UNIVERSITY OF CALIFORNIA, MERCED

Adaptive molecular evolution within a highly diverse group of fishes, rockfishes (*Sebastes*)

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

Doctor of Philosophy

in

Quantitative and Systems Biology

Joseph Heras

Committee in charge:

Professor Michael N. Dawson, Chair
Professor David Ardell
Professor Monica Medina
Professor Miriam Barlow
Professor Andres Aguilar

2014

UNIVERSITY OF CALIFORNIA, MERCED
Graduate Division

The Dissertation of Joseph Heras is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Faculty Advisor: _____

Miriam Barlow

Committee Members: _____

Chair: Michael N. Dawson

_____

David Ardell

_____

Monica Medina

_____

Andres Aguilar

_____

Date

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Dr. Andres Aguilar, who gave me an opportunity to test my limits as a scientist in training, who was helpful and allowed me to develop a project that I thought was worth crafting into a dissertation. My dissertation committee: Drs. Michael N. Dawson, Monica Medina, and David Ardell, who were helpful throughout the Ph.D. dissertation process. Dr. Miriam Barlow, who provided guidance and supported my decisions as I completed my degree.

I would like to thank my fellow graduate student colleagues, Jason Baumsteiger and Andres Martinez, who provided comic relief and have been there for me throughout the majority of my Ph.D. program. To my sister Eva Marie, my brother Patrick, my brother-in-law Jamshid Danesh, and my nephew Brandon as they brought me cheer during the toughest of times. My good friend, Mark D. Dela who spent countless hours in my presence as we conversed about the content of my dissertation over a cup of coffee at Dripp. I would like to thank Omar Ureta, Victor Landa, and Benjamin C. Morrow who assisted with sampling off the coast of Santa Cruz, California. Marissa Chavez who was there to listen to my dilemmas and brought the best out of me. Aaron and Elizabeth Razo, Chris and Tiffany Galvan, Mike and Danielle Tatum for spiritual support. I would also like to thank Jennifer Liberto who contributed to developing bioinformatic scripts that were used to analyze some of my datasets for this dissertation. The Dawson Laboratory group: Sarah Abboud, Liza Gomez-Delgalio, Lauren Schiebelhut, Holly Swift, and Julia Vo who were readily available to provide feedback on any of my presentations, whether for a conference or preparation of my qualifying exam. I would like to thank Corey Cain for constructively criticizing my dissertation proposal. My Tío Raphael and Tía Paty Peña, and Tío Arturo and Tía Elisa Heras for providing accommodations while I was finishing the last of my analyses at California State University, Los Angeles. Carrie King who provided insight on my dissertation and assisted me with executing some of my ideas. I would like to thank Edwin H. Gibb for providing computational assistance on some of the analyses that were essential for my dissertation to be completed.

Lastly, I would like to dedicate this dissertation to my parents, Jose A. Heras and Eva Heras, who were supportive and patient with me, and encouraged me to pursue my dreams.

VITA

EDUCATION

2005	Bachelor of Science in Biology, University of California, Riverside
2007	Master of Science in Biology, California State University, Los Angeles
2010	Advanced to Candidacy, Doctor of Philosophy in Quantitative and Systems Biology, University of California, Merced

PUBLICATIONS

Heras J, Koop BF, and Aguilar A. 2011. A transcriptomic scan for positively selected genes in two closely related marine fishes: *Sebastes caurinus* and *S. rastrelliger*. Marine Genomics. 4(2): 93-98.

Baumsteiger J, Swift HF, Lehman JM, Heras J, and Gomez-Daglio L. 2010. Getting to the root of phylogenetics. Frontiers of Biogeography. 2(3): 68-69.

De Ley P, De Ley IT, Morris K, Abebe E, Mundo M, Heras J, Waumann D, Olivares AR, Burr J, Baldwin JG, and Kelly WK. 2005. An integrated approach to fast and informative morphological vouchering of nematodes for application in molecular barcoding. Philosophical Transactions of the Royal Society B. 360(1462): 1945-1958.

Adaptive molecular evolution within a highly diverse group of fishes, rockfishes (*Sebastes*)

by

Joseph Heras

University of California, Merced, 2014

Advisor: Prof. Miriam Barlow

ABSTRACT OF THE DISSERTATION

The molecular processes that govern speciation are not completely understood, especially within marine ecosystems. Insight on the topic of speciation can be provided by focusing on adaptive radiations, because of the rapid amount of lineages that have arisen within a short natural historical time frame. Rockfishes (genus: *Sebastes*) are known as a system for adaptive radiations within marine systems and have been considered an ancient species flock. This dissertation concentrates on marine rockfishes from the subgenus (*Pteropodus*) and closely related congeners to make inferences about speciation and adaptation within marine systems. Brain, kidney, testes, ovaries, and spleen tissues were selected from multiple rockfish species to extract total RNA in order to sequence the transcriptome with the use of Sanger and next generation sequencing techniques. These tissues types are intended to give us a broad spectrum of the rockfish transcriptome and were utilized to identify genes subject to positive selection. Pair-wise and multiple species comparisons were conducted to identify orthologous sequence pairs between/among species and estimated nonsynonymous (K_a) and synonymous (K_s) substitutions.

In the first study, a comparative analysis of Expressed Sequence Tags (ESTs) from *Sebastes caurinus* and *S. rastrelliger* was conducted to identify candidate genes under positive selection within the subgenus *Pteropodus.* Genes with elevated K_a/K_s values belonged to the following functional categories: immune function, metabolism, longevity, and reproductive behavior, indicating that adaptive divergence at immunological, physiological, and reproductive loci may be important in the diversification of this group of fishes. In the second study, ESTs from *S. goodei* and *S. saxicola*, a pair of related congeners to *Pteropodus*, were utilized for identifying positive selection with the use of testes and ovary tissue types. Gonadal tissues were selected from these two species to elucidate patterns of adaptation and speciation because of their recognition to contribute to reproductive barriers between species. Sequence divergence was estimated within the untranslated regions (UTRs) between these two species and was compared with the rate of divergence within coding regions of these genes to gain a clear

depiction of neutral substitutional mutation rates. Orthologous gene pairs between the two species were identified and tested for positive selection. In addition, a candidate gene approach from the zona pellucida (ZP) family was selected to determine whether these genes are under positive selection. In the third study of this dissertation, a multi-transcriptomic comparison was conducted, in which three *Pteropodus* species were analyzed with the addition of *S. rastrelliger* and *S. caurinus* datasets to identify a pattern of positive selection within this subgenus. Brain tissue was used to gain the most diverse set of transcripts from the rockfish genome. Genes under positive selection belonged to a variety of gene functions that included sensory perception, growth, and metabolism. In addition we identified 10 sequences under positive selection that were part of the phosphatidylinositol signaling system pathway, although genes under positive selection had a broad range of gene functions and genes identified under positive selection across all five *Pteropodus* species were not identified in the first study.

These collections of studies are intended to further advance the field of evolutionary biology by providing support of which functional genes are important for adaptation and speciation. Currently, there is no rockfish genome available, in which by sequencing the transcriptome this provides a foundation for future genomic projects to gain a better understanding about the genus. This dissertation was developed to determine whether traditional gene categories, which are known to be under positive selection such as immune function, reproduction, and apoptosis, also play an important role in marine speciation. With transcriptomic data from multiple species within *Sebastes*, we obtained a suite of candidate genes under positive selection, which can be used to assess whether these genes are responsible for the radiation and how adaptation and speciation occurred across the entire genus of *Sebastes*.

Table of Contents

List of Figures

List of Tables

Chapter 1

Elucidating patterns of adaptive evolution within marine rockfishes (*Sebastes*) via the use of transcriptomic scans

1.1 What makes *Sebastes* an ideal system for understanding molecular evolutionary processes?

Adaptive evolution is the end result of where advantageous alleles become fixed (positive selection) within a population (Yang et al. 2000). From this process, do loci under positive selection share a function that can be detected throughout multiple lineages in order to identify the driving force of adaptation and speciation? The use of adaptive radiations, a single ancestral lineage that stems into multiple lineages that are diverse and occupy a variety of different ecological habitats (Gavrilets and Vose, 2005), can be used to assess whether the same loci are under positive selection within multiple lineages (Sugawara et al. 2002). Identifying loci under selection from multiple lineages can provide support that these genomic regions contribute to adaptation. Adaptive radiations provide an opportunity to assess whether certain loci are responsible for adaptation because these multiple lineages are replicates of the speciation process, which can be used to identify genetic patterns that contributed to the radiation (Burford and Bernardi, 2008).

Within teleost fishes, the iconic example for adaptive radiations has been African rift lake cichlids that occupy a diverse set of habitats and have extensive morphological variation (Moore 1903; Greenwood, 1981). Just within Lake Victoria, there are 200 species that radiated within the last few thousand years (Johnson et al. 1996). Other examples of adaptive radiations include freshwater sculpin from Lake Baikal in Russia (Berg, 1965) and pupfishes from Lake Chichancanab of Mexico (Humphries and Miller, 1981). Most studies of adaptive radiations within teleost fishes occur in lacustrine systems and limited examples of adaptive radiations have been described within marine systems. Ice fishes within the suborder Notothenioidei and marine rockfishes from the genus *Sebastes*, the system of interest for this dissertation, are considered examples of ancient species flocks, groups of species that arose due to a past adaptive radiation (Johns and Avise, 1998). The environmental differences observed between lacustrine and marine systems are the absolute barriers found within lacustrine systems that would allow for allopatry to occur (Puebla, 2009). In contrast, dispersal of marine larvae provides an opportunity for marine organisms to disperse over long distances, which would reduce the chances of allopatric speciation to occur.

Rockfishes (*Sebastes*) inhabit temperate waters of the Pacific and Atlantic Oceans, but the center of diversity is located in the North Pacific with 96 resident species (Love et al. 2002) that occupy habitats from the coast of Baja California, Mexico to Alaska, United States. There are 22 subgenera recognized within *Sebastes* (Kendall, 2000), which includes *Pteropodus* the subgenus of interest within this study. Marine rockfishes have matrotrophic viviparity, the process of where the eggs are fertilized

internally and the mother provides nutrition to the developing embryo and the offspring are expelled as larvae (Love et al. 2002), which is not a common life history trait in a majority of extant bony fishes. Currently, three species within *Sebastes* (Boccaccio rockfishes, *S. paucispinus*; canary, *S. pinniger*; and yelloweye, *Sebastes ruberrimus*) have been designated as overfished and within the United States are federally registered as threatened species by the National Oceanic and Atmospheric Administration (Jacobsen, 2013). The common trait among all rockfish species is a suborbital stay, a bone that extends from the third infraorbital bone to the preoperculum bone.

Similar to cichlids, marine rockfishes exhibit an array of coloration, which possibly evolved due to differences in depth and visibility. Seehausen et al. (1998) showed within cichlids, females select mates based on color when color differences are visible. Within rockfishes, there is very limited information about their reproductive behavior, but there is a diverse array of coloration within marine rockfishes. The diversity in morphological traits within this genus and limited information about reproductive behavior provides an opportunity to determine whether certain loci are associated with mate choice, a potential contributing factor to reproductive isolation and eventually speciation.

