicipes* putative polypeptide N-acetylgalactosaminyltransferase 9 (LOC119097062), transcript variant X3, misc_RNA	72.4
S9466654_m.109508	0.0010	0.0899*	348	XM_037220628	PREDICTED: *Pollicipes pollicipes* uncharacterized LOC119097583 (LOC119097583), transcript variant X2, mRNA	67
TRINITY_DN10101_c1_g1_i11_m.154605	0.0010	0.0669*	594			

TRINITY_DN117406_c0_g1_i1_m.172924	0.0975	0.0016*	540	XM_037233541	PREDICTED: *Pollicipes pollicipes* E3 ubiquitin-protein ligase ZNRF2-like (LOC119109787), mRNA	47.7
TRINITY_DN16492_c0_g1_i1_m.128952	0.0010	0.0320*	297	XM_037227238	PREDICTED: *Pollicipes pollicipes* oxidation resistance protein 1-like (LOC119103634), transcript variant X2, mRNA	75.6
TRINITY_DN20265_c0_g1_i1_m.133370	50.000	0.3679	408	XM_012900718	*Acytostelium subglobosum* LB1 hypothetical protein partial mRNA	35.4
TRINITY_DN4457_c0_g1_i2_m.159994	0.0010	0.0711*	648	XM_017263736	PREDICTED: *Drosophila elegans* ras-related protein Rab-11A (LOC108140755), mRNA	89.7
TRINITY_DN45169_c0_g1_i6_m.138253	50.000	0.3679	297	XM_032974990	PREDICTED: *Petromyzon marinus* stress response protein NST1-like (LOC116954458), transcript variant X2, mRNA	41.5
TRINITY_DN45228_c0_g2_i1_m.137559	0.0010	0.0927*	417	XM_037233772	PREDICTED: *Pollicipes pollicipes* uncharacterized LOC119110025 (LOC119110025), mRNA	86.6
TRINITY_DN4825_c0_g1_i4_m.151136	0.0010	0.0838*	759	XM_026423284	PREDICTED: *Frankliniella* occidentalis zinc finger and SCAN domain-containing protein 2-like (LOC113206968), transcript variant X1, mRNA	35.1
TRINITY_DN48479_c0_g2_i1_m.162257	50.000	0.4882	420	XM_037217963	PREDICTED: *Pollicipes pollicipes* larval/pupal cuticle protein H1C-like (LOC119095067), mRNA	64.7
TRINITY_DN57675_c0_g2_i1_m.137060	50.000	0.5863	717	XM_037234201	PREDICTED: *Pollicipes pollicipes* nuclear envelope integral membrane protein 1-like (LOC119110390), mRNA	81.3
TRINITY_DN70576_c0_g1_i2_m.156738	0.0010	0.0948*	318	XM_037233928	PREDICTED: *Pollicipes pollicipes* CD109 antigen-like (LOC119110191), mRNA	91.1
TRINITY_DN73108_c0_g1_i1_m.151285	0.0010	0.0945*	351	XM_037234980	PREDICTED: *Pollicipes pollicipes* homeobox protein Hox-B7-like (LOC119111151), mRNA	100
TRINITY_DN92966_c0_g1_i1_m.149860	0.0010	0.0828*	453	XM_037222706	PREDICTED: *Pollicipes pollicipes* calcium-dependent secretion activator-like (LOC119099637), partial mRNA	90.7

we proposed that metabolic genes contribute heavily to the overall divergence of populations in *N. scabrosus*. Marden (2010) and Skibinski & Ward (2004) have pointed out that these genes are often targets of natural selection and the strength of selection is associated with diversity; how these metabolic loci function in maintaining species cohesion across large geographic ranges is of equal interest (Nunez et al 2020).

However, working with non-model organisms is challenging even to identify particular loci as there are no functional annotations and the divergence with 'model' organisms is quite large. In this instance, *Notochthamalus* (Cirripedia: Balanomorpha: Chthamalidae) is likely 100 ma divergent (Pérez-Losada et al., 2014) (from any barnacle species with a genome at all, like *Balanus amphitrite* (Kim et al., 2019), or the well-studied *Semibalanus balanoides* (Cirripedia: Balanomorpha: Balanidae). Divergence times to other well-characterized arthropods (e.g. *Daphnia, Drosophila*) exceed 500 million years. Targets of selection are thus perhaps unlikely to be uniquely identifiable, as opposed to highly conserved regions, as annotation is indirect in these non-model organisms. Nevertheless, we have identified sequences we believe correspond to several allozyme and metabolic genes in *Notochthamalus scabrosus*.

Alcohol dehydrogenase (*Adh*), had a negative value of Tajima's D in both populations suggesting that selection is acting to reduce variation for this gene in both populations. The *Dehydrogenase reductase* locus, which was the best Blast hit for the Drosophila *Adh* sequence, had an above average Kxy and negative Tajima's D, the combination of negative Tajima's D and high Kxy suggests this locus is under different selective

pressures in the Northern and Southern populations. The partial transcript of mannose-6-phosphate isomerase (*Mpi*), has a negative Tajimas's D for the Arica population, however Kxy for *Mpi* was zero suggesting this locus may be subject to purifying selection. *Mpi* has been a particular focus for evolutionary ecology in marine organisms, especially barnacles, as it appears to indicate the need for distinct enzymatic function in heterogeneous habitats (Schmidt et al 2000). In contrast to these allozyme genes, phosphoglucomutase (*Pgm*) appears to be evolving neutrally or may be under balancing selection in the Argentina population. This gene may be of particular interest to investigate further as Nunez et al., 2020 found a Tajima's D of -1.2 for *Pgm* in *Semibalanus balanoides*.

While our data shows clear differences between the Northern and Southern populations, it also captures the variation between individuals within a population. As previous studies have shown ample dispersal of pelagic larvae across the species range, loci which show differentiation between Northern and Southern populations are candidates to evaluate whether drift or selection are driving that divergence. Though our initial focus was on recognized metabolic enzymes as in Skibinski & Ward (2004), clearly a diverse range of transcribed genes respond to the huge spatial and environmental distance between these range endpoints.

To further investigate which genes may be experiencing the most diversifying selection we examined conserved domain of transcripts with high F_{ST}. Of the 28,636 transcripts we were able to calculate F_{ST} for, 535 transcripts had an $F_{ST} = 1$. For these transcripts, we

searched for conserved protein domains using Interproscan, and identified 224 conserved domains in these transcripts. Additionally, we calculated Ka/Ks using fixed differences in the coding region of these transcripts. Of note, one of these transcripts (NODE_47854_m.20445) contained an Alcohol dehydrogenase N-terminal protein domain, though after filtering steps this locus lacked sufficient data to calculate Ka/Ks; this is not one of the transcripts we identified as *Adh* using BLAST. Among these transcripts we also identified several other protein domains associated with metabolic genes such as Galactose mutarotase N-terminal barrel, 6-phosphogluconate dehydrogenase, 3-hydroxyisobutyrate dehydrogenase, AAA ATPase AAA+ lid domain, and Nuclear respiratory factor-1 activation binding domain. The presence of these metabolic related protein domains in our transcripts with an F_{ST} of 1 suggests that metabolism is likely under divergent selection between northern and southern populations. This may also indicate that our transcriptome is composed of fragmented or duplicated transcripts, as some of these domains would be traditionally found in the metabolic loci we examined.

Of the transcripts with an F_{ST} of 1 that we were able to calculate Ka/Ks for, most have a Ka/Ks near zero suggesting purifying selection is constraining non-synonymous substitutions, one locus had a Ka/Ks of approximately 1 suggesting it is evolving neutrally, while 9 transcripts had a Ka/Ks greater than 1 suggesting they may experience directional selection between the populations. Though we note that due to a combination of missing data and the filtering steps required prior to calculating these statistics our analysis was limited by very few variant sites per transcript and had limited statistical

power to interpret these loci that may be under directional selection. Interestingly, one transcript with an elevated Ka/Ks appears to be a stress response gene, suggesting the differential selective pressure of environment between the northern and southern populations may be driving divergence at this locus. Additionally, two of the loci with an F_{ST} of 1 and were cuticle proteins, one of which had a Ka/Ks near zero and the other had an elevated Ka/Ks. Since one these had high population differentiation but not fixed non-synonymous differences in the coding region, there may be SNPs that are non-coding that account for the population differentiation we are observing and may also affect expression of these transcripts.

These findings as well as those from Ewers-Saucedo 2016, illustrate the evolutionary divergence between the Northern and Southern extent of *N. scabrosus*. These populations are separated by a large geographic distance, with a strong genetic transition in southern Chile, and a large difference in oceanic and intertidal environments all driving divergence between these populations. The northern part of Chile experiences warmer water, greater terrestrial aridity, less persistent upwelling (Thiel, 2007), and is at the 'downstream' end of the distribution relative to current-driven gene flow. The southern population sampled here, from Argentina, is distinct in almost every way.

Our findings here illustrate the strong selective forces acting on populations at the ends of this species range, creating genetically differentiated populations in the face of high dispersal. The data presented here, including the *de novo* transcriptome assembly of *N. scabrosus*, provide a resource for not only understanding divergence within *N. scabrosus*,

but for chthamalid barnacles overall. This family of barnacles is globally distributed and exhibits extraordinary levels of cryptic diversity (Wares, 2020) that guide our understanding of intertidal biogeography (Ewers-Saucedo et al., 2016). As it becomes more clear that biogeography involves not just vicariance and dispersal as primary mechanisms but also adaptation, this deeper exploration will be important for recognizing global patterns.

References

Álvarez-Noriega, M., Burgess, S. C., Byers, J. E., Pringle, J. M., Wares, J. P., & Marshall, D. J. (2020). Global biogeography of marine dispersal potential. *Nature Ecology & Evolution*, *4*(9), 1196–1203.

Antonovics, J. (1976). The nature of limits to natural selection. *Annals of the Missouri Botanical Garden*, 224–247.

Connell, J. H. (1961). The influence of interspecific competition and other factors on the distribution of the barnacle *Chthamalus stellatus*. *Ecology*, 710–723.

Darwin, C. (1854). A monograph on the subclass. *Cirripedia, with Figures of All the Species. Ray Soc. Publ.(London)*.