There are extensive differences in life history traits among species within *Sebastes*, such as longevity and the number of eggs produced (Love et al. 2002). This information can provide insight on the genetic mechanisms of how this may have contributed to the adaptive radiation within this group. There are many speculations about which environmental factors impact lifespans within living organisms. Studies done on freshwater killifish from Trinidad (*Rivulus hartii*) demonstrated that exposure to predation prompts killifish generations to be smaller in size and have faster growth rates. In comparison between high-predation exposures to no-predation exposure, killifish from high-predation sites were more effective at producing more eggs than killifish from no-predation sites in accordance with a high ration of food supply (Walsh and Reznick, 2008). These pressures of obtaining resources and levels of exposure to predators may be a driving factor for the differences that we see in marine rockfishes both in size, fecundity, and lifespans.

Sebastes have high fecundity, the lowest observed egg count is 18,000 produced by calico (*S. dalli*) rockfishes and the highest egg count of 2,700,000 is observed in yelloweye rockfishes (*S. ruberrimus*) (Love et al. 2002). Not all eggs are fertilized and genetic studies have demonstrated multiple paternity occurs within multiple rockfish species (Sogard et al. 2008; Hyde et al. 2008). In addition, there is an extensive difference in lifespans within rockfishes; the shortest-lived rockfish species is calico rockfish (*S. dalli*) at 12 years and the longest-lived rockfish is rougheye (*S. aluetianus*), which have a maximum lifespan of 205 years (Love et al. 2002). Since we see a wide range of differences in longevity and egg production throughout this genus, this can provide insight into adaptation and speciation for this group. This speciose group with a diverse set of morphological traits and unique life history traits makes marine rockfishes ideal for studying adaptive radiations, because we can assess whether certain traits are responsible for the rapid diversification of lineages.

1.2 Candidate species within *Sebastes* to understand adaptive evolution

This dissertation focuses on identifying patterns of positive selection within coding gene sequences between/among species of *Sebastes* to make inferences about adaptation and speciation. *S. goodei* (chilipepper) and *S. saxicola* (stripetail) and the subgenus *Pteropodus* were of interest for identifying protein-coding regions for the identification of positive selection as an indicator for adaptation and speciation. These studies demonstrate that candidate genes are subjected to positive selection within closely related species and to assess whether these genes are found under positive selection across the subgenus *Pteropodus* (Eigenmann and Beeson, 1893) with the use of different tissue types and also to identify genes under positive selection within distantly related congeners (*S. goodei* and *S. saxicola*) with the use of gonadal tissue. The divergence between *S. goodei* (subgenus *Sebastodes*; Gill, 1861) and *S. saxicola* is estimated to be over 6 mya (Hyde and Vetter, 2007). The use of reproductive tissue from these two congeners was to make inferences about reproductive isolation mechanism that contributed to the radiation within this genus (Chapter 3).

This subgenus (*Pteropodus*) was selected for this dissertation (Chapters 2 and 4) because this group is nested within a diverse genus and all species within the subgenus form a monophyletic clade in current published phylogenies (Li et al. 2006; Hyde and Vetter, 2007). All species within this subgenus reside in the Northeast Pacific and contain common characteristics such as mottled color patterns and strong head spines (Li et al. 2006). In addition, the members of this subgenus generally occupy nearshore and shallow shelf habitats and are found sympatric among each other along most of the coast of California (Li et al. 2006). Based on geological and genetic data, *Pteropodus* has been suggested to evolve around 3 million years ago whereas the genus *Sebastes* arose close to 8 million years ago (Hyde and Vetter, 2007). Current phylogenies constructed for species within *Sebastes* (Hyde and Vetter, 2007; Li et al. 2006) show that *Pteropodus* contains the following species: *S. atrovirens* (kelp), *S. auriculatus* (brown), *S. caurinus* (copper), *S. chrysomelas* (black and yellow), *S. carnatus* (gopher), *S. dalli* (calico), *S. maliger* (quillback), *S. nebulosus* (China), and *S. rastrelliger* (grass). Hyde and Vetter (2007) have suggested that future analyses of species flocks within rockfish species should focus on monophyletic and geographically constrained clades (subgenera) such as *Acutomentum*, *Pteropodus*, *Sebastocles*, or *Sebastomus*, because of the invariable speciation rates found within this genus. In congruence with Hyde and Vetter's (2007) suggestions, this dissertation provides insight on the molecular evolution of the subgenus *Pteropodus*, as well a broad scale perspective on loci subject to positive selection between the distantly related congeners *S. goodei* and *S. saxicola*.

1.3 Speciation and adaptation within marine systems

Marine systems offer a unique source for understanding evolutionary processes because of the opportunity for extensive planktonic larvae stages to disperse widespread throughout the ocean. In addition, there are limited geographical barriers that would inhibit gene flow within marine ecosystems. Therefore, speciation events within these ecosystems would not be solely a consequence of allopatric speciation (Ingram, 2011), where sympatry or parapatry would likely be the processes of speciation. In order for these processes of speciation to occur within marine systems, reproductive isolation

mechanisms would have to be ongoing, such as differences in spawning time, mate recognition, environmental tolerance, and gamete compatibility. These processes have been suggested to contribute to marine speciation (Palumbi, 1994), which is likely driven by divergent or disruptive selection (Puebla, 2009). In this dissertation, the focus is to identify patterns of adaptive evolution within marine rockfishes and make inferences of how these patterns are associated with ecological speciation and mechanisms for reproductive isolation.

1.4 Our understanding of the evolution of marine rockfishes (*Sebastes*)

The taxonomic assignment of the species within the genus *Sebastes* has puzzled ichthyologist for over a century (Kendall 2000). This taxonomic confusion stems from the characterization of morphological traits in rockfishes and how to sort out each species based on these traits, which were described from species that resided from the Pacific Ocean. In 1861, Dr. Theodore Gill had noticed that Bocaccio (*S. paucispinis*) appeared to be morphologically different from the remainder of the Pacific rockfish species. This was the support for a separate genus, *Sebastodes*, and this lead to a great deal of confusion of which species would be included in this genus and eventually 15 genera were generated. However, Dr. Kiyomatsu Matsubara was able to make sense of the morphological differences and lumped all marine rockfishes back into one genus *Sebastes*. He made these inferences based on the anatomy, morphology, and taxonomy of rockfishes and similar species (Love et al 2002).

The phylogenetic reconstruction of the genus *Sebastes* provided insight on the resolution of the taxonomic classification of rockfishes, and there are strongly supported and phylogenetic studies using genetic markers have provided support one genus (Johns and Avise, 1998; Rocha-Olivares et al. 1999; and Hyde and Vetter, 2007). Hyde and Vetter (2007) provided a comprehensive phylogenetic review of the genes that included 99 species within *Sebastes*, with the use of seven mitochondrial genes and two nuclear genes. This included the fossil record and the estimated time of origin of the genus as well as lineages within the genus. This information has provided a robust understanding of the evolutionary relationships of the species within *Sebastes*, however what are the genetic mechanisms that prompted this adaptive radiation within this genus? Only a few studies have addressed how functional genes have contributed to this radiation within *Sebastes* (Sivasundar and Palumbi, 2010; and Johansson and Banks, 2011).

Many studies use candidate genes to study adaptive evolution (Gerrard and Meyer, 2007), such as genes associated with mate choice (a prezygotic reproductive isolation mechanism), where searching for molecular constituents associated with mate choice would likely be associated with vision (Sugawara et al. 2005). Sivasundar and Palumbi (2010) demonstrated within rhodopsin genes adaptive changes that were associated with depth in marine rockfishes. In comparison to African rift lake cichlids, the difference visual sensitivity is dependent on differential gene expression within cone opsin genes, where within any given species there is a subset of genes expressed (Carleton and Kocher, 2001; Parry et al. 2005). Sivasundar and Palumbi (2010) used 32 species from *Sebastes* and sequenced the rhodopsin gene, where they identified 14 amino acid replacements out of the 240 amino acid fragment studied that included six out of the nine species within *Pteropodus*. All *Pteropodus* (*S. auriculatus*, *S. chrysomelas*, *S.*

carnatus, *S. caurinus*, *S. maliger*, and *S. nebulosus*) species investigated in the Sivasundar and Palumbi (2010) study showed the same amino acid replacement in comparison to the other congeners found at lower depths. This suggests that the ancestral lineage to *Pteropodus* was most likely a near shore species that radiated into the multiple lineages that we see today.

Another candidate gene studied in regards to adaptive evolution of rockfishes is the olfactory receptor Type 2 gene (Vlr-like Ora 2; Johansson and Banks, 2011). Vlr-like Ora genes have been suggested to be chemoreceptors for pheromones within fishes. This gene was a likely candidate for positive selection because of the potential to be part of the mechanisms for reproductive isolation. The speculation that these genes would be under positive selection is due to the reproductive behavior of marine rockfishes, in which odor can play an important aspect of sexual selection within marine rockfishes. Although not much is known about reproductive behavior, observations have described male rockfish courtship as swimming in front of the female before copulation is permitted (Helvey, 1982). With the males swimming in front of the female, the male may release pheromones in order to initiate copulation. Johansson and Banks (2011) demonstrated within five rockfish species, which included two *Pteropodus* species (*S. caurinus* and *S. maliger*) that nine sites were under positive selection. However, these sites do not pertain to the ligand binding site, which indicates that these changes maybe a result of relaxed selection. Both studies from Sivasundar and Palumbi, (2010) and Johansson and Banks (2011) are examples of using single candidate genes to investigate patterns of adaptive evolution. However, there are more than likely thousands of genes within the rockfish genome that show signatures of positive selection, and the interactions of these genes would provide a stronger statement about how natural selection and adaptive evolution has impacted this group of fishes.

1.5 Adaptive evolution: our understanding of a natural phenomena

Adaptive evolution is the end result of certain alleles that proliferate within a population because these alleles are advantageous and increase the chances of an individual to reproduce and survive (Swanson, 2003). Charles Darwin made keen observations of how specific traits were advantageous, as seen in his description of the diversity of finches on the Galapagos Islands, which inhabited a variety of ecological niches (Lack, 1983). However, the molecular constituents of these traits were not revealed to Darwin due to the restraints of technology available during his lifetime. Today, the advancement in nucleotide sequencing technology has provided the opportunity for genetic information to be sequenced from a multitude of species. In addition, this sequence information offers the opportunity to quantify the amount of genetic differences at the population, species, and higher taxonomic levels. These genetic differences stem from mutations that have occurred and accumulated over time which contribute to the processes of evolution in various forms. These mutations that are favorable become fixed within the population are considered to be under positive selection (Hughes, 1999).