Deutsch, C., Penn, J. L., & Seibel, B. (2020). Metabolic trait diversity shapes marine biogeography. *Nature*, 1–6.

Ewers-Saucedo, C., Pringle, J. M., Sepúlveda, H. H., Byers, J. E., Navarrete, S. A., & Wares, J. P. (2016). The oceanic concordance of phylogeography and biogeography: A case study in *Notochthamalus*. *Ecology and Evolution*, *6*(13), 4403–4420.

Hare, M. P., Nunney, L., Schwartz, M. K., Ruzzante, D. E., Burford, M., Waples, R. S., . . . Palstra, F. (2011). Understanding and Estimating Effective Population Size for Practical Application in Marine Species Management. *Conservation biology, 25*(3), 438-449. doi:10.1111/j.1523-1739.2010.01637.x

Häussermann, V., & Försterra, G. (2009). Marine benthic fauna of Chilean Patagonia. *Nature in Focus, Santiago*.

Huson, D. H., & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution, 23*(2), 254-267.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., & Nuka, G. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240.

Kim, J.-H., Kim, H., Kim, H., Chan, B. K., Kang, S., & Kim, W. (2019). Draft genome assembly of a fouling barnacle, *Amphibalanus amphitrite* (Darwin, 1854): The first reference genome for Thecostraca. *Frontiers in Ecology and Evolution*, *7*, 465.

Laughlin, K. M., Ewers, C., & Wares, J. P. (2012). Mitochondrial lineages in *Notochthamalus scabrosus* as indicators of coastal recruitment and interactions. *Ecology and Evolution*, *2*(7), 1584–1591.

MacManes, M. D. (2018). The Oyster River Protocol: A multi-assembler and kmer approach for de novo transcriptome assembly. *PeerJ*, *6*, e5428.

Marden, J. H. (2013). Nature's inordinate fondness for metabolic enzymes: Why metabolic enzyme loci are so frequently targets of selection. *Molecular Ecology*, *22*(23), 5743–5764.

Nunez, J. C., Flight, P. A., Neil, K. B., Rong, S., Eriksson, L. A., Ferranti, D. A., Rosenblad, M. A., Blomberg, A., & Rand, D. M. (2020). Footprints of natural selection at the mannose-6-phosphate isomerase locus in barnacles. *Proceedings of the National Academy of Sciences*, *117*(10), 5376–5385.

Pappalardo, P., Pringle, J. M., Wares, J. P., & Byers, J. E. (2015). The location, strength, and mechanisms behind marine biogeographic boundaries of the east coast of North America. *Ecography*, *38*(7), 722–731.

Pérez-Losada, M., Høeg, J. T., Simon-Blecher, N., Achituv, Y., Jones, D., & Crandall, K. A. (2014). Molecular phylogeny, systematics and morphological evolution of the acorn barnacles (Thoracica: Sessilia: Balanomorpha). *Molecular Phylogenetics and Evolution*, *81*, 147–158.

Pringle, J., M. , & Wares, J., P. (2007). Going against the flow: maintenance of alongshore variation in allele frequency in a coastal ocean. *Marine Ecology Progress Series, 335*, 69-84. Retrieved from https://www.int-res.com/abstracts/meps/v335/p69-84/

Schmidt, P. S., Bertness, M. D., & Rand, D. M. (2000). Environmental heterogeneity and balancing selection in the acorn barnacle *Semibalanus balanoides*. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *267*(1441), 379–384.

Schmidt, P. S., & Rand, D. M. (1999). Intertidal microhabitat and selection at Mpi: Interlocus contrasts in the northern acorn barnacle, *Semibalanus balanoides*. *Evolution*, *53*(1), 135–146.

Schmidt, P. S., & Rand, D. M. (2001). Adaptive maintenance of genetic polymorphism in an intertidal barnacle: Habitat-and life-stage-specific survivorship of Mpi genotypes. *Evolution*, *55*(7), 1336–1344.

Sexton, J. P., McIntyre, P. J., Angert, A. L., & Rice, K. J. (2009). Evolution and ecology of species range limits. *Annual Review of Ecology, Evolution, and Systematics*, *40*.

Skibinski, D. O., & Ward, R. D. (2004). Average allozyme heterozygosity in vertebrates correlates with Ka/Ks measured in the human-mouse lineage. *Molecular Biology and Evolution*, *21*(9), 1753–1759.

Thiel, M. (2007). *The Humboldt current system of northern and central Chile*.

Wares, J. P. (2020). Small, flat, and gray: Cryptic diversity in chthamalid barnacles in the global context of marine coastal biogeography (Cirripedia: Balanomorpha: Chthamalidae). *The Journal of Crustacean Biology*, *40*(1), 1–16.

Zakas, C., Binford, J., Navarrete, S. A., & Wares, J. P. (2009). Restricted gene flow in Chilean barnacles reflects an oceanographic and biogeographic transition zone. *Marine Ecology Progress Series*, *394*, 165–177.

Zakas, C., & Wares, J. P. (2012). Consequences of a poecilogonous life history for genetic structure in coastal populations of the polychaete *Streblospio benedicti. Molecular ecology, 21*(22), 5447-5460. doi:10.1111/mec.12040

Zakas, C., Jones, K., & Wares, J. P. (2014). Homogeneous nuclear background for mitochondrial cline in northern range of *Notochthamalus scabrosus*. *G3: Genes, Genomes, Genetics*, *4*(2), 225–230.

Zhang, Z., Li, J., Zhao, X.-Q., Wang, J., Wong, G. K.-S., & Yu, J. (2006). KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics, proteomics & bioinformatics, 4*(4), 259-263.

Zhang, Z., Xiao, J., Wu, J., Zhang, H., Liu, G., Wang, X., & Dai, L. (2012). ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochemical and biophysical research communications, 419*(4), 779-781.

CHAPTER 5

NOVEL DIVERSITY AND MOLECULAR EVOLUTION IN THE VERTEBRATE DNA METHYLTRANSFERASE GENE FAMILY[4]

Introduction

The role of genomic modifiers as a means of regulating plastic and developmental responses has been under intense study in the past decade. Much of this work has focused on the addition of methyl groups to DNA, a mechanism that is controlled by DNA methyltransferase (Dnmt) genes (Jurkowska, Jurkowski, & Jeltsch, 2011; Klose & Bird, 2006). Dnmts regulate the addition and maintenance of methyl groups to DNA and RNA (Jurkowska et al., 2011; Klose & Bird, 2006). Addition of methyl groups to nucleic acids change how those molecules interact with other nucleic acids and proteins effecting gene expression (Jurkowska et al., 2011; Klose & Bird, 2006) and chromosome stability. DNA methylation is a component of developmental differentiation and metabolic regulation, but can also promote clear phenotypic differences in morphology, physiology, or behavior (Adams, Vinkenoog, Spielman, Dickinson, & Scott, 2000; Robert Kucharski, Maleszka, Foret, & Maleszka, 2008; R Kucharski, Maleszka, & Maleszka, 2016). Though many aspects of the role of DNA methylation and its regulation are still unclear, studies which mutate or delete Dnmts have shown that functional Dnmts are essential for life

[4] Bobier, K. E., Freeman, B., and Wares, J.

(Adam J. Bewick, Zachary Sanchez, et al., 2019; Li, Bestor, & Jaenisch, 1992; Okano, Bell, Haber, & Li, 1999).

One of the most common chemical modifications to DNA in eukaryotes is 5 methylcytosine, the addition of methyl groups to the 5' position of the DNA cytosine ring, from here in we will refer to this specific modification as DNA methylation. Proper control of DNA methylation is important for development and vigor in many species. In addition to cellular mechanisms that regulate DNA methylation, there can also be a strong influence of environmental factors on an organism's methylome (the methylation state of nucleic acids in the genome). Many studies have exposed organisms to extreme environments, such as heat shock or concentrated toxins, and observed changes in either the organism's methylome or the methylome of the organism's progeny. Despite these laboratory studies it is still unclear how variation in natural environment can change a methylome and how these environmental differences interact with genes regulating the methylome, or how genetic variation in these genes changes these interactions. One study linked differences in DNA methylation patterns to behavioral isolation between populations of darters, a diverse group of Perciform fishes (T. A. Smith, Martin, Nguyen, & Mendelson, 2016). DNA methylation patterns can also change with age in many species (Parrott & Bertucci, 2019), adding demographic variation within populations. DNA methylation therefore has the potential to contribute to species diversity through ecological speciation or sexual selection.

The extent to which variation in methylation patterns among individuals and locations highlights the importance of plasticity for species to persist in broad environmental niches (Stamp & Hadfield, 2020). Additionally, studies have found differences in DNA methylation between wild fishes and captive reared fishes. Recent studies have shown that Pacific salmon raised in hatcheries have notable differences in their DNA methylation patterns compared to wild salmon (Gavery, Nichols, Goetz, Middleton, & Swanson, 2018; Le Luyer et al., 2017). These differences can be correlated to phenotypic differences between hatchery and wild salmon, such as fitness, swimming endurance, and predator avoidance. The dynamic nature of DNA methylation and its importance for gene regulation and genome stability point to the potential importance of DNA methylation in phenotypic plasticity and adaptive potential.

The role and regulation of DNA methylation can vary widely across taxa. In mammals and plants, methylation in the promoter region is generally associated with transcriptional suppression, while gene body methylation is associated with increased gene expression in mammals. The relationship between gene expression and gene body methylation in plants is still not fully understood, with some evidence supporting a positive relationship and other evidence showing no relationship (Adam J. Bewick, Zhang, Wendte, Zhang, & Schmitz, 2019; Muyle & Gaut, 2018; Schmitz, Lewis, & Goll, 2019). Despite this shared pattern between plants and animals, fungi appear to lack gene body methylation (Adam J Bewick et al., 2019) suggesting the similar effects of gene body methylation in plants and animals potentially derived independently in these two groups or was lost in fungi. Recent studies have emphasized the importance of epigenetic marks (histone

modification and DNA methylation) for chromosome packaging, genome stability, and gene regulation. *Caenorhabditis elegans* and several budding yeast species have lost their Dnmt genes and lack 5mC-DNA methylation, for example, limiting the generality of what we can learn from these model organisms (Capuano, Mülleder, Kok, Blom, & Ralser, 2014; Colot & Rossignol, 1999; Ponger & Li, 2005; Simpson, Johnson, & Hammen, 1986). Groups of insects, plants, and fungi have had losses and duplications of methyltransferase genes (Adam J Bewick et al., 2019; Adam J Bewick, Vogel, Moore, & Schmitz, 2017). Species that have naturally lost methyltransferase genes appear to lack negative effects from these losses and the addition of functional Dnmts to their genome can be lethal (Lyko et al., 1999). Conversely when these genes are mutated in closely related species, severe effects including infertility and death are observed (Adam J. Bewick, Zachary Sanchez, et al., 2019).