Many studies have identified loci throughout the genomes under positive selection from a variety of taxa (Andres et al. 2013; Elmer et al. 2010; Schartl et al. 2013). Through studies of molecular evolution there are certain genes or loci that are

consistently found under strong purifying (negative) selection (i.e. histones), where selective pressures prevent mutations to occur due to the reduction in fitness of the organism. On the converse, occasionally mutations arise that are advantageous which allows for the process of positive selection to occur. These mutations allow the organism to thrive within a given environment where the end result would be adaptive evolution. Traditional studies which have concentrated on identifying these genomic regions under positive selection focus on coding regions in which the identification of nonsynonymous and synonymous differences (Miyata et al. 1980; Nei and Gojobori, 1986; Goldman and Yang, 1994, McDonald and Kreitman, 1991) are identified. Loci under positive selection have been suggested to be the driving mechanism to provide adaptive evolution. The most traditional estimate of adaptive evolution is through the use of d_n/d_s or K_a/K_s (both are in reference to the same analysis) rate ratio as a demonstration of positive selection. An estimate of d_n/d_s ratio greater than one is an indication of positive selection, and a d_n/d_s ratio less than one is indicative of purifying selection. This d_n/d_s test is considered to be conservative test because nonsynonymous substitutions would be considered deleterious, and thus the rates of nonsynonymous substitutions are much lower than synonymous substitutions (Eyre-Walker, 2006).

Contemporary studies on adaptive evolution have focused on identifying elevated nonsynonymous substitutions as compared to synonymous substitutions within coding genes. Initial models to identify adaptive evolution within coding sequences were developed by Muse and Gaut, 1994; and Nei and Gojobori, 1986. These methods were developed to take coding sequences and estimate the amount of synonymous substitutions and non-synonymous substitutions. The ratio of nonsynonymous over synonymous substitutions (ω) is a form to assess the type of selection operating on the coding gene. These methods are considered approximate methods in which the amounts of sites of synonymous and nonsynonymous are counted, then the synonymous and nonsynonymous differences are counted, and lastly corrections are made for multiple substitutions (Yang and Bielawski, 2000). A more recent method has been developed, which is based on maximum likelihood (ML), where parameters (i.e. sequence divergence, d_n/d_s ratio, and transition/transversion rate ratio) are estimated from the data based on ML (Yang and Bielawski, 2000). The estimation of positive selection within coding regions has been extensive in which the estimation of selection can be identified at specific sites via a Naïve Empirical Bayes or through Bayes Empirical Bayes, in which the former ignores sampling errors within the parameter estimates while the latter accounts for these errors by using a prior (Yang et al. 2005).

Novel methods have been developed that identify positive selection include parallel amino acid changes, increase the rate of insertion/deletion substitutions, accelerated gene loss, and enhance gene expression noise (Zhang, 2010). In addition, evolutionary studies have called attention to the functional differences in gene expression (Oleksiak et al. 2002), which may play a strong role in adaptation but are recently being uncovered with the use of next generation sequencing. Lastly, changes within the development of an organism appear to be of great importance especially if these are novel genes identified within the newly divergent species (Chen et al. 2010).

1.6 Caveats of using d_n/d_s tests

Novel methodologies are essential to unveil most of the properties of adaptive evolution (Zhang, 2010). The use of d_n/d_s tests has been greatly criticized because the estimation of positive selection as an indicator of adaptive evolution can be misleading, as there can be reasons other than positive selection for rapid changes within the coding region (Hughes, 2007). One speculation for the increase of nonsynonymous substitutions over synonymous substitutions would be because of relaxed selection within a given region of the protein (Hughes, 2007). If the region of the protein is not a binding site or does not have structural function, then purifying nor positive selection would be operating on these sites. Hughes (2007) also suggested that adaptive evolution can be a result of a single nucleotide polymorphism, or the result of gene silencing/deletion, and the change in expression of genes. Hoekstra and Coyne (2007) have reviewed studies where mutations lead responsible for morphological changes were associated with to loss of gene function, deletion or frameshifts, and even loss of phosphorylation sites. These changes will not be measured within analyses of d_n/d_s. Although d_n/d_s may not detect all patterns of adaptive evolution, within genomic scans there are repetitive gene categories found under positive selection. As more genomic information is uncovered and the advancement of bioinformatic tools, we can use multiple approaches to make assessments about positive selection, and thus make stronger inferences about how these genes are impacted by environmental pressures.

1.7 Comparative genomics within teleost fishes

With the use of next generation sequencing, there are currently eleven fish species genomes sequenced and readily available on Ensembl (http://www.ensembl.org/). This knowledge of a variety of fish genomes has provided an essential resource for economic use and also evolutionary studies (Sarropoulou and Fernandes, 2010). Model fish genomes also serve for identifying orthologous and paralogous genes that are essential for making inferences about gene function within non-model organisms. If paralogous genes are assumed to be orthologous across different species, this can lead to misleading conclusions about function and mutation rates (Dolinski and Botstein, 2007). Studies of teleost fishes have indicated that an entire genome duplication, known as the Fish Specific Genome Duplication (FSGD), occurred ~350 million years ago (Meyer and Van de Peer, 2005). With a duplicated genome, this provides the opportunity for novel gene functions to occur via neofunctionlization or subfunctionalization (Wolfe, 2001). There are currently about ~25,000 different species of teleost fishes presently known (Taylor et al. 2003), many more than any other vertebrate group (Meyer and Van de Peer, 2005). The diversity of this speciose group has been suggested to occur because of the FSGD, which has allowed for adaptive evolution to occur within novel genes in response to a variety of novel habitats (Meyer and Van de Peer, 2005). There is contradicting evidence that shows that FSGD is not solely the source for adaptive evolution (Santini et al. 2009). Santini et al. (2009) showed that diversification of teleost fishes due to the FSGD was probably 10% and that the roughly 88% of the diversification of teleost fishes within recent lineages were not likely due to the FSGD. As more genomic information becomes available can help resolve this paradox of how the diversification of fishes evolved.

1.8 Caveats of different sequencing platforms and tissue types

There is a distinct difference in the amount of genomic information obtained from different sequencing technologies; where 96,000 bp can be obtained from a Sanger sequencing run whereas over 20 Gb of sequence information can be processed via next generation technologies (www.illumina.com). Pop and Salzberg (2008) reviewed the limitations of new sequencing technology, which are associated with the assembly and annotation process. The reduction in sequencing costs and the unprecedented amount of sequencing information obtained from next generation sequencing technology (i.e. 454 Life Sciences and Solexa/Illumina) enables scientists to sequence genomes from non-model organisms. However the short read sequences provide a bioinformatic dilemma where these short reads are aligned there is the possibility of aligning homologous sequence due to repetitive elements identified. The assembly of closely related sequences is considered "transcript shadowing" (Trapnell et al. 2010). The amounts of assembly errors are different from what is identified in Sanger sequencing technology. Some of the caveats of next generation sequencing is the annotation process in which genes will be accurate if found in other species. However, with sequencing errors there may be the issue of inframe stop codons that would be difficult to distinguish from pseudogenes. To resolve this issue with today's technological resources would be to resequence multiple individuals from the same species, especially non-model organisms and use of different alignment algorithms, which can provide support for the proper identification of reading frames and assembly of loci. In addition, the need for the advancement of characterizing novel coding and protein sequences in order to keep up with the unprecedented amount of genomic information is essential. This would provide insight on the function of novel genes and improve the annotation process that would provide a robust understanding of adaptation and/or speciation at the genomic level.

Another caveat is the information provided via transcriptome sequencing of specific tissues. Tissue transcriptomes only provide insight about that specific tissue which is a fraction of the genome of an organism (~%5 of the genome; www.genome.gov). Only through sequencing the entire genome can provide an overview of all the genes that are subjected to positive selection. In addition, sequencing all tissue types and different developmental stages and compare these genes between species can lead to identifying patterns in development or tissues that have undergone adaptive evolution. This would lead to whether a gene under positive selection is expressed throughout multiple tissue types or specific tissue types. Also, whether certain genes under positive selection are expressed during certain developmental stages as well. As sequencing technology prices are reduces and readily available, searching through multiple tissue transcriptomes provides a broader outlook on which genes are under positive selection.

1.9 Future implications

This dissertation provides a first glance at transcriptomes of multiple species and a series of genes under natural selection. This study is a foundation for future studies on how these genes are correlated with life history or environmental pressures. From these transcriptomic analyses, only a fraction of the rockfish genome has been revealed. However, the series of candidate genes found under positive selection within this

dissertation can be used to develop population/species level analyses to validate patterns of adaptation. One example would be to use candidate genes that are associated with longevity and sequence these genes from multiple species within *Sebastes* that have various lifespans in order to identify whether parallel amino replacements match the maximum age of each species. Sivasundar and Palumbi (2010) conducted a study on opsin genes to determine whether amino acid replacements in these genes were based on depth. Another pathway for continuing this research would be to assess the candidate genes under positive selection from incipient species. This will determine whether some of the genes contribute to the adaption process. In addition, candidate genes under positive selection can be sampled from multiple populations from a species. Afterwards, the assessment of F_{st} outliers can give indication of whether these loci are under local natural selection at the population level. From this information, an assessment can be done to determine whether F_{st} outliers correlate with environmental patterns (i.e. difference in temperature or depth). Another approach and difficult to be conducted on live rockfishes, would be conducting genetic crosses in order to identify genes that are associated with specific adaptive traits such as pelvic reduction in threespine sticklebacks (*Gasterosteus aculeatus*; Shapiro et al. 2004). Lastly, the characterization of gamete proteins in order to understand how certain male sperm proteins interact with female egg proteins and determine if specific regions on these male proteins increase/decrease the chances of fertilization. These would be the next steps of identifying adaptation and speciation to further advance this field of comparative genomics.

1.10 Concluding remarks

The use of transcriptomics, a subset of genes expressed within the genome, is a powerful resource for rapidly identifying coding genes from the genome as opposed to sequencing the genome with no prior knowledge. In addition, the advancement in knowledge of the constituents of a genome can provide insight on loci that are associated with risk factors for genetic disorders, variation in immune function, specialization (i.e. dietary tolerances), and reproduction (Ryder, 2005). These factors may provide insightful information not only for evolutionary biology but for conservation management as well. This dissertation is intended to advance our understanding of a commercially important group of marine fishes and reveal how natural selection is operating at the transcriptomic level within this speciose group. With recent advancements in genetic sequencing technology and computational capabilities, we can further advance the field of adaptive evolution by seeking out the genomic regions which have allowed specific organisms to adapt to a particular niche.

This dissertation was developed to be a "spy glass" approach in which analyzing distinct species for genetic differences that contributed to the speciation process. However, the dilemma with this approach is that the changes, which have lead to adaptation, occurred during the split of the two lineages back in evolutionary time and many other contributing factors have been ongoing such as genetic drift (Via, 2009). The opposite side of the spectrum would be to investigate molecular evolution with a "magnifying glass" approach. This approach is geared toward observing changes as two lineages are starting to diverge. This can be demonstrated with the use of population genetic studies of partially isolated ecotypes of marine rockfishes. The loci identified

under selection in this dissertation can be used for future studies in understanding the protein function and also to test whether these loci at the population level are found under selection as well. Overall, this dissertation has provided depth and information about the genomic components of marine rockfishes, and a novel quantitative approach to estimate positive selection on a transcriptomic level in this speciose group of fishes.