The DNA methyltransferases which mediate 5mC are composed of several subfamilies with varying functions: Dnmt1 and Dnmt3. Dnmt1 is a maintenance methyltransferase, it adds methyl-groups to the newly synthesized strand during DNA replication. Dnmt3 genes have a de novo methylation function and can add methyl groups to previously unmethylated DNA. Dnmt3 was duplicated in early vertebrate evolution, likely during the 2R vertebrate genome duplication, leading to two paralogs, Dnmt3A and Dnmt3B. Trdmt1, previously known as Dnmt2, has sequence homology with DNA methyltransferases, however most evidence to date suggests it has limited functions in DNA methylation and instead methylates tRNA (Goll et al., 2006).

Our understanding of DNA methyltransferase genes in vertebrates pulls heavily from a few mammal species, including mouse and human, and one fish species, Danio rerio (zebrafish). It's reasonable to assume this diverse lineage may harbor unacknowledged variation in the DNA methyltransferase gene family as we observed in other groups. In fact, by just comparing mouse, human, and zebrafish we can readily observe that there has been an expansion of Dnmt3 paralogs in zebrafish compared to mammals. We know there is variation in the function, importance, and presence of genes in the Dnmt gene family even among closely related taxa, such as among beetles and fungi (Adam J Bewick et al., 2019; Adam J Bewick et al., 2017). However, we don't know how this gene family varies among fishes (from here on I will use the term fishes to refer to the paraphyletic group of vertebrates including Agnatha, Chondrichthyes, and Osteichthyes, but excluding Tetrapoda) despite fishes composing greater than 50% of vertebrate species.

Diversity of this gene family among fishes is of interest because, contrasts in evolution and retention in other taxa, as above, demonstrate the potential for variation in utilization of these genes among fishes and other vertebrates. One clear addition has been the gain of a calponin homology (CH) domain in two of four Dnmt3b paralogs (dnmt3bb.2 and dnmt3ba) with unknown function described in zebrafish (Shimoda, Yamakoshi, Miyake, & Takeda, 2005). By examining Dnmt genes across fishes and vertebrates we can clarify the evolutionary origins of this protein domain in Dnmts and characterize how widespread it is among fishes.

A huge taxonomic gap remains in our understanding of the importance, effects, and transmissibility of DNA methylation. Ray-finned fishes are the most diverse clade of vertebrates (Helfman, Collette, Facey, & Bowen, 2009), yet have been largely overlooked in studies of DNA methylation. A recent study has examined the evolution of methylation related genes in chordates, though their study was limited to 27 species (Liu, Hu, Panserat, & Marandel, 2020). Here, we investigate the molecular evolution of the methyltransferase gene family in vertebrates with a focus on fishes. By definition, Dnmt genes have a DNA methyltransferase domain, a functional or structural unit of a folded protein, that has conserved motifs at the protein sequence level (Ryazanova, Abrosimova, Oretskaya, & Kubareva, 2012). We can utilize available annotated genomes from jawless fishes, ray-finned fishes, lobe-finned fishes, and tetrapods, representing 500 million years of divergence (Helfman et al., 2009), to identify and compare orthologous copies across these taxa. This broad taxonomic sample allows us to assess the retention of methylation genes as well as gene duplications and losses in the DNMT gene family. Perhaps the most interesting aspect of this investigation will be examining how the gene family has evolved since the whole genome duplication (320-350mya) (Glasauer & Neuhauss, 2014) in the lineage of teleost fishes, as well as subsequent duplication events such as the salmonid whole genome duplication.

Examination of the Dnmt gene family in fishes provides a unique opportunity to study evolution of a gene family after known gene and genome duplication events across a diverse array of taxa. In this study I examine the evolution of DNA methyltransferase (Dnmt) genes across vertebrates and compare sequence divergence for 232 vertebrate

species. The goals of this study were to 1) identify and quantify Dnmt genes in vertebrate lineages, 2) characterize retention and loss of genes after duplication events, 3) estimate selection pressure of Dnmt gene sequence, 4) formulate a new hypothesis for the evolution of Dnmt genes in relation to duplication events for vertebrates 5) determine the evolutionary origin of CH domains in Dnmt3b genes.

Methods

To assess the gene family evolution of Dnmts in vertebrates, we identified annotated proteins from genomes in the Ensembl Genome Browser (Fernández & Birney, 2010) and NCBI (Pruitt, Tatusova, & Maglott, 2007) that contained a C-5 cytosine specific DNA methylase protein domain (Protein family database (Pfam) identifier: PF00145). Protein sequences from each genome were run through InterProScan version 5.41-78.0 (Finn et al., 2016; P. Jones et al., 2014) to identify protein domains from the Protein Family (Pfam) database (Finn et al., 2016). The output from InterProScan was then run through a custom Perl script to pull out sequences that specifically contained the C-5 cytosine specific DNA methylase protein domain (Protein family database (Pfam) identifier: PF00145). We identified methyltransferase domain containing protein and transcript sequences from genomes of 231 vertebrate species including 1 jawless fish, 3 cartilaginous fishes, 1 lobe finned fish, 40 bony/ray finned fishes, 3 amphibians, 13 reptiles, 64 birds, and 106 mammals, as well as an outgroup species (*Ciona intestinalis*). When genomes for a species were available on both Ensembl and NCBI, I used the Ensembl version for consistency.

For each protein sequence identified as contacting a DNA methylase domain, we used custom scripts to search the gff file to extract the gene and transcript ids. If genes had multiple transcripts annotated, the sequence with the longest open reading frame were utilized in analyses. The Dnmt1 protein is characterized by DMAP binding domain, a Dnmt1-RFD (Cytosine_MeTrfase1_RFD), a zinc finger CXXC, a two interval BAH domain, and a DNA methylase domain. Dnmt3 proteins generally have a PWWP motif, an ADD domain, and a DNA methylase domain. The PWWP domain is named for its Pro-Trp-Trp-Pro motif, and binds to Histone-4 methylated at lysine-20 (Qiu, Sawada, Zhang, & Cheng, 2002; Stec et al., 1998; Wang et al., 2009).

All Dnmt transcript sequences (2197 sequences) with a DNA methylase domain were initially aligned with PASTA (Practical Alignment using SATe and Transitivity) (Mirarab et al., 2015) using Muscle (Edgar, 2004) as the merger. This alignment was used to build an initial consensus tree with Tamura-Nei Neighbor joining methods and 100 bootstrap replicates in Geneious (Kearse et al., 2012). I used this initial alignment and phylogeny, along with available gene annotations and BLAST analyses, to separate the sequences into groups corresponding to the Dnmt1, Trdmt1, and Dnmt3 gene subfamilies. Sequence assignment was then manually checked to assess any discrepancies with annotations. This phylogenetic approach to identifying which sub family a sequence belonged to was more accurate than relying on BLAST hits alone as there is a great deal of sequence variation within and across the Dnmt subfamily and many sequences are poorly annotated. Once split into groups corresponding to Dnmt1, Trdmt1 (Dnmt2), and Dnmt3, coding sequences were realigned using ClustalW 2.1

(Larkin et al., 2007) with a BLOSUM (Henikoff & Henikoff, 1992) cost matrix, a gap open cost of 10, and gap extension cost of 0.1, in the translation alignment tool in Geneious (Kearse et al., 2012). Due to high sequence divergence along the 3' end of the gene, sequences were trimmed to the region containing the PWWP and DNA methylase domains from ~200bp upstream of the PWWP domain to the end of the transcript. IQ-TREE2 (Minh et al., 2020) was used to build maximum likelihood phylogenetic trees, with ModelFinder and 1000 ultrafast bootstrap replicates. The best fit model of substitution was selected based on BIC. Substitution models and outgroups used for each tree are listed in Table 5.1.

To combat long branch attraction at the base of the Dnmt3a phylogeny, we utilized the -g option in IQTREE2 to constrain trees by specifying the branching pattern at the root based on known broad-scale phylogenetic relationships. The constrained tree was compared to the unconstrained tree to test for significant difference in the likelihood of the topology using maximum likelihood, RELL approximation, and Approximately Unbiased tests (Kishino, Miyata, & Hasegawa, 1990; Shimodaira, 2002) implemented in IQTREE2.

Table 5.1 – Outgroup sequences for each subfamily gene tree.

Gene Subfamily	Outgroup Species	Substitution Model	Outgroup Sequence ID
Dnmt1	*Ciona intestinalis*	TIM+F+R9	ENSCINT00000010797.4
Dnmt2	*Ciona intestinalis*	TIM3e+R10	ENSCINT00000008432.3
Dnmt3	*Petromyzon marinus*	GTR+F+R10	ENSPMAT00000008525.1
Dnmt3a	*Petromyzon marinus*	GTR+F+R6	ENSPMAT00000007628.1
Dnmt3b	*Petromyzon marinus*	GTR+F+R7	ENSPMAT00000008525.1

The constraint tree used for dntm3a is ((ENSLACT00000014792.1,ENSDART00000151921.2),(XM_020512624.1,XM_007900434.1),ENSPMAT00000007628.1). The goal of this constraint tree is to specify the Chondrichthyes clade should be a sister clade to the clade containing bony fish and tetrapods, and all of those clades combined should be sister to the sea lamprey outgroup.

Tests for selection

All tests for selection were conducted with the HyPhy (hypothesis testing using phylogenies) version 2.5.15 (Kosakovsky Pond et al., 2020; Pond & Muse, 2005) phylogenetic software. Alignments were checked in HyPhy to identify sites that were only gaps or had stop codons. Stop codons at the end of transcripts were removed; with all available up-to-date data there were no stop codons; for current analysis one sequence has missing information for some amino acids due to initial sequence quality. We evaluated patterns of selection in Hyphy, including To assess selection across Dnmt transcripts including gene wide episodic diversifying selection (BUSTED - Branch-site Unrestricted Statistical Test for Episodic Diversification) (Murrell et al., 2015), episodic diversifying section along a subset of branches (aBSREL - adaptive Branch-Site Random Effects Likelihood) (Kosakovsky Pond et al., 2011; M. D. Smith et al., 2015), episodic positive selection (sites evolving under positive selection for a portion of branches) (MEME - Mixed Effects Model of Evolution) (Murrell et al., 2012).

Identifying other genes in zebrafish with CH domains

We used a phylogenetic approach in order to investigate the evolutionary origin of a Calponin Homology (CH) domain on a subset Dnmt3b genes in Neopterygii fishes including spotted gar and all teleosts, we examined the genome of *Danio rerio* for CH domain containing genes. Interproscan (P. Jones et al., 2014) was used to identify genes containing CH domains (PF00307) in the Danio rerio genome. The protein sequences for just the CH domain region were aligned with MUSCLE (Edgar, 2004) implemented in Geneious 10.1.2 (Kearse et al., 2012) with default parameters. Substitution models were assessed in ModelFinder (Kalyaanamoorthy, Minh, Wong, von Haeseler, & Jermiin, 2017), the best model was determined based on BIC. A maximum likelihood tree was created using IQ-TREE (Nguyen, Schmidt, Von Haeseler, & Minh, 2015) with a LG+R4 model of substitution and 1000 ultrafast bootstrap replicates (Hoang, Chernomor, Von Haeseler, Minh, & Vinh, 2018).

Results

We identified 2201 protein sequences, corresponding to 919 genes containing DNA methylase domains in 222 vertebrate species plus our chordate outgroup *Ciona intestinalis*.

Dnmt1 is evolutionarily constrained to be single copy in vertebrates

Dnmt1 appears to be maintained in single copy in vertebrates despite species in other groups such as insects and fungi having increased copy number of this maintenance gene. A handful of species (12 of 211) have multiple annotations for Dnmt1 or Dnmt1 like

genes, half of those appear to be mis-annotations or pseudogenes with early stop codons. One marsupial (*Notamacropus eugenii*) and five bony fish species, including two salmonids (*Oncorhynchus kisutch*, *Salmo sar*) and three cave fish (*Sinocyclocheilus anshuiensis*, *Sinocyclocheilus grahami*, *Sinocyclocheilus rhinocerous*), appear to have two copies that are full length or near full length.

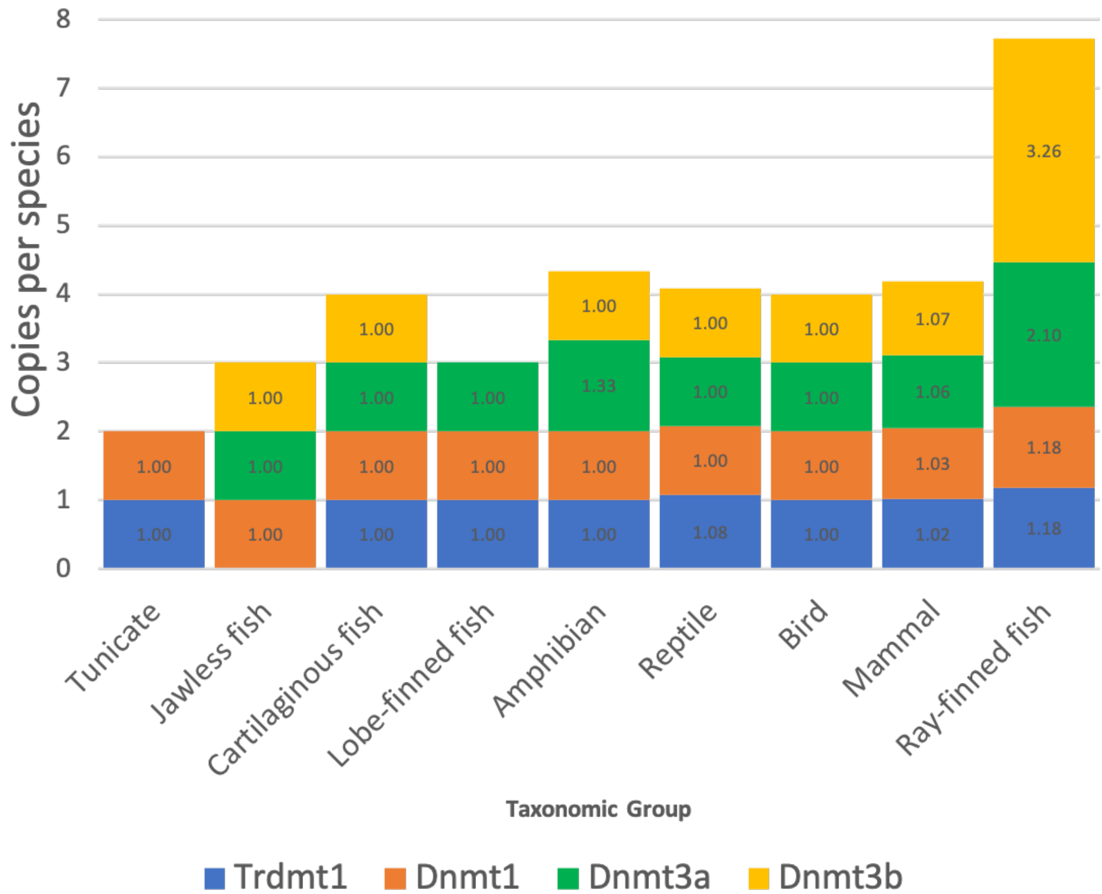

Figure 5.1 – Copies per species of genes with a DNA methylase domain (PF00145) for each gene subfamily by taxonomic group.

Figure 5. 2 - Dnmt1 Maximum likelihood tree. Branch weight is bootstrap support. Clades have been collapsed generally at the Order level.

Dnmt1 is also highly conserved at the sequence level, with 66% sequence identity across an alignment of 210 transcript sequences representing over 500 million years of evolution. In addition to the DNA methylase domain that characterizes methyltransferase

Trdmt1 like Dnmt1 appears to largely be constrained to be maintained in single copy sequences, Dnmt1 genes in vertebrates also consistently contain a DMAP binding domain, a Dnmt1-RFD (Cytosine_MeTrfase1_RFD), a zinc finger CXXC, and a two interval BAH domain. The transcript sequences of Trdmt1 share 64.6% pairwise identity across vertebrate species, and code for a protein primarily consisting of a DNA methylase domain. We did not identify a Trdmt1 sequence in *Petromyzon marinus*. Of the 216 species we found Trdmt1 sequences for 10 had two copies. Seven of these species with two Trdmt1 sequences were bony fishes (*Cyprinus carpio*, *Gasterosteus aculeatus*, *Lates calcarifer*, *Notothenia coriiceps*, *Sinocyclocheilus anshuiensis*, *Sinocyclocheilus grahami*, *Sinocyclocheilus rhinocerous*), two were mammals (*Octodon degus*, *Propithecus coquereli*), and one was a turtle (*Pelodiscus sinensis*). Four of the seven bony fish species with two copies of Trdmt1 belong to the cypriniform clade containing carp and barbels and are known to have shared whole genome duplication ~10mya (Xu et al., 2019).

Dnmt3 is highly variable within and between species

Dnmt3 appears to have undergone repeated duplication and loss in ray finned fishes. This is partially a result of the teleost whole genome duplication which doubled the copies of Dnmt3a and Dnmt3b genes, however, there appears to be additional shared and lineage

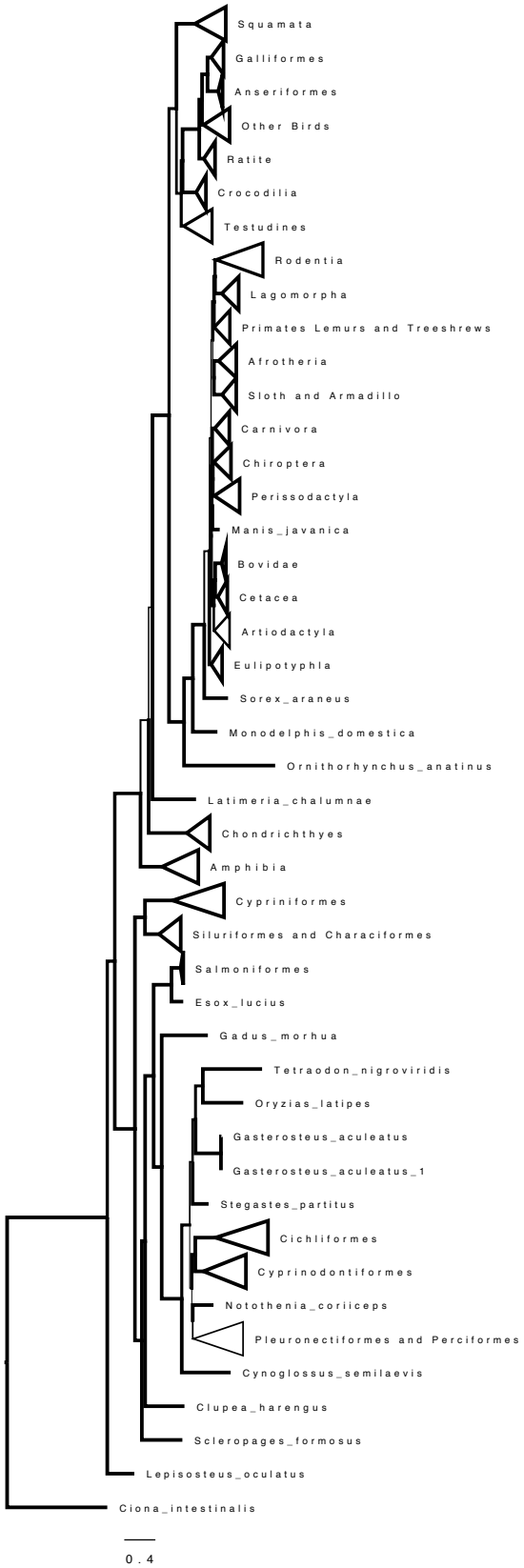

Figure 5.3 – Trdmt1 maximum likelihood tree. Branch weight is bootstrap support. Clades have been collapsed generally at the Order level.

specific gains and losses of Dnmt3 genes. Duplication events are more extensive for Dnmt3B compared to Dnmt3A, with at least one tandem duplication of Dnmt3B occurring prior to the teleost whole genome duplication and resulting in the gain of a CH domain associated with the duplication. This initial duplication is present in the genomes of spotted gar and teleost fishes but is absent in cartilaginous fish and tetrapod genomes. Subsequent duplication and retention of Dnmt3s has occurred in some fishes. These additional gene duplications occur primarily in taxa known to have undergone recent whole genome duplications, such as Salmoniforms and Cypriniforms. Though the genus *Poecilia* (Künstner et al., 2016), which is not known to have a recent whole genome duplication, has also retained recently duplicated copies of Dnmt3ab. In addition to duplication events retained copies of Dnmt3 are highly variable outside of a conserved region containing a PWWP domain and the DNA methylase domain. This variation includes differences in the protein domains coded for by the 5' end of the gene.
In addition to duplications of Dnmt3 in the genomes of fishes a few duplications have occurred in other vertebrates, including several independent expansions of Dnmt3 in Amphibians, Marsupials, Rodents, and Primates.