Chapter 2

A transcriptomic scan for positively selected genes in two closely related marine fishes: *Sebastes caurinus* and *S. rastrelliger*

Joseph Heras[1], Ben F Koop[2], and Andres Aguilar[1,3*]

1- School of Natural Sciences and Graduate Group in Quantitative and Systems Biology, University of California Merced, 5200 N. Lake Rd., Merced, CA USA 95344

2- Department of Biology – Centre for Biomedical Research, University of Victoria, Victoria, B.C. Canada V8W 3N5

3-Sierra Nevada Research Institute, University of California Merced, 5200 N. Lake Rd., Merced, CA USA 95344

* - Author for correspondence

e-mail: aaguilar2@ucmerced.edu

Tel: 1-209-228-4057

Fax: 1-209-228-4060

Key Words: Expressed sequence tags; comparative genomics; positive Darwinian selection; adaptive radiation; rockfish

Abstract

Comparative genomic analyses can provide valuable insight into functional evolutionary divergence among closely related species. Here we employ a comparative evolutionary analysis of expressed sequence tags (ESTs) from two closely related species of marine fishes (genus *Sebastes* - rockfish). *Sebastes* is a highly diverse group of marine fishes that inhabit a wide array of marine habitats and study of this group can provide insights into speciation in the marine environment. ESTs were developed for *S. caurinus* (23,668 from brain, kidney, and spleen tissues) and *S. rastrelliger* (11,207 from brain and pituitary tissues). Following assembly we were able to identify, with high confidence, 257 orthologous sequence pairs between the two species through a reciprocal best hit blast search. Analysis of functional divergence between orthologs revealed that 19.46% had K_a/K_s values greater than 0.5 and 8.17% had K_a/K_s values greater than one, identifying a large pool of candidate genes to further study adaptive divergence in the group. Genes with elevated K_a/K_s values belonged to the following functional categories: immune function, metabolism, longevity, and reproductive behavior, indicating that adaptive divergence at immunological, physiological, and reproductive loci may be important in the diversification of this group of fishes. This study provides the ground work to better understand the molecular evolution of genes involved in a radiation of marine fishes.

2.1 Introduction

Comparative genomic studies have provided a framework for understanding genetic and genomic differences for a wide array of organisms (Clark et al. 2007; Bustamante et al. 2005; Schranz et al. 2007). Such studies often identify orthologous gene pairs between two or more organisms in order to gain a better perspective of the functional differences due to adaptive divergence (Voolstra et al. 2009). The identification of positive Darwinian selection between orthologs can be used to localize targets of adaptive evolution (Bustamante et al. 2005; Clark et al. 2007; Warren et al. 2008; and Elmer et al. 2010).

In respect to selection, novel traits permit species to adapt to their respective environments. Studies on natural selection at the molecular level have identified functional differences in gene expression (Oleksiak et al. 2002) and/or mutational differences found within DNA coding regions, which may lead to novel protein functions (Yang et al. 2000). These novel functions, if advantageous and fixed within population(s), serve as agents for natural selection (Hughes, 1999). One method to assess selection at the molecular level would be to identify selected genes that encode for advantageous polymorphisms. Selection can be measured through the assessment of nonsynonymous (K_a) and synonymous substitution (K_s) rates within coding genes (Miyata et al. 1980; Nei and Gojobori, 1986; Goldman and Yang, 1994). Studies that have focused on identifying selection via K_a/K_s have mostly found genes under positive Darwinian selection that pertain to reproduction, immune response, sensory perception, and apoptosis (Ellegren, 2008).

This study aims to identify adaptive evolutionary patterns in rockfishes (genus *Sebastes*), a group of marine fishes that are considered a species flock (Johns and Avise, 1998) and will be useful in understanding the genetic basis of adaptation and speciation. Rockfishes have internal fertilization, as opposed to the majority of teleost fishes that have external fertilization, and this feature may have contributed to speciation within this group (Love et al. 2002; Hyde and Vetter, 2007). Rockfishes exemplify the term "ancient species flock" due to the radiation that occurred within this genus approximately 5 million years ago (Johns and Avise, 1998). Rockfishes inhabit temperate waters of the Pacific and Atlantic Oceans, but the focal point of diversity is located in the North Pacific, with 96 residential species (Love et al. 2002) and approximately 105 species found world-wide (Hyde and Vetter, 2007). What sets rockfishes apart from other teleost radiations (e.g. African lake cichlids and cottids from Lake Baikal) are the different selective pressures found in marine systems in comparison to lacustrine systems (Puebla, 2009). Marine systems differ greatly due to the presence of planktonic larvae, which can increase gene flow over longer distances (Puebla, 2009). Also, rockfishes can serve as a model for studying evolution in marine habitats because they provide replicates of divergence within a given environment or time-frame (Burford and Bernardi, 2008) and there are a limited number of examples of rapid radiations found within marine environments (e.g. Antarctic notothenioids - Eastman and McCune 2000; McCartney et al. 2003).

The two species in this study, *S. caurinus* (Copper rockfish) and *S. rastrelliger* (Grass rockfish) belong to the subgenus *Pteropodus* and the two species are estimated to have diverged less than 3 million years ago (Hyde and Vetter, 2007). The distribution of

S. caurinus ranges from the northern Gulf of Alaska (Kachemak Bay) to Islas San Benito (central Baja California, Love et al. 2002), whereas *S. rastrelliger* ranges from Yaquina Bay (Oregon) to Bahia Playa Maria (central Baja California) (Love et al. 2002; Li et al. 2006). *S. caurinus* can be found in barely subtidal waters to 183 m (600 ft), while *S. rastrelliger* can often be found in intertidal/subtidal waters to 46m (150 ft, Love et al. 2002). Lastly, *S. caurinus* is predominantly absent from nearshore environments throughout most of Southern and Baja California, whereas *S. rastrelliger* is prevalent in these regions.

Our objectives in this study were to first identify orthologous gene pairs from transcriptomic datasets (Expressed Sequence Tags - ESTs) from *S. caurinus* and *S. rastrelliger* and second to identify genes under positive Darwinian selection. These genes under positive Darwinian selection may provide a foundation for identifying selection that contributed to the adaptive radiation found within this group of marine fishes.

2.2 Methods

2.2.1 EST Generation and Qualitative Analysis

Tissue samples from *S. caurinus* were obtained from a fisherman from Sidney, British Columbia, and *S. rastrelliger* tissue samples were provided by Dr. N Sherwood (Biology, University of Victoria). Brain, kidney, and spleen tissue and brain and pituitary tissue from *S. caurinus* and *S. rastrelliger* were immediately flash frozen and stored at -80°C. The difference in tissue utilized for this study was based on the best source of intact mRNA for each species and tissue availability. For RNA extraction, total RNA (Trizol reagent; Invitrogen, Carlsbad, CA, USA) and poly(A)+ RNA (FastTrack MAG kit; Invitrogen, Carlsbad, CA, USA) was extracted from the flash frozen tissue. Mixed tissue libraries were normalized by the duplex-specific nuclease normalization method (Evrogen, Moscow, Russia) and directionally cloned into the AL-17.3 (Evrogen, Moscow, Russia) vector.

Plasmid DNAs were extracted and BigDyeTM Terminator (ABI, Foster City, CA, USA) cycle sequenced on ABI 3730 sequencers using conventional procedures and the following primers: 5'-T18-3', M13 forward (5′-GTAAAACGACGGC CAGT-3′), and M13 reverse 5′-CAGGAAACAGCTATGAC-3′) or M13 reverse (-24) (5′-AACAGCTATGACCATG- 3′) or (Koop et al. 2008). Base calling and trimming of vector, poly-A tails, and low quality regions were addressed as described by Rise et al. (2004). Briefly, base-calling from chromatogram traces was performed using Phred (Ewing and Green, 1998; Ewing et al., 1998). Vector, poly-A tails, and low quality regions were trimmed from EST sequences; sequences that had less than 100 good quality bases after trimming were discarded. Initial assembly of ESTs into contigs used Phrap (http://bozeman.mbt.washington.edu), under stringent clustering parameters (minimum score = 100; repeat stringency = 0.99). A second stage assembly used the consensus sequences (with quality scores) from the first stage and parameters of 96% repeat frequency and 300 minscore to build initial contigs and consensus sequences (Koop et al. 2008). The EST resources have been submitted to GenBank with the following accession numbers: *S. caurinus* GenBank GI 213089055 – 213112722 and *S. rastrelliger* GenBank GI 156615282 – 156604076.

2.2.2 EST analyses: Masking, Clustering, Assembling, and Reciprocal Best Hit BLASTs

Each EST dataset was processed through REPEATMASKER OPEN-3.2.8 (Smit et al. 2010) in order to mask repeats found within each database by using the slow sensitive mode and source species of "Teleost Fishes". After masking, each dataset was processed through CAP3 (Huang and Madan, 1999) for further contig assembly. The settings used for CAP3 assembly included an overlap length cutoff of greater than 30 bp, percent identity cutoff of 75, and a max overhang percentage of 20. After CAP3 assembly, a Reciprocal Best Hit BLAST search was conducted on contigs from the two databases (*S. caurinus* and *S. rastrelliger*) by using TBLASTX (NCBI blast version, 2.2.17) from the BLAST package (Altschul et al. 1997). In order to gain greater confidence of detecting orthologs, PERL scripts were developed to obtain ortholog pairs with a percent coverage of 70% and an e-value of $1.0E^{-5}$ from the BLAST results of the two databases.

2.2.3 Annotations

All putative ortholog pairs were annotated in BLAST2GO (Conesa et al. 2005) to the SwissProt database by using BLASTX (NCBI blast version, 2.2.17), an E-value of 1.0 E^{-5}, 20 Blast Hits and a High Scoring Pair length cutoff of 33. The annotation parameters contained an E-value Hit Filter of $1.0E^{-6}$, annotation cutoff of 55, and a gene ontology (GO) weight of 5.

2.2.4 Identification of the Open Reading Frame, Alignments, and Estimating K_a/K_s of Putative Orthologs

In order to identify the open reading frame from the putative orthologs, all sequences were BLAST against the SWISS-PROT database (Bairoch and Apweiler, 1997) plus five protein datasets from fugu, medaka, green spotted pufferfish, stickleback, and zebrafish in the Ensembl database (Hubbard et al. 2005) (Ensemble 58) by using BLASTX (NCBI blast version, 2.2.17) with an E-value of $1.0E^{-10}$. The BLASTX file and putative ortholog file were processed through ORF-PREDICTOR and TARGETIDENTIFIER (Min et al. 2005a; Min et al. 2005b) to identify the open reading frame and determine whether the ESTs were composed of the full-length coding sequences in comparison to the annotations. Based on the translated output provided by ORF-PREDICTOR, ortholog pairs were aligned and prepared with PAL2NAL v12.2 (Suyama et al 2006), which include PERL scripts that incorporated CLUSTAL W 2.0.10 (Larkin et al. 2007) for alignment. Once the putative ortholog pairs were aligned, according to the open reading frame, an estimation of nonsynonymous (K_a) and synonymous (K_s) substitution ratios were calculated between ortholog pairs in KAKS_CALCULATOR V1.2 (Zhang et al. 2007) by using the YN model (Yang and Nielsen, 2000). All candidate orthologs with a K_s value greater than 0.1 were removed due to the possibility of including paralogs (Bustamante et al. 2005). To determine if there was any enrichment of GO categories in the positively selected group versus the non-selected group of orthologs, Fisher's Exact Test was performed which includes multiple testing corrections (two tailed) using the GOSSIP package (Blüthgen et al. 2005) and a False Discovery Rate was used as the corrected p-value for each GO term.