Duplicate copies of Dnmt3a due to the teleost whole genome duplication are generally conserved, with only 6 of 40 bony fishes having only one copy of Dnmt3a. The average number of Dnmt3a copies for bony fish is 2.1, with some species having as many as 4 copies. We did not observe any consistent loss of a Dnmt3a copy across Orders of bony fish.

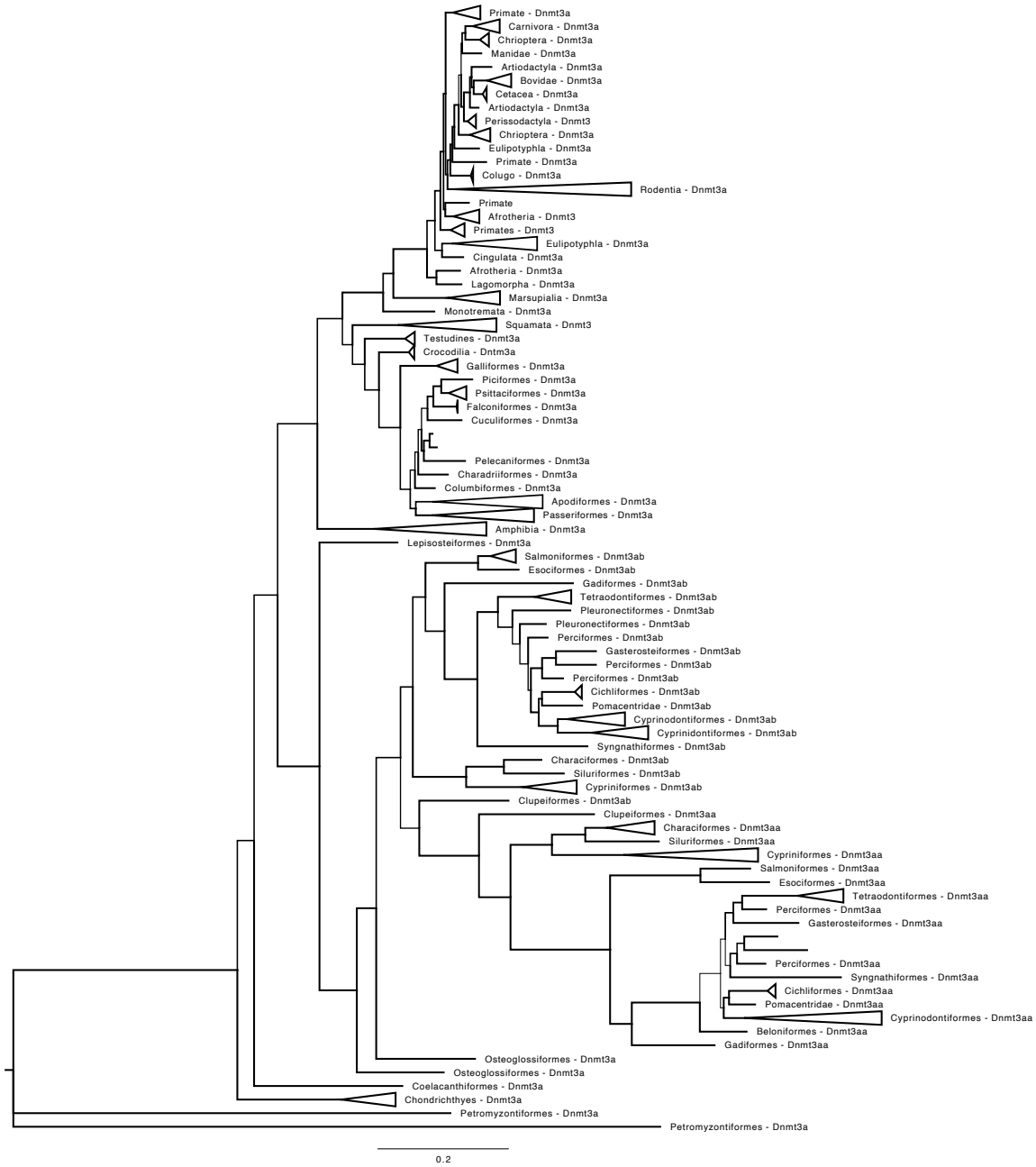

Figure 5.4 – Dnmt3a Maximum likelihood tree. Branch weight is bootstrap support. Clades have been collapsed generally at the Order level.

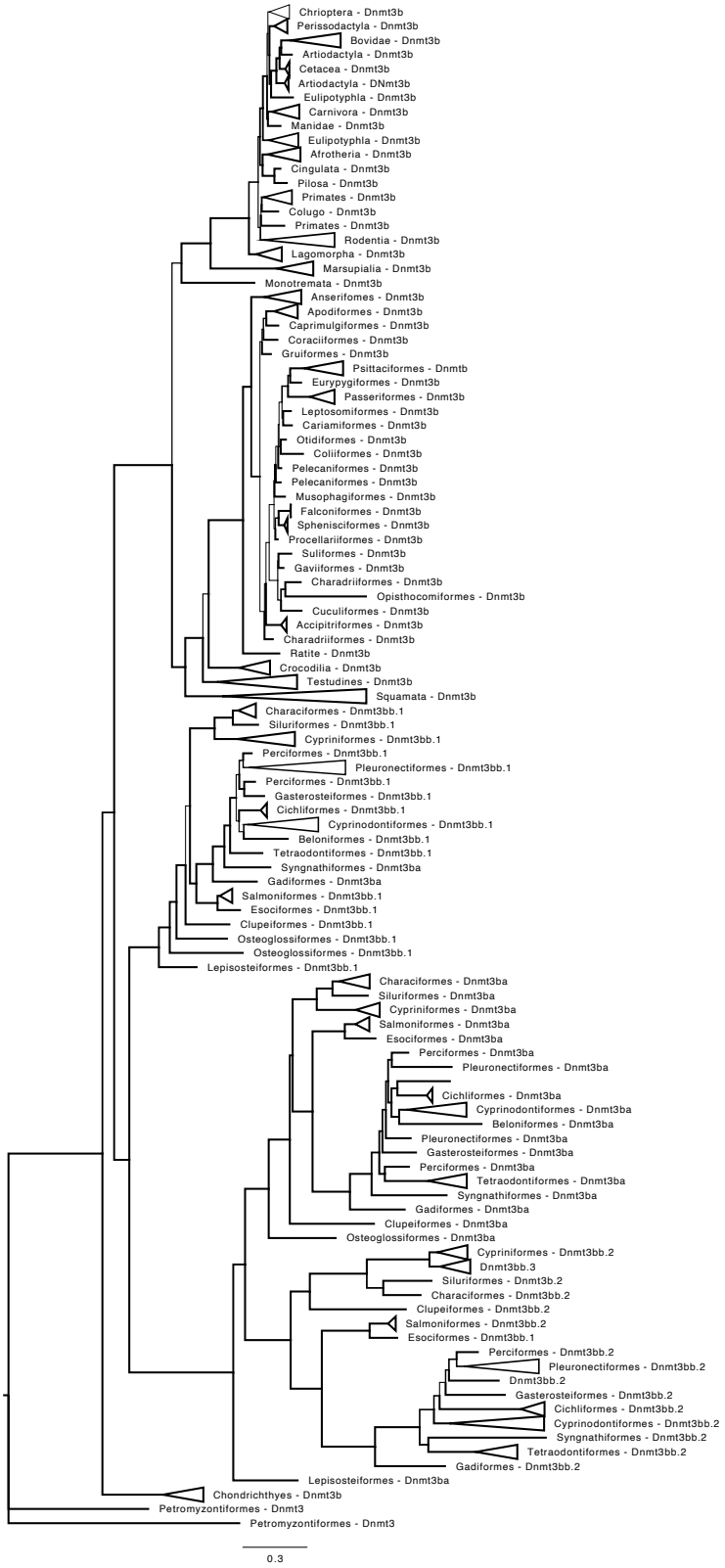

Figure 5.5 – Dnmt3b Maximum likelihood tree. Branch weight is bootstrap support. Clades have been collapsed generally at the Order level

The number of Dnmt3b copies recovered from bony fish genomes varied from 0 to 8 with an average of 3.3 +/- 1.4 std dev copies, and a mode of 3 copies. The three primary clades of Dnmt3b genes correspond to those named Dnmt3ba, Dnmt3bb.1, and Dnmt3bb.2 in the Danio rerio genome. We will continue referencing gene clades based on the naming system from *Danio rerio*.

Tests for selection

For Dnmt3A BUSTED found evidence of gene wide episodic diversifying selection (LRT p ≤0.05). Using aBSREL we identified 23 of 415 branches in the Dnmt3a phylogeny that exhibited episodic diversifying selection (LRT p ≤0.05). Of 949 sites, MEME found evidence of diversifying selection at 85 and 61 sites with a p-value threshold of 0.10 and 0.05 respectively.

For Dnmt3B BUSTED found evidence for gene wide episodic diversifying selection (LRT p ≤0.05). For Dnmt3b we identified 51 of 538 branches showing evidence if episodic diversifying selection using aBSREL (LRT p ≤0.05). Of 1277 sites, MEME found evidence of diversifying selection at 197 and 155 sites with a p-value threshold of 0.10 and 0.05 respectively.

Calponin Homology (CH) domain containing genes in *Danio rerio*

By examining the genes in the *Danio rerio* genome, that contain Calponin Homology domains, we can try to identify the origin of the CH domain in two of *D. rerio*'s Dnmts. Through phylogenetic analyses of the CH domain of these genes we have been able to

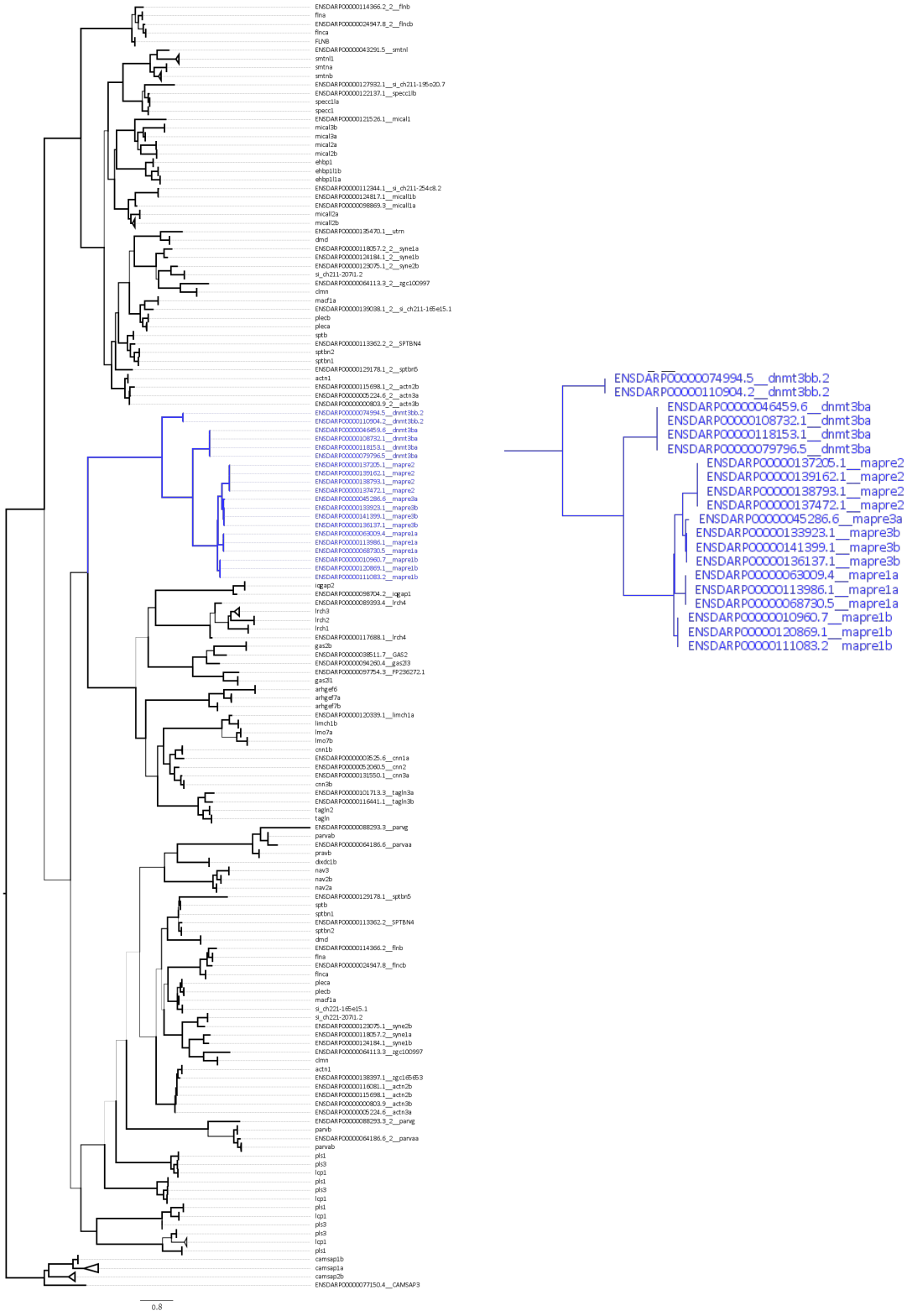

Figure 5.6 - Left: Maximum likelihood tree made with IQTREE of CH domain protein sequences for all genes containing a CH domain in the Danio rerio genome. The blue portion of this tree contains Danio rerio Dnmt genes with CH domains and Mapre genes. Right: Mapre and Dnmt CH domains. New trees of CH domain containing genes in danio genome. Branch weight indicates bootstrap support.

identify the most similar CH domain sequences in a group of Mapre genes. Among the 276 CH domain sequences in the *D. rerio* genome, we are able to consistently recover a monophyletic clade containing Dnmt3ba and Dnmt3bb.2, as well as Mapre1b, Mapre1b, Mapre2, Mapre3a, and Mapre3b (Figure 5.6). This suggests shared ancestry of the CH domain among these two gene families. In other words, a Mapre gene is the likely origin of the first CH domain in a fish Dnmt gene.

Dnmt3 Evolution Hypothesis

Our results suggest that the existing hypothesis for the duplication events that have led to the current 6 orthologs of Dnmt3 in zebrafish needs to be revised and genialized to apply to all bony fish species. We identified the gain of the Calponin homology (CH) domain – a family of actin-binding domains – on some Dnmt3b genes occurred in a common ancestor of gars and Teleosts prior to the teleost whole genome duplication. Additionally, this gain of a CH domain at the 5' end of the Dnmt3b gene seems to correspond with a tandem duplication of Dnmt3b that can be found in gar as well as teleost genomes. I propose a model starting with an ancestral Dnmt3 gene that was duplicated in the vertebrate whole genome duplication, resulting in two paralogs, Dnmt3a and Dnmt3b. Then following the divergence of lobed finned fishes (coelacanth, lung fish, and tetrapods) and ray finned fishes a tandem duplication of Dnmt3b occurred. During this duplication event the second copy of Dnmt3b (corresponding to Dnmt3bb.2 in zebrafish) gained a region that codes for a Calponin Homology domain on the upstream (3') end of the gene. Based on phylogenetic analyses, this CH domain shares as common ancestor with CH domains of Mapre genes suggesting a Mapre gene as the source of the CH

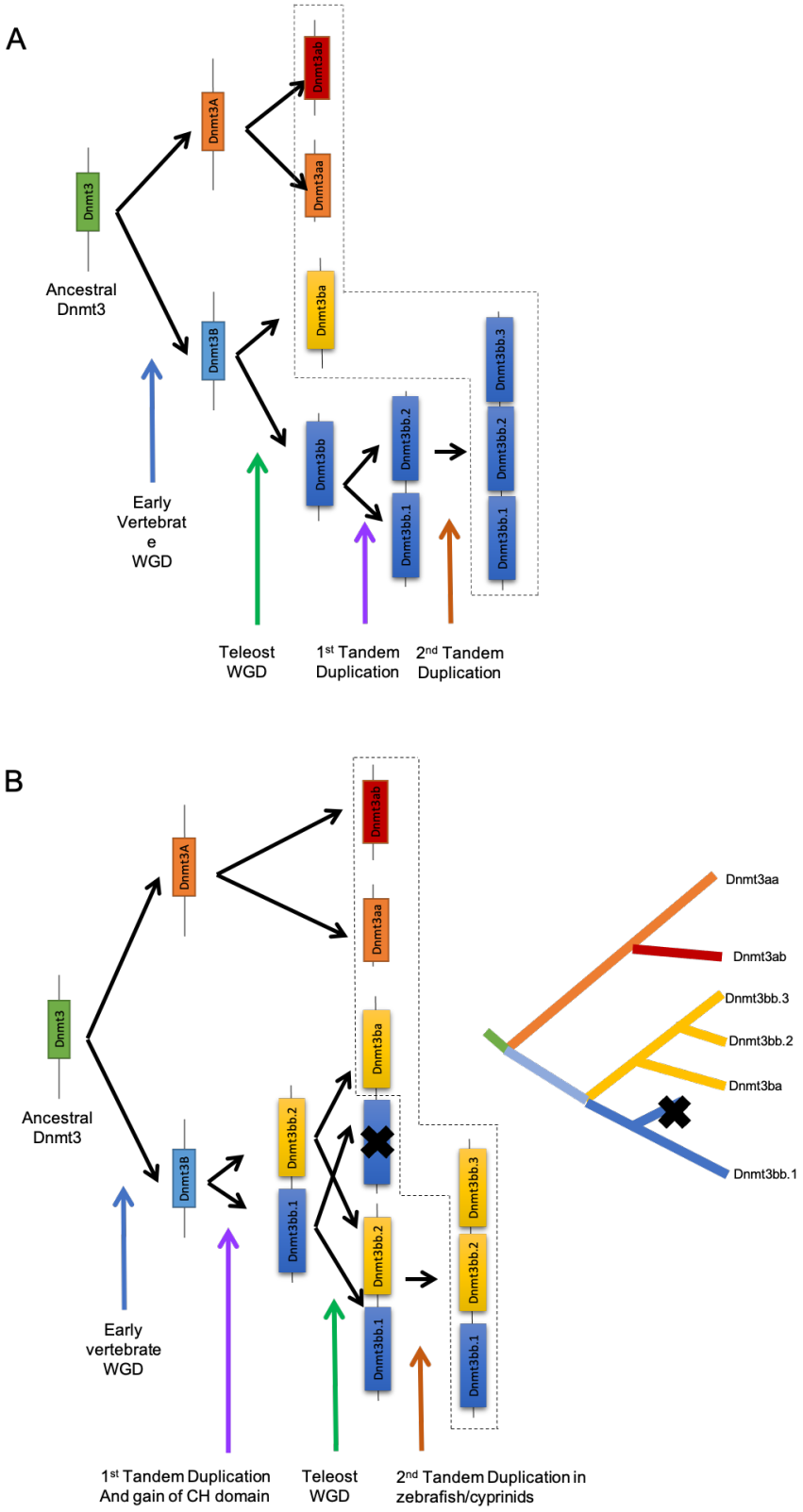

Figure 5.7 – Prior hypothesis of Dnmt3 evolutionary relationships in vertebrates (A), revised hypothesis presented here, with diagram of genetic relationships between gene copies (B). Dashed lined polygons encompass extant gene copies.