2.3 Results and Discussion

2.3.1 EST Generation

Prior to identifying ortholog pairs, the following ESTs were generated: 23,668 from *S. caurinus* (brain, kidney, and spleen) and 11,207 from *S. rastrelliger* (brain and pituitary) tissue, and were deposited into GenBank (GE796994-GE820661 and EW975926-EW987132). For *S. caurinus* and *S. rastrelliger*, the average read length (Phred 20) was 842.2 and 866.3 bp respectively. The average final trimmed length was 747.0 and 740.6 bp.

2.3.2 Assembly and Annotation – Entire Dataset

In REPEATMASKER Open-3.2.8 (Smit et al. 2010) there were 22 small RNAs, 0 satellites, 1,737 simple repeats, and 3,551 low complexity repeats in the *S. caurinus* EST dataset and there were 8 small RNAs, 0 satellites, 1,062 simple repeats, and 1,762 low complexity repeats found for *S. rastrelliger* dataset. These repeats were masked prior to assembly in CAP3 (Huang and Madan, 1999). From the CAP3 assembly, there were 5,686 contigs and 3,183 singletons for *S. caurinus* and 3,461 contigs and 1,161 singletons for *S. rastrelliger*. There were 8,869 and 4,622 transcripts (contigs and singletons) from the *S. caurinus* and *S. rastrelliger* respectively and each transcriptome was annotated by comparison to the SWISS-PROT database.

The annotations (GO categories) were similar between the two transcriptome sets, given that the *S. caurinus* ESTs involved an additional tissue and contained more sequences. These two transcriptome datasets were assorted according to three major divisions (Figure 2.1). There were 22 and 21 GO categories that belong to the first division (biological process) for *S. caurinus* and *S. rastrelliger*, respectively. For *S. caurinus* and *S. rastrelliger*, the top two biological process categories were metabolic (16% and 12% of sequences) and cellular processes (21% and 16% of sequences), respectively. Within the second major division (molecular function), both species contained 12 categories and the two major categories were binding (51% and 48% of sequences) and catalytic activity (26% and 23% of sequences), respectively. Lastly, the third division (cellular component), there were 6 categories for *S. caurinus* and *S. rastrelliger*, with the two most abundant categories being cell (45% and 39% of sequences) and organelle (30% and 31% of sequences), respectively. The GO categories represented in all three domains in this study were similar in comparison to Elmer et al. (2010) comparative study of two crater lake cichlid species transcriptomes.

2.3.3 Open Reading Frame Identification and Putative Ortholog Annotation

We identified 310 putative ortholog pairs identified with PERL scripts. This low number of orthologous pairs was most likely due to our rigorous RBHB and alignment length criteria. Within these 310 pairs, TARGETIDENTIFIER (Min, 2010) found 162 putative full-length sequences, 7 3'-sequenced partial sequences, 25 partial sequences (5'-sequenced partial), 15 short full-length sequences, and 1 ambiguous sequence. By using TARGETIDENTIFIER (Min, 2010), this allowed us to determine whether we had putative full or partial length sequences for our K_a/K_s analysis.

From the 310 putative ortholog pairs, there were 193 pairs (62%) annotated in BLAST2GO (Conesa et al. 2005) by using to the SWISS-PROT database. The remaining 117 (38%) orthologs were not annotated for one of the following reasons: absence of BLAST hit(s), contained BLAST hit(s) but did not have any GO annotations associated with the hit(s), or did not meet the annotation criteria according to BLAST2GO.

2.3.4 K_a/K_s (Selection)

Out of the 310 candidate orthologs, there were 257 that had a K_s value less than 0.1, which is an indication of strong purifying selection and an alignment length greater than 150 nucleotides. Additionally, there were 29 putative pairs that contained a K_a/K_s between 0.5 and 1, which is an indication of weak purifying selection, and 21 pairs with a $K_a/K_s > 1$, which is an indication of positive selection (Table 1 and Figure 2.2). We did not find any evidence for enrichment of any of the GO terms with the Fisher's Exact Test (FDR corrected p-values > 0.05) based on genes identified under positive selection).

2.3.5 Genes Under Positive Selection

The most notable functions of the positively selected genes that we found in this study were immune response, metabolism, longevity, and reproduction (Table 2.1). Genes involved in immune response include inducible protein gig2, which has been known to provide an antiviral response in teleost fish (*Carassius auratus*, Jiang et al. 2009), and Collectin-12, which displays several functions that are associated with host innate immune systems (Holmskov et al. 1994). These genes may be under positive selection primarily due to the co-evolutionary arms race between hosts and pathogens (Ellegren, 2008). The finding of immune related loci in our pool of selected loci is not surprising given the multitude of other studies that have shown concordant results (Heger and Ponting, 2007; Kosiol et al. 2008).

Genes involved in metabolic functions such as beta-2-glycoprotein 1 / apolipoprotein h, glutathione s-transferase a4, and lipocalin precursor were found under positive Darwinian selection. Selective forces shaping metabolic pathways may be important in this group of fishes, given the different environmental regimes an individual may experience over its life time (Seibel and Drazen, 2007). Likewise, the radiation with the sub-genus *Pteropodus* has involved a diverse array of near-shore habitats, which may drive evolution at metabolically important genes.

Longevity related genes (elongation factor 2 and eukaryotic translation initiation factor 3 subunit f) were also found to be under positive selection. Both of these genes are known to be integral components of mRNA translation, and are associated with longevity in *Caenorhabditis elegans* (Tavernarakis, 2007; Chen et al. 2007). This is particularly interesting because members of *Sebastes* have a broad range of life spans. The shortest lived species, *S. dalli*, is estimated to live up to 12 years and the longest lived species, *S. aleutianus*, which can live up to 205 years (Love et al. 2002). Between *S. caurinus* and *S. rastrelliger* there are differences in lifespan, with maximum ages estimated to be 50 and 23 years respectively for the two species (Love et al. 2002). Other genes involved in longevity may have an interesting evolutionary history in this group of fishes.

Lastly, we found one gene under positive selection that mediates sexual reproduction - male-specific protein (MSP). MSP is linked to sexual behavior and aggressiveness in tilapia (*Sarotherodon galilaeus*) males and contains putative pheromone-like properties (Machnes et al. 2008). Machnes et al. (2008) conducted a BLAST search for homologous sequences with MSP tilapia, in which the *S. rastrelliger* EST sequence was retrieved along with homologs from other teleosts fishes, a bird (chicken), and mammals. There have been a limited amount of studies on rockfish reproductive behavior and there has been only speculation in regards to the production of pheromones by male rockfish in their urinary bladder, which stimulate mating behaviors (Shinomiya and Ezaki, 1991; Love et al. 2002). The function of MSP in rockfish is currently unknown, but the fact that this gene was found to be under positive selection and its known function in cichlids suggests that MSP may also play a role in rockfish reproductive behavior.

There are limitations in identifying positive selection via K_a/K_s estimation in functional gene regions. Hughes (2007) suggests that current methods for detecting selection are not capable of identifying selection via single nucleotide polymorphism (SNP), deletion or silencing of genes, or differences in gene expression. It is plausible that such changes may also contribute to the adaptive radiation found within *Sebastes.* However, our initial scan in rockfish, which may provide insight into which genes are under positive Darwinian selection for this group, also identify functional gene categories of selected genes have also been detected in numerous other genomic scans (Heger and Ponting, 2007; Kosiol et al. 2008; Clark et al. 2007). Future work should attempt to understand not only the types and strength of selection on these (and other candidates) within this group of fishes, but also the extent to which specific loci may have been subject to episodic selection within restricted radiations within the genus and regulatory changes that may have also contributed to divergence within the group.

2.4 Conclusion

This study has provided an initial glimpse of natural selection between two species of rockfish. Genes with elevated K_a/K_s values belonged to the following functional categories (immune function, metabolism, longevity, and reproductive behavior), indicating that adaptive divergence at immunological, physiological, and reproductive loci may be important in the diversification of this group of fishes. What remains to be seen, is if the loci that we have found to be under positive selection are restricted to the *Pteropodus* group or are under selection across a wider array of *Sebastes* species. This study provides the ground work to better understand the molecular evolution of genes involved in the radiation within this group of marine fishes.

2.5 Acknowledgements

We would like to thank D. Ardell (UCM) for assistance with developing PERL scripts and fruitful discussion regarding ortholog identification. We would also like to thank Kris von Schalburg and Amber Messmer for their help with library construction and sequencing. Tissue for *S. rastrelliger* was provided by Dr. N Sherwood (UVic). This work was partially supported by grants from Natural Sciences and Engineering Research Council (NSERC) to BFK, the UC Merced Graduate Division to JH and the National Science Foundation to AA (DEB-0719475).