domain now found on Dnmt3bs in fishes. A possible explanation for the addition of this domain to Dnmt3b and simultaneous tandem duplication is non-homologous recombination. This gain of a protein domain suggests this copy of Dnmt3b may have gained novel function. During the teleost whole genome duplication Dnmt3a and the tandem copies of Dnmt3b were duplicated again. This resulted in two copies on Dnmt3a in teleosts (corresponding to Dnmt3aa and Dnmt3ab in zebrafish) and four copies of Dnmt3b. Soon after the teleost whole genome duplication one of these copies of Dnmt3b was lost likely as a result of a chromosomal rearrangement. This resulted in three copies of Dnmt3b that are present in teleost fishes (corresponding to Dnmt3ba, Dnmt3bb.1, and Dnmt3bb.2. In addition to these three copies of Dnmt3b some lineages of teleost fishes have had subsequent duplication and retention events of these genes including another tandem duplication in the lineage leading to zebrafish resulting in three tandem Dnmt3b genes on one chromosome, as well as whole genome duplications in salmonid and carp lineages.

Discussion

Though some other organisms such as yeast, drosophila, and C. elegans lack DNA methyltransferase genes we found Dnmt genes in all vertebrate genomes examined, Trdmt1 was also present in most genome examined. We found three monophyletic super clades corresponding to Dnmt1, Trdmt1, and Dnmt3. Dnmt1 and Trdmt1(Dnmt2) were found in single-copy in all genomes examined with a few exceptions noted in the results. Several species of vertebrates appear to have multiple copies of Dnmt1, however upon closer examination many of these extra copies appear to be mis-annotations or

pseudogenes with early stop codons. Suggesting that duplication of Dnmt1 in vertebrates results in one copy being maintained by selection and the other copy accumulating mutations until function is lost. These are likely the result of recent genome duplications as each of these taxonomic groups have experience a recent genome duplication. This makes it likely that each is a real gene with one copy per species currently undergoing the process of becoming a pseudogene. Further investigation into sequence variation and gene expression levels for each copy of Dnmt1 in these species may further elucidate the strength of selection in maintaining Dnmt1 in single copy soon after duplication. This suggests a biological importance to keeping Dnmt1 in single copy in vertebrates. Interestingly, Dnmt1 is known to have undergone copy number expansion in other organisms such as insects and fungi. Therefore, vertebrates appear to have a unique evolutionary constraint that is maintaining Dnmt1 in single copy.

The majority of species analyzed retain Trdmt1 in single copy, though several species appear to have two copies, these appear to be the result of recent duplications (Figure 5.3). Like Dnmt1, these second copies may be due to mis-assemblies or mis-annotations, however at least one duplication appears to be real as the duplication is shared within a taxonomic clade. The cypriniform subclade that includes three species of *Sinocyclocheilus* as well as *Cyprinus carpio*. *Sinocyclocheilus* and *Cyprinus* shared a genome duplication event approximately 10 million years ago (Xu et al., 2019). We found two Trdmt1 genes for each of the species in this clade.

Zebrafish have two homologs of Dnmt3a, and four homologs of Dnmt3b, partially a result of the teleost whole genome duplication and additional duplications of Dnmt3b (Campos, Valente, & Fernandes, 2012). Two copies of Dnmt3a have been maintained for ~350 million years, suggesting that these gene copies have potentially diverged and have gene function split between the gene copies (subfunctionalization) or one copy may have evolved a new function (neofunctionalization) while the other copy retained its original function. (Duarte et al., 2006; Kuo & Kissinger, 2008; Lan & Pritchard, 2016; Lynch & Force, 2000).

In order to have a significant contribution to evolution methylation must be heritable between generations (Mendizabal, Keller, Zeng, & Soojin, 2014; Verhoeven, vonHoldt, & Sork, 2016). Methylation inheritance mechanisms are known for some groups of organisms, for example imprinting in mammals (Bartolomei & Tilghman, 1997; Okano et al., 1999), but remain to be investigated for many organisms.

Though this study is limited by the accuracy and completeness of genomes available in public databases, we still found interesting patterns of evolution including gene duplications, and rearrangements. Due to variation in the quality of genome assemblies used in this project, when we failed to recover an ortholog we expected in our dataset we inferred this as missing data rather than interpreting it as a loss of a gene.
Having a better understanding of what genes are controlling these changes could provide tools for managing captive populations, such as salmon, that are being reared for aquaculture and restoration purposes.

Comment on Limitations

There was some evidence of additional copies of Dnmt1 in several species; these appear to be pseudogenes, mis-assemblies, or mis-annotations. Online databases of genome data are highly valuable resources, this project would have been impossible without them, yet they still have some extreme limitations. While some genomes for model organisms are highly polished, have been created with a variety of sequencing technologies, and validated through linkage maps, experimentally, and computationally, many others have been assembled with limited sequencing data, limited prior knowledge of genomic structure, and limited computational analyses. This leads to highly variable completeness and quality of genome sequences that are available to researchers. In addition to sequence completeness and errors, annotations of genomic databases are also majorly prone to errors (Schnoes, Brown, Dodevski, & Babbitt, 2009). Because of these limitations, I have refrained from inferring any gene losses without a consistent taxonomic signal of loss. Further, we did not depend on functional annotations to identify DNA methylase domain containing gene, but rather searched protein sequences of all genes for this domain. This analysis is still error prone by the fact it is dependent of protein domain databases to compare sequences to.

References

Adams, S., Vinkenoog, R., Spielman, M., Dickinson, H. G., & Scott, R. J. (2000). Parent-of-origin effects on seed development in *Arabidopsis thaliana* require DNA methylation. *Development, 127*(11), 2493-2502.

Bartolomei, M. S., & Tilghman, S. M. (1997). Genomic imprinting in mammals. *Annual review of genetics, 31*(1), 493-525.

Bewick, A. J., Hofmeister, B. T., Powers, R. A., Mondo, S. J., Grigoriev, I. V., James, T. Y., . . . Schmitz, R. J. (2019). Diversity of cytosine methylation across the fungal tree of life. *Nature ecology & evolution, 3*(3), 479.

Bewick, A. J., Sanchez, Z., McKinney, E. C., Moore, A. J., Moore, P. J., & Schmitz, R. J. (2019). Dnmt1 is essential for egg production and embryo viability in the large milkweed bug, *Oncopeltus fasciatus*. *Epigenetics & Chromatin, 12*(1), 6. doi:10.1186/s13072-018-0246-5

Bewick, A. J., Vogel, K. J., Moore, A. J., & Schmitz, R. J. (2017). Evolution of DNA methylation across insects. *Molecular Biology and Evolution, 34*(3), 654-665.

Bewick, A. J., Zhang, Y., Wendte, J. M., Zhang, X., & Schmitz, R. J. (2019). Evolutionary and Experimental Loss of Gene Body Methylation and Its Consequence to Gene Expression. *G3: Genes|Genomes|Genetics, 9*(8), 2441-2445. doi:10.1534/g3.119.400365

Campos, C., Valente, L. M., & Fernandes, J. M. (2012). Molecular evolution of zebrafish dnmt3 genes and thermal plasticity of their expression during embryonic development. *Gene, 500*(1), 93-100.

Capuano, F., Mülleder, M., Kok, R., Blom, H. J., & Ralser, M. (2014). Cytosine DNA methylation is found in Drosophila melanogaster but absent in Saccharomyces cerevisiae, *Schizosaccharomyces pombe*, and other yeast species. *Analytical chemistry, 86*(8), 3697-3702.

Colot, V., & Rossignol, J. L. (1999). Eukaryotic DNA methylation as an evolutionary device. *Bioessays, 21*(5), 402-411.

Duarte, J. M., Cui, L., Wall, P. K., Zhang, Q., Zhang, X., Leebens-Mack, J., . . . Altman, N. (2006). Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of *Arabidopsis*. *Molecular Biology and Evolution, 23*(2), 469-478.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research, 32*(5), 1792-1797.

Fernández, X. M., & Birney, E. (2010). Ensembl Genome Browser. In *Vogel and Motulsky's Human Genetics* (pp. 923-939): Springer.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., . . . Sangrador-Vegas, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic acids research, 44*(D1), D279-D285.

Gavery, M. R., Nichols, K. M., Goetz, G. W., Middleton, M. A., & Swanson, P. (2018). Characterization of genetic and epigenetic variation in sperm and red blood cells from adult hatchery and natural-origin steelhead, *Oncorhynchus mykiss. G3: Genes, Genomes, Genetics, 8*(11), 3723-3736.

Glasauer, S. M., & Neuhauss, S. C. (2014). Whole-genome duplication in teleost fishes and its evolutionary consequences. *Molecular genetics and genomics, 289*(6), 1045-1060.

Goll, M. G., Kirpekar, F., Maggert, K. A., Yoder, J. A., Hsieh, C.-L., Zhang, X., . . . Bestor, T. H. (2006). Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science, 311*(5759), 395-398.

Helfman, G., Collette, B. B., Facey, D. E., & Bowen, B. W. (2009). *The diversity of fishes: biology, evolution, and ecology*: John Wiley & Sons.

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences, 89*(22), 10915-10919.

Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution, 35*(2), 518-522.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., . . . Nuka, G. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics, 30*(9), 1236-1240.

Jurkowska, R. Z., Jurkowski, T. P., & Jeltsch, A. (2011). Structure and function of mammalian DNA methyltransferases. *Chembiochem, 12*(2), 206-222.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods, 14*(6), 587.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., . . . Duran, C. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics, 28*(12), 1647-1649.

Kishino, H., Miyata, T., & Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution, 31*(2), 151-160.