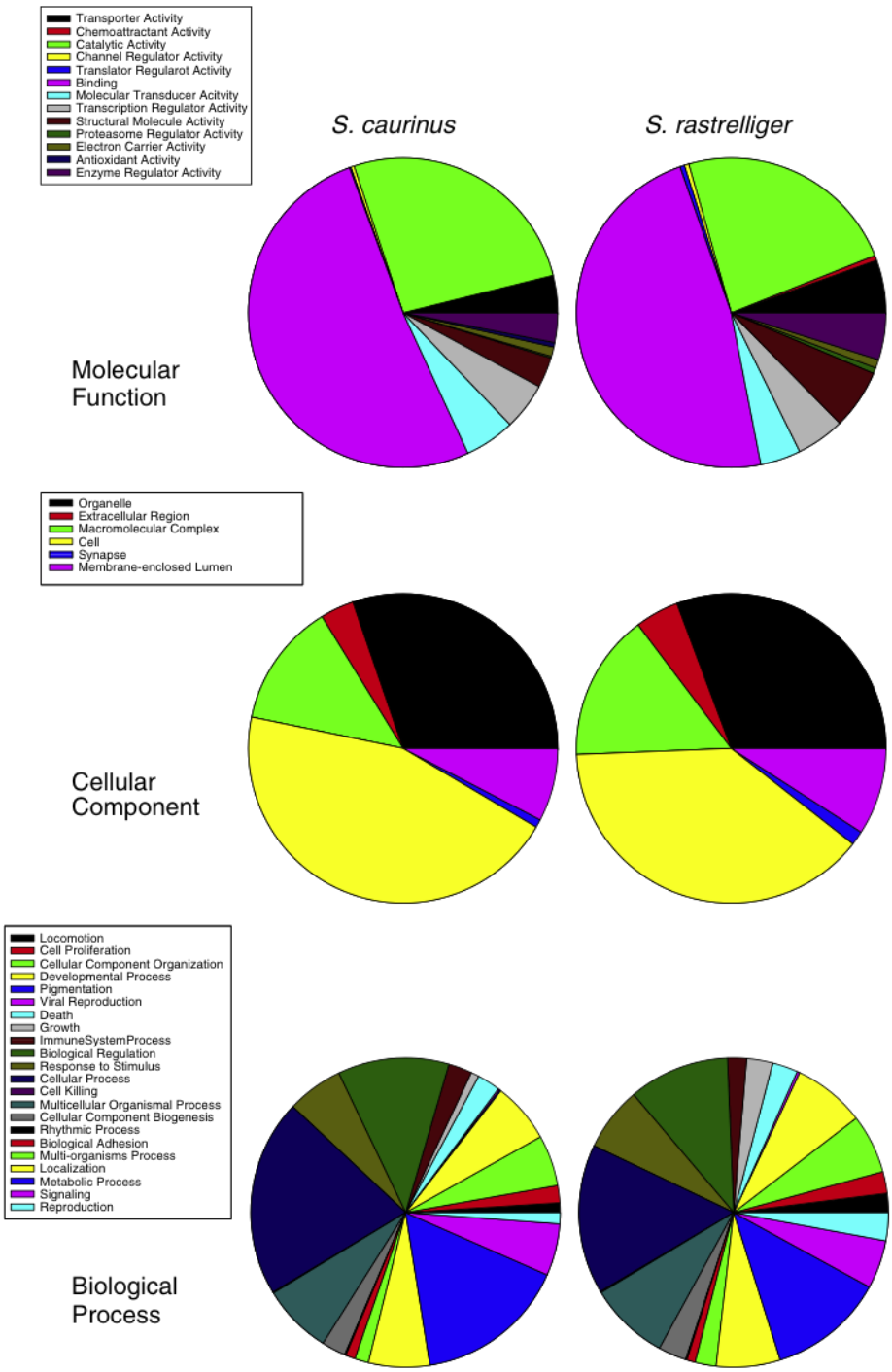

Figure 2.1: Results of the level two annotation to the SWISSPROT database for the two EST libraries.

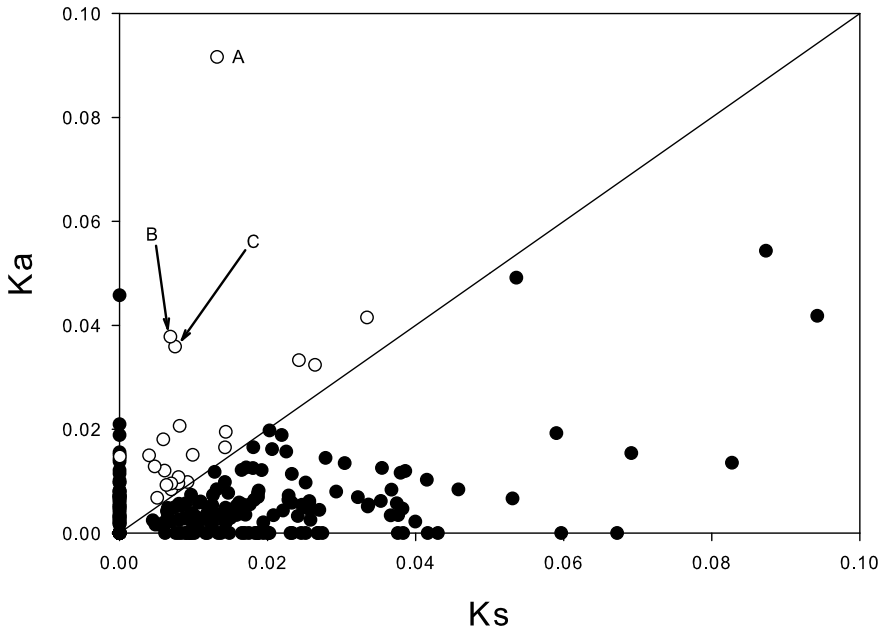

Figure 2.2: Results of the K_a/K_s analysis for the 257 orthologous gene pairs between *S. caurinus* and *S. rastrelliger*. The solid line indicates a K_a/K_s value of 1, values above the line are considered under positive selection. Three genes with high K_a/K_s values are indicated (A: Interferon inducible protein gig2; B: No annotation available; C: 14kDa apolipoprotein).

K_a/K_s	Seq. Description	Hit Accession Number	*S. caurinus* E-value	*S. caurinus* Similarity	*S. rastrelliger* E-value	*S. rastrelliger* Similarity	Gene Ontology[1]
1.069	protein ccsmst1 precursor	A8WGU8	$5.06E^{-13}$	60	$6.63E^{-12}$	68	C:integral to membrane; C:membrane
1.157	NA	NA	NA	NA	NA	NA	NA
1.198	eukaryotic translation initiation factor 3 subunit f	A5A6I3	$5.04E^{-120}$	89	$7.58E^{-107}$	89	C:cytosol; C:eukaryotic translation initiation factor 3 complex; F:protein binding; P:translational initiation; F:translation initiation factor activity
1.201	lipocalin precursor	Q01584	$2.41E^{-24}$	56	$3.36E^{-25}$	55	F:cysteine-type endopeptidase inhibitor activity; F:binding; F:prostaglandin-D synthase activity; P:transport; F:transporter activity; P:lipid metabolic process
1.225	NA	NA	NA	NA	NA	NA	NA
1.241	coiled-coil-helix-coiled-coil-helix domain-containing protein 5	Q9BSY4	$3.30E^{-25}$	71	$6.53E^{-26}$	73	NA
1.325	Elongation factor 2	Q90705	0	92	0	92	C:cytoplasm; F:translation elongation factor activity; F:GTP binding; F:protein binding; P:response to chemical stimulus; C:ribonucleoprotein complex; P:translation; F:GTPase activity
1.347	complement c1q-like protein 4	Q4ZJM9	$2.37E^{-11}$	50	$1.78E^{-11}$	50	P:negative regulation of cellular process
1.358	s100 calcium-binding protein a13	Q99584	$3.88E^{-10}$	62	$4.98E^{-10}$	61	F:RAGE receptor binding; F:drug binding; C:cytosol; F:copper ion binding; F:zinc ion binding; P:regulation of cell shape; F:calcium ion binding; P:interleukin-1 alpha secretion; C:extracellular space; P:positive regulation of I-kappaB

							kinase/NF-kappaB cascade; P:response to electrical stimulus; P:response to copper ion; C:perinuclear region of cytoplasm; P:positive regulation of cell proliferation; F:fibroblast growth factor 1 binding; F:protein homodimerization activity; C:nucleus
1.361	Calcium and integrin-binding protein 1	Q99828	$3.24E^{-63}$	86	$4.67E^{-70}$	85	C:apical plasma membrane; P:apoptosis; C:nucleoplasm; P:integrin-mediated signaling pathway; F:protein binding; P:cell adhesion; P:double-strand break repair; F:calcium ion binding; C:filopodium; C:endoplasmic reticulum
1.373	beta-2-glycoprotein 1 / apolipoprotein h	P17690	$6.83E^{-63}$	55	$1.49E^{-62}$	55	P:anatomical structure morphogenesis; F:protein binding; P:regulation of fibrinolysis; P:organ development; P:localization; P:primary metabolic process; C:plasma lipoprotein particle; F:lipid binding; P:positive regulation of blood coagulation; P:negative regulation of apoptosis
	male-specific protein	Q6SKG4	$2.34E^{-54}$	71	$1.09E^{-54}$	71	NA
1.522	collectin-12	A5PMY6	$1.27E^{-44}$	73	$2.05E^{-21}$	92	F:scavenger receptor activity; F:sugar binding; P:innate immune response; P:phagocytosis, recognition; F:low-density lipoprotein binding; F:pattern recognition receptor activity; C:integral to membrane
1.973	glutathione s-transferase a4	P24472	$3.45E^{-66}$	73	$7.90E^{-63}$	72	P:metabolic process; F:glutathione transferase activity
2.534	novel protein	Q4SRN3	$1.38E^{-38}$	64	$1.28E^{-38}$	67	NA
2.697	corepressor interacting with rbpj / cbf1-	Q5ZI03	$3.93E^{-82}$	88	$3.39E^{-81}$	87	C:nucleolus; C:nuclear speck; P:RNA splicing; C:histone deacetylase complex; F:transcription corepressor activity; F:transcription factor activity; P:mRNA processing; P:negative

	interacting corepressor						regulation of transcription, DNA-dependent; C:cytoplasm
3.050	integrin beta-2 / cell surface adhesion glycoproteins lfa-1 cr3 subunit beta / complement receptor c3 subunit b	P05107	$6.53E^{-33}$	60	$2.69E^{-37}$	61	F:glycoprotein binding; P:regulation of peptidyl-tyrosine phosphorylation; F:protein kinase binding; P:leukocyte cell-cell adhesion; C:plasma membrane; P:neutrophil chemotaxis
3.739	histamine n-methyltransferase	P50135	$1.09E^{-53}$	63	$5.12E^{-57}$	63	P:respiratory gaseous exchange; F:methyltransferase activity
4.778	14 kda apolipoprotein	B9V2X5	$5.97E^{-52}$	78	$2.99E^{-46}$	78	NA
5.502	NA	NA	NA	NA	NA	NA	NA
6.959	interferon-inducible protein gig2	XP_001344537[2]	$9.47E^{-42}$	75	$4.17E^{-42}$	74	F:NAD+ ADP-ribosyltransferase activity

Table 2.1: Annotation information for the 21 ortholog pairs with Ka/Ks > 1. Hit accessions are from the SWISS-PROT database unless otherwise noted. NA indicates an annotation was not available for the ortholog pair.

1- C: Cellular Component; F: Molecular Function; P: Biological Process

2- Accession from GenBank

Chapter 3

Gonadal transcriptomics elucidate patterns of adaptive evolution within marine rockfishes (*Sebastes*)

Joseph Heras[1], Kelly McClintock[1], Shinichi Sunagawa[2], and Andres Aguilar[3*]

1- School of Natural Sciences and Graduate Group in Quantitative and Systems Biology, University of California Merced, 5200 N. Lake Rd., Merced, CA USA 95344

2- European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany

3- Department of Biological Sciences, California State University, Los Angeles, 5151 State University Dr. Los Angeles, CA 90032

* - Author for correspondence: Joseph Heras, School of Natural Sciences, University of California, Merced, Merced, USA, (209) 228-4508, jheras@ucmerced.edu

Data deposition: Raw reads from *S. goodei* were deposited to dbEST under the accession numbers JZ693907-JZ704944. Short reads from *S. saxicola* were deposited to the Short Read Archive under the accession SRR1212396.

Abstract

The genetic mechanisms of speciation and adaptation in the marine environment are not well understood. The rockfish genus *Sebastes* provides a unique model for studying speciation because there are limited examples of species flocks within marine systems. Using marine rockfishes, we identified signatures of natural selection from transcriptomes developed from gonadal tissue of two rockfish species (*Sebastes goodei* and *S. saxicola*). We predicted orthologous transcript pairs, and estimated their distributions of nonsynonymous (K_a) and synonymous (K_s) substitution rates. We identified 144 genes out of 1,079 orthologous pairs under positive selection, of which 11 are functionally annotated to reproduction based on gene ontologies (GOs). One orthologous pair of the zona pellucida gene family, which is known for its role in the selection of sperm by oocytes, out of ten was identified to be evolving under positive selection. In addition to our results in the protein coding-regions of transcripts, we found substitution rates in 3' and 5' UTRs to be significantly lower than Ks substitution rates implying negative selection in these regions. Our study provides insight on the mechanisms of adaptive evolution within a highly speciose group.

Key Words: Bioinformatics, Orthologs, Positive Selection, Reproductive Genes, Untranslated Region, and Zona Pellucida

3.1 Introduction

Genomic information can increase our understanding of the molecular evolutionary processes that drive speciation (Elmer et al. 2010). Comparative genomic and transcriptomic studies have provided a framework for understanding how genes and genomic sequences relate to adaptation and phenotypic evolution at the organismal level (Ellegren 2008). Many of these comparative studies (Clark et al. 2007; Elmer et al. 2010; Heger and Ponting, 2007) identify coding genes that are subject to rapid divergence and positive selection, a process where mutations are advantageous and favored. Either a single or the accumulation of advantageous mutations can contribute to the process of adaptive evolution. The identification of positive selection at the molecular level has been frequently estimated by the calculation of nonsynonymous (K_a) and synonymous (K_s) substitutions, in which a K_a/K_s ratio greater than one is an indication of positive selection and a value less than one is indicative of negative or purifying selection, the purging of deleterious alleles (Miyata et al. 1980; Nei and Gojobori, 1986; Goldman and Yang, 1994). Genes under positive selection are generally categorized within comparative genomic studies under processes such as biosynthesis, development, metabolic processes, immune function and reproduction (Clark et al. 2007; Elmer et al. 2010; Heger and Ponting 2007).

Identifying the mechanisms of speciation within marine systems has been a daunting and difficult task. Most studies (post Mayr; 1942) have focused on identifying geographic barriers that would prompt allopatric speciation (Taylor and Hellberg, 2005). However, within marine ecosystems there are limited geographic barriers that would inhibit gene flow, which suggests that speciation events within these systems would not be solely a consequence of allopatric speciation (Ingram, 2011). This concept suggests a marine-speciation paradox, where incipient species that come into contact frequently would prevent allopatric speciation (Bierne et al. 2003), in which marine rockfishes provide an ideal model system for understanding what are the mechanisms that contribute to the speciation process. Ingram (2011) demonstrated that the adaptive radiation of rockfish species stem from the consequence of parapatry, in which phylogenetic evidence exhibits differences in depth and depth-related morphology among species within this group. If the evolutionary processes within rockfishes are a consequence of ecological differences, sexual selection can still play an important role within the speciation process. Although limited in the literature, there is evidence that supports that both sexual selection and environmental heterogeneity can play an important role in speciation (Maan and Seehausen, 2011). Our understanding of these evolutionary mechanisms is incomplete and a lucid description of the molecular evolutionary processes within marine organisms can help resolve this marine-speciation paradox.

Divergent sexual selection can facilitate the speciation process via reproductive traits that form a barrier between incipient species and result in reproductive isolation (Endler and Basolo, 1998; Swanson and Vacquier, 2002; Masta and Maddison, 2002). Other factors like spawning time, mate recognition, environmental tolerance, and gamete compatibility are thought to contribute to the marine speciation process (Palumbi, 1994). Several molecular evolutionary studies have demonstrated that genes associated with reproduction (i.e. genes that encode for gamete recognition proteins) have rapidly diverged between closely related taxa (Aagaard et al. 2010; Lee and Vacquier, 1995;

Turner and Hoekstra, 2006). Swanson and Vacquier (2002) suggested that the rapid divergence within reproductive genes may stem from a single or combination of selective pressures such as sperm competition, sexual selection and sexual conflict. Levitan and Ferrell (2006) demonstrated how sperm competition operates within male and female sea urchins (*Strongylocentrotus franciscanus*) in which mating pairs that had the most common bindin (sperm protein) genotypes had higher reproductive success in the presence of low polyspermy – the fertilization of an egg with multiple sperm. However, when polyspermy levels were high, males and females with unmatched bindin genotypes had the selective advantage. This depicts an "arms race" between sperm and egg proteins, in which sperm competition is a source of directional selection, and egg proteins are also under selective pressures to develop barriers against polyspermy (Palumbi, 2009). Although a vast amount of information supports the rapid diversification of reproductive genes, very little is known about the forms of selection operating on these genes (Pujolar and Pogson, 2011). Most studies on gamete evolution within marine systems have been demonstrated with free spawning organisms (Aagaard et al. 2010; Clark et al. 2007b; Levitan and Ferrell, 2006). In contrast, marine rockfishes are viviparous and the evolutionary processes of gamete recognition proteins within this group are unknown. However, Sogard et al. 2008 demonstrated multiple paternity within *S. atrovirens* (kelp rockfish), which permits the opportunity for selective forces to operate on reproductive proteins (e.g. sperm competition and the prevention of polyspermy).

A prime candidate for understanding reproductive barriers at the molecular level is the zona pellucida (ZP) gene family, which encodes for glycoproteins that create the acellular vitelline envelope around the oocyte (Evans, 2000; Modig et al. 2007; Spargo and Hope, 2003). The function of ZP proteins varies in fishes and includes uptake of nutrition, functional buoyancy (Podolsky, 2002), protection of the growing oocyte, species-specific binding, and guidance of the sperm to the micropyle (Dumont and Brummet, 1980). There are at least eight ZP genes in many fish species (Tian et al. 2009) that belong to three subfamilies: ZPB, ZPC, and ZPAX (Modig et al. 2007). The subfamily ZPA is missing from fishes, which may be due to a gene deletion (Smith et al. 2005) and subfamilies ZPC and ZPB are known to contain gene duplicates (Goudet et al. 2008). Selection has been tested in ZPC genes in 6 teleost fish species, but the results have been inconclusive due to the lack of robustness in the statistical methods used (Berlin and Smith, 2005). In this study, we wanted to address more closely the hypothesis that genes in the ZP family may provide a reproductive barrier between closely related species.

Rockfishes (genus *Sebastes*) are a prime system for understanding adaptive radiations and the mechanisms of speciation within marine systems (Burford and Bernardi, 2008). Adaptive radiations involve rapid divergence of multiple lineages, which serve as replicates of speciation within a given environment or time frame (Burford and Bernardi, 2008). *Sebastes* spp. has been considered an ancient species flock (Johns and Avise, 1998), a group of closely related species with a monophyletic origin (Hyde and Vetter, 2007). The genus arose around 8 mya, contains 22 recognized subgenera, (Kendall, 2000), and approximately 105 species found worldwide (Hyde and Vetter, 2007). In addition, this genus is composed of species that are morphologically and ecologically divergent (Ingram, 2011; Magnuson-Ford et al. 2009), with a center of

diversity for this group is located in the Northeast Pacific (Love et al. 2002). Though many studies have concentrated on describing species-level variation (Johns and Avise, 1998; Rocha-Olivares et al. 1999; Li et al. 2006; and Hyde and Vetter, 2007), very few studies have investigated the genetic mechanisms that have contributed to this radiation (Sivasundar and Palumbi, 2010; Heras et al. 2011).

In this study, our aims were to identify and characterize genes subject to positive selection between two marine fish species in *Sebastes*. We used a comparative transcriptomic approach, in which we characterized and compared transcriptomes generated from gonadal tissues of the two species *S. goodei* and *S. saxicola*. We selected *S. goodei* (chilipepper - Eigenmann and Eigenmann, 1890) and *S. saxicola* (stripetail – Gilbert, 1890) based on the extensive amount of evolutionary time since their most recent common ancestor (estimated to be greater than 6 million years ago [mya] Hyde and Vetter 2007), which can give a broader depiction of which functional genes have diverged within this genus. In addition, gonadal tissues were selected for this study to locate highly divergent reproductive genes, which can serve as candidates for investigating positive selection across the entire genus. Our reasoning for the two different sequencing methods was that each library was prepared with the latest sequencing technology that was available at the time. We annotated the function of expressed genes using gene ontology (GO), and identified signatures of positive selection from estimates of K_a/K_s ratios for ortholog pairs that we annotated between these two species. Genes that were found to be evolving under positive selection were further analyzed in the context of their orthologs in model fishes and ESTs from our earlier study (Heras et al. 2011). In addition to identifying selection through analysis of coding regions, we additionally estimated genetic divergence between the two species in untranslated regions (UTRs). Overall, this study was developed to understand how differences at the genomic level contribute to adaptive evolution within this speciose group.

3.2 Materials and Methods

3.2.1 EST Sequencing and Assemblies: *S. goodei*

Ovary and testes were collected from fresh dead *S. goodei* individuals (one per sex), placed immediately in RNAlater, and stored at -80ºC. Complementary DNA (cDNA) isolation and library construction was performed by BIO S&T (Montreal, Canada). Total RNA was extracted with TRIzol (Invitrogen, Carlsbad, CA), and cDNA was synthesized according to the SMART cDNA library construction kit (Clontech, USA). The resulting cDNAs were full-length enriched, and possess SfiI A&B at the 5' and 3' ends which facilitated directional cloning. Double-stranded cDNAs were obtained by primer extension. Double-stranded cDNAs were digested with SfiI, afterwards only fragments greater than 0.5 kb were purified with a gel purification kit.

Purified cDNA was ligated to SfiI-digested and Calf intestinal phosphatase (CIPed) vectors by overnight incubation at 16°C. The ligation mixture was desalted and electroporated in ElectroMax DH10B cells (Gibco-BRL, USA). Quality control (average cDNA insert sizes and recombinant rate) was performed prior to mass transformation. Transformed cells were distributed into 96-deep-well plates for amplification at about 2,300 recombinants per well.

Cells were plated onto LB-agar (amplicillin and x-gal) plates. Clones were prepared for sequencing in two ways. Method 1 – positive colonies were picked directly into 96-well plates that contained LB broth + ampicillin. Cultures were grown overnight at 37^{o}C with moderate shaking. The Montage Plasmid Miniprep HTS kit (Millipore) was used to isolate plasmid DNA. Sequencing on purified plasmid DNA was done with M13 (-20 and +40) primers at JGI as implemented in Sunagawa et al. (2009). Method 2 – cDNA libraries were produced by double-stranded cDNA, which was size fractionated to obtain long reads. Afterwards, cDNA inserts were cloned into the vector pExpress1 (Express Genomics, Frederick, MD), and electroporated into *E. coli* strain DH10B. Libraries contained ~96% recombinants with an average insert size of 1.95 kb. Libraries were sequenced on 96-well capillary sequencing platforms (ABI 3700) located at the DOE Joint Genome Institute (JGI, Walnut Creek, CA) and at the Genome Core Facility at the University of California, Merced, CA.

Expressed Sequence Tags (ESTs) were cleaned and assembled with an automated pipeline (EST2uni, Forment et al. 2008), which includes base calling (PHRED), vector trimming and low quality bases removal with LUCY (Chou and Holmes, 2001), and repeat masking with REPEATMASKER-OPEN 3.0 (Smit et al. 2010). Afterwards, the assembly of sequencing reads into unique consensus sequences (unigenes, Forment et al. 2008) was conducted with CAP3 (Huang and Maden, 1999), and functional annotations were conducted with BLAST (Altschul et al. 1990), in which the hits are then parsed so that a description is listed for each unigene. The unigene datasets are composed of high quality and clean sequences, which are assembled into contigs and singletons (Forment et al. 2008). These sequences can be found at GenBank with the following accession numbers: *Sebastes goodei* JZ693907-JZ704944. Unigenes were processed again with CAP3 (Huang and Maden, 1999) to correct for putative assembly errors and then used for the comparative transcriptomic analysis against the *S. saxicola* dataset.

3.2.2 EST Sequencing and Assemblies: *S. saxicola*

Ovary tissue was collected from a single fresh dead *S. saxicola* individual, placed immediately in RNAlater, and stored at -80ºC. Complementary DNA (cDNA) isolation and library construction for 454 sequencing was performed by BIO S&T (Montreal, Canada). The library was sequenced at the University of South Carolina Environmental Genomics Core facility on a Roche 454 sequencer. The library was sequenced on a ½ of a titer plate.

The *S. saxicola* raw reads and base quality information from the 454 GS FLX sequencing run were first extracted and clipped using the SFF_EXTRACT 0.2.8 (Blanca and Chevreux 2010) script. Further removal of adaptors and contamination, such as low quality bases and poly (A) stretches, was achieved by using SNOWHITE 1.1.4 (Barker et al. 2010), a pipeline that implements aggressive cleaning with SEQCLEAN (http://compbio.dfci.harvard.edu/tgi/software) and TAGDUST 1.12 (Lassmann et al. 2009). Reads were then processed through REPEATMASKER-OPEN 3.0 (Smit et al. 2010) using the CROSS_MATCH (Downloaded June 2010; Green, 2010) search engine to search the "teleost fish" database and mask repetitive elements. A primary *de novo* assembly was initially done using the 454 default settings in MIRA 3.2.0 (Chevreux et al. 2004) with a minimum percent identity of 94%. A secondary assembly was performed on the contigs produced from MIRA 3.2.0 (Chevreux et al. 2004) and all remaining singletons in CAP3 (Huang and Maden 1999). A minimum overlap of 25 bp and a minimum %ID of overlap of 95% was used in the secondary assembly. Finally all contigs less than 300 bp in length were removed before additional analyses.

3.2.3 Annotation of the *S. goodei* and *S. saxicola* Datasets

Both EST datasets were annotated in BLAST2GO (Conesa et al. 2005) with the following Blast parameters: BLASTX to the SWISS-PROT database (Bairoch and Apweiler 1997), an E-value of 1.0 x E^{-6}, 20 BLAST hits, and a High Scoring Pair length cutoff of 33 nt. The annotation parameters were an E-value hit filter of 1.0 x E^{-6}, annotation score cutoff of 55, and a gene ontology (GO) weight of 5. A two-tailed Fisher's Exact Test was used in BLAST2GO (Conesa et al. 2005) to determine whether there was enrichment of GO terms for the orthologous pairs that contained a $K_a/K_s > 0.5$ in comparison to orthologous pairs that were conservative ($K_a/K_s < 0.1$).

3.2.4 Detection of Orthologs from the *S. goodei* and *S. saxicola* Datasets and Estimation of Selection

BLASTX (NCBI blast version, 2.2.17) from the standalone BLAST package (Altschul et al. 1990) was used to identify homologs in both *S. goodei* and *S. saxicola* ESTs against the SWISS-PROT database (Bairoch and Apweiler, 1997, downloaded June 2011) with five teleost fish datasets from fugu (*Takifugu rubripes*), medaka (*Oryzias latipes*), green spotted pufferfish (*Tetraodon nigroviridis*), stickleback (*Gasterosteus aculeatus*), and zebrafish (*Danio rerio*) in the Ensembl database (Hubbard et al. 2005, Ensembl 63). Afterwards the BLASTX reports and the EST sequences were processed through ORFPREDICTOR (Min et al. 2005), which identifies putative open reading frames and translates nucleotide sequences into protein sequences. The translated protein

datasets from *S. goodei* and *S. saxicola* were used in INPARANOID 4.0 (O'Brien et al. 2005) to identify orthologs and avoid the inclusion of paralogs. *Danio rerio* (Ensembl dataset, Zv9) was used as an outgroup for removing potential false orthologs. Orthologous pairs were aligned based on the putative open reading frame using PAL2NAL 12.2 (Suyama et al. 2006) and Perl scripts that include CLUSTAL W 2.0.10 (Larkin et al. 2007). K_a and K_s were calculated for the orthologous pairs between *S. goodei* and *S. saxicola* in KAKS_CALCULATOR 1.2 (Zhang et al. 2007) by using the YN model (Yang and Nielsen, 2000).

3.2.5 Ortholog Identification and Positive Selection

We used 5,336 and 18,505 contigs from *S. goodei* and *S. saxicola* ESTs respectively for the identification of orthologs and the K_a/K_s analyses. There were 1,559 orthologs detected with INPARANOID 4.0 (O'Brien et al. 2005). Once processed through KAKS_CALCULATOR v1.2 (Zhang et al. 2007), pairs were removed from our analyses if the alignment length was less than 150 bp and/or the K_a/K_s values were greater than 50. Ortholog pairs with a K_s value less than 0.1 were further analyzed, which has been used as a benchmark to avoid inclusion of paralogs (Bustamante et al. 2005). We also included a second set of ortholog pairs with K_s values within the range of 0.1-0.5 (Figure 3.1). This allowed us to determine whether the $K_s > 0.1$ benchmark should be extended for our analyses.

3.2.6 PAML Analyses and Zona Pellucida Phylogeny Construction

We analyzed three different datasets, in which we tested for adaptive evolution with the PAML v4.4 (Yang, 1997) software package. We used CODEML which is part of the PAML v4.4 (Yang, 1997) package and tested for positive selection with M7 and M8 models (Yang et al. 2000) and conducted Likelihood Ratio Tests (LRTs) between the two models. We conducted a TBLASTX search with additional datasets from *Oreochromis niloticus, Oryzias latipes, Sebastes rastrelliger*, and *Sebastes caurinus* to identify orthologs. Only ortholog pairs of length 65 codons or greater, and 85% identity were utilized for our analysis. The first dataset consisted of orthologs from *Oreochromis niloticus* (Nile Tilapia), *Oryzias latipes* (Medaka) and the two focal *Sebastes* species, which contained eleven genes. Orthologs were identified for genes with elevated K_a/K_s values. These two model species were chosen due to their close relationship to rockfishes when we analyzed our ZP phylogenies. The second dataset included additional orthologs identified from a previous study (*S. caurinus* and *S. rastrelliger*; Heras et al. 2011) to further validate signatures of adaptive evolution within the genus that contained four gene pairs.

A third dataset contained sequences from the zona pellucida (ZP) gene family with 5 gene pairs. Sequences annotated to this family were used to construct a phylogeny of the ZP gene family and a fine-scaled analysis of positive selection. *S. goodei* and *S. saxicola* sequences were trimmed and translated within ORF PREDICTOR (Min et al. 2005). Based on the annotations (assignment of ZP subfamily), the longest ESTs from the two *Sebastes* species were used for phylogenetic analyses and the following subfamilies were identified: ZPAX, ZPB, and ZPC. ZP subfamilies (one sequence alignment dataset for

ZPAX and ZPB, and another for ZPC) were aligned with MAFFT 6 (http://mafft.cbrc.jp/alignment/server/) and a ZPA homolog from *Xenopus levis* was used as an outgroup. These sequences along with teleosts fish sequences with known ZP annotation (Modig et al. 2007) and the top TBLASTX hits from GenBank were translated and aligned in MAFFT 6 (http://mafft.cbrc.jp/alignment/server/). After alignment, sequences were processed through PROTTEST 3 (Abascal et al. 2005) to determine a model of protein evolution. Phylogenies were constructed with aligned sequences and a selected protein model in PHYML 3.0 (Guindon et al. 2010). If both a *S. goodei* and *S. saxicola* homologous pair were present, they were processed in KAKS_CALCULATOR v1.2 (Zhang et al. 2007) by using the YN model (Yang and Nielsen, 2000) to estimate positive selection.

3.2.7 UTR Divergence

We were interested in the neutral substitutional mutation rate within our transcriptomic datasets. In addition, we expected the UTR regions to be highly divergent only if paralogs were identified in our ortholog search between the two datasets. This will give an indication that our K_s cut-off at 0.5 is valid. We developed scripts, which were used to remove 5' and 3' UTRs from the orthologous pairs and conduct a pairwise alignment in MUSCLE v3.7 (Edgar, 2004). Lastly, we estimated sequence divergence using a Jukes-Cantor model as suggested by Elmer et al. (2010). Only pairs greater than 50 bp were used for our analyses. Only pairs that contained both a 5' and 3' estimate were removed to prevent a partial paired analysis and we conducted a pair-wise BLAST of the orthologs to assess the quality of our alignments. BLAST scores of 90 bits or greater were included for our divergence analysis. Coding regions were reprocessed through KAKS_CALCULATOR 1.2 (Zhang et al. 2007) and K_s values were estimated with a Jukes-Cantor correction in order to make comparisons. Simple pairwise t-tests and Wilcoxon Rank Sum Tests were calculated between and within coding regions and UTRs by using R (http://www.R-project.org).

3.3 Results

3.3.1 Sequence Statistics and Annotation

The *S. goodei* ESTs contained 2,370 and 13,824 raw sequences respectively and a mean EST length of 655.9 and 614 bp respectively (Table 3.1). We assembled 6,139 unigenes, which were composed of 664 singletons and 630 contigs from ovary tissue, and 2,849 singletons, and 1,996 contigs from testes tissue. When processed through a second run of CAP3 (Huang and Madan, 1999), from the 6,139 sequences were reduced to 5,336 contigs and used for our comparative analyses with *S. saxicola*.

The *S. saxicola* ESTs contained a primary assembly of 311,289 reads and 295,114 clean reads. The primary assembly contained 85,431 singletons and 51,310 contigs with 71% redundancy (Table 3.2). From these 136,741 sequences, a second assembly was processed and contained 41,174 singletons, 14,090 primary contigs, and 23,475 secondary contigs. Only sequences that were assembled into contigs and greater than 300 bp were used for our comparative analyses resulted in 3,112 primary contigs and 15,393 secondary contigs were used with a total of 18,505 contigs.

There were 2,480 and 8,763 sequences from *S. goodei* and *S. saxicola* datasets that were annotated. Within the *S. goodei* and *S. saxicola* datasets, there were Gene Ontologies (GO) terms within the biological process domain, that belonged to the cellular process, metabolic process, biological process, multicellular organismal process, developmental process, cellular component organization, response to stimulus, localization, signaling, cellular component biogenesis, reproduction, death, growth, cell proliferation, immune system process, and multi-organism