Klose, R. J., & Bird, A. P. (2006). Genomic DNA methylation: the mark and its mediators. *Trends in biochemical sciences, 31*(2), 89-97.

Kosakovsky Pond, S. L., Murrell, B., Fourment, M., Frost, S. D., Delport, W., & Scheffler, K. (2011). A random effects branch-site model for detecting episodic diversifying selection. *Molecular Biology and Evolution, 28*(11), 3033-3043.

Kosakovsky Pond, S. L., Poon, A. F., Velazquez, R., Weaver, S., Hepler, N. L., Murrell, B., . . . Nekrutenko, A. (2020). HyPhy 2.5—A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Molecular Biology and Evolution, 37*(1), 295-299.

Kucharski, R., Maleszka, J., Foret, S., & Maleszka, R. (2008). Nutritional control of reproductive status in honeybees via DNA methylation. *Science, 319*(5871), 1827-1830.

Kucharski, R., Maleszka, J., & Maleszka, R. (2016). *A possible role of DNA methylation in functional divergence of a fast evolving duplicate gene encoding odorant binding protein 11 in the honeybee.* Paper presented at the Proc. R. Soc. B.

Künstner, A., Hoffmann, M., Fraser, B., A. , Kottler, V., A. , Sharma, E., Weigel, D., & Dreyer, C. (2016). The Genome of the Trinidadian Guppy, *Poecilia reticulata*, and Variation in the Guanapo Population. *PLoS ONE, 11*(12), e0169087-e0169087. doi:10.1371/journal.pone.0169087

Kuo, C.-H., & Kissinger, J. C. (2008). Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites *Plasmodium* and *Theileria*. *BMC evolutionary biology, 8*(1), 108.

Lan, X., & Pritchard, J. K. (2016). Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science, 352*(6288), 1009-1013.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., . . . Lopez, R. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics, 23*(21), 2947-2948.

Le Luyer, J., Laporte, M., Beacham, T. D., Kaukinen, K. H., Withler, R. E., Leong, J. S., . . . Bernatchez, L. (2017). Parallel epigenetic modifications induced by hatchery rearing in a Pacific salmon. *Proceedings of the National Academy of Sciences, 114*(49), 12964-12969.

Li, E., Bestor, T. H., & Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell, 69*(6), 915-926.

Liu, J., Hu, H., Panserat, S., & Marandel, L. (2020). Evolutionary history of DNA methylation related genes in chordates: new insights from multiple whole genome duplications. *Scientific Reports, 10*(1), 1-14.

Lyko, F., Ramsahoye, B. H., Kashevsky, H., Tudor, M., Mastrangelo, M.-A., Orr-Weaver, T. L., & Jaenisch, R. (1999). Mammalian (cytosine-5) methyltransferases cause genomic DNA methylation and lethality in *Drosophila. Nature genetics, 23*(3), 363-366.

Lynch, M., & Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics, 154*(1), 459-473.

Mendizabal, I., Keller, T., Zeng, J., & Soojin, V. Y. (2014). Epigenetics and evolution. *Integrative and comparative biology, 54*(1), 31-42.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution, 37*(5), 1530-1534.

Mirarab, S., Nguyen, N., Guo, S., Wang, L.-S., Kim, J., & Warnow, T. (2015). PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *Journal of Computational Biology, 22*(5), 377-386.

Murrell, B., Weaver, S., Smith, M. D., Wertheim, J. O., Murrell, S., Aylward, A., . . . Smith, D. M. (2015). Gene-wide identification of episodic selection. *Molecular Biology and Evolution, 32*(5), 1365-1371.

Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., & Pond, S. L. K. (2012). Detecting individual sites subject to episodic diversifying selection. *PLoS Genet, 8*(7), e1002764.

Muyle, A., & Gaut, B. S. (2018). Loss of Gene Body Methylation in Eutrema salsugineum Is Associated with Reduced Gene Expression. *Molecular Biology and Evolution, 36*(1), 155-158. doi:10.1093/molbev/msy204

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution, 32*(1), 268-274.

Okano, M., Bell, D. W., Haber, D. A., & Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell, 99*(3), 247-257.

Parrott, B. B., & Bertucci, E. M. (2019). Epigenetic aging clocks in ecology and evolution. *Trends in Ecology & Evolution, 34*(9), 767-770.

Pond, S. L. K., & Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. In *Statistical methods in molecular evolution* (pp. 125-181): Springer.

Ponger, L. c., & Li, W.-H. (2005). Evolutionary diversification of DNA methyltransferases in eukaryotic genomes. *Molecular Biology and Evolution, 22*(4), 1119-1128.

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research, 35*(suppl 1), D61-D65.

Qiu, C., Sawada, K., Zhang, X., & Cheng, X. (2002). The PWWP domain of mammalian DNA methyltransferase Dnmt3b defines a new family of DNA-binding folds. *Nature structural biology, 9*(3), 217-224.

Ryazanova, A. Y., Abrosimova, L., Oretskaya, T., & Kubareva, E. (2012). Diverse domains of (cytosine-5)-DNA methyltransferases: structural and functional characterization. In *Methylation-From DNA, RNA and Histones to Diseases and Treatment*: InTech.

Schmitz, R. J., Lewis, Z. A., & Goll, M. G. (2019). DNA Methylation: Shared and Divergent Features across Eukaryotes. *Trends in Genetics, 35*(11), 818-827. doi:10.1016/j.tig.2019.07.007

Schnoes, A. M., Brown, S. D., Dodevski, I., & Babbitt, P. C. (2009). Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS computational biology, 5*(12).

Shimoda, N., Yamakoshi, K., Miyake, A., & Takeda, H. (2005). Identification of a gene required for de novo DNA methylation of the zebrafish no tail gene. *Developmental dynamics, 233*(4), 1509-1516.

Shimodaira, H. (2002). An Approximately Unbiased Test of Phylogenetic Tree Selection. *Systematic biology, 51*(3), 492-508. doi:10.1080/10635150290069913

Simpson, V. J., Johnson, T. E., & Hammen, R. F. (1986). *Caenorhabditis elegans* DNA does not contain 5-methylcytosine at any time during development or aging. *Nucleic acids research, 14*(16), 6711-6719.

Smith, M. D., Wertheim, J. O., Weaver, S., Murrell, B., Scheffler, K., & Kosakovsky Pond, S. L. (2015). Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Molecular Biology and Evolution, 32*(5), 1342-1353.

Smith, T. A., Martin, M. D., Nguyen, M., & Mendelson, T. C. (2016). Epigenetic divergence as a potential first step in darter speciation. *Molecular ecology*, n/a-n/a. doi:10.1111/mec.13561

Stamp, M. A., & Hadfield, J. D. (2020). The relative importance of plasticity versus genetic differentiation in explaining between population differences; a meta-analysis. *Ecology Letters, 23*(10), 1432-1441.

Stec, I., Wright, T. J., van Ommen, G.-J. B., de Boer, P. A., van Haeringen, A., Moorman, A. F., . . . den Dunnen, J. T. (1998). WHSC1, a 90 kb SET domain-containing gene, expressed in early development and homologous to a Drosophila dysmorphy gene maps in the Wolf-Hirschhorn syndrome critical region and is fused to IgH in t (1; 14) multiple myeloma. *Human molecular genetics, 7*(7), 1071-1082.

Verhoeven, K. J., vonHoldt, B. M., & Sork, V. L. (2016). Epigenetics in ecology and evolution: what we know and what we need to know. *Molecular ecology*.

Wang, Y., Reddy, B., Thompson, J., Wang, H., Noma, K.-i., Yates III, J. R., & Jia, S. (2009). Regulation of Set9-mediated H4K20 methylation by a PWWP domain protein. *Molecular cell, 33*(4), 428-437.

Xu, P., Xu, J., Liu, G., Chen, L., Zhou, Z., Peng, W., . . . Sun, Y. (2019). The allotetraploid origin and asymmetrical genome evolution of the common carp *Cyprinus carpio*. *Nature Communications, 10*(1), 1-11.

CHAPTER 6

CONCLUSIONS

In this dissertation, I presented several resources and studies that I produced to examine mechanisms of divergence, from biogeographic isolation to gene family evolution. In Chapter 2, I assembled the mitochondrial genome of *Notropis lutipinnis*, and conducted a phylogenetic analysis that showed a high level of paraphyly among genera within this group of Leucisidae fishes, highlighting the need to closely examine and update these taxonomic relationships with genomic markers.

In my third chapter I examined genome-wide divergence for population pairs of four species thought to have been isolated by a geologic stream capture event. Here we found that two of the species appear to have diverged at least several hundred thousand years ago, the other two appear to have a much more recent divergence between the two river systems. This suggests that the history of dispersal and isolation between these river systems is likely more complicated than previously thought. The addition of more samples per population as well as more species could provide sufficient information to more clearly examine the genetic divergence of species thought to have been isolated by this steam capture. The original intent of this study was to include eight species and at least twelve individuals per population and compare the genomic data to large datasets of coalescent simulations. Unfortunately, due to the coronavirus pandemic and related lockdown we were unable to finish making sequencing libraries for this project. We hope

that when laboratories are more accessible again, we will be able to gather additional sequencing data and further assess the questions presented here.

In Chapter 4, we hoped to identify genes that may be under selection and driving divergence between the northern and southern extent of the species range of *Notochthamalus scabrosus*. One of our original goals for this project was to identify metabolic loci that had been used previously in studies using allozymes. Though we identified some partial transcripts that appear to correspond to some of these genes, the majority we were unable to identify in our dataset due to high divergence between *N. scabrosus* and other species with well-annotated genomes. Despite these challenges we were able to identify several hundred loci with high population structure and for some of these we were able to calculate Ka/Ks and evaluate selection at the molecular level between populations. Though many of these were not metabolic loci, we did find additional loci of interest to investigate further, such as a cuticular protein.

In my final analytical chapter, I examined the evolution of the DNA methyltransferase gene family across vertebrates. Although we identified a general pattern of conservation in sequence and copy number, we found an interesting pattern of duplication and gene expansion in Dntm3s across Actinopterygian fishes and described a new hypothesis for the evolutionary relationships of these novel copies and the gain of a Calponin Homology domain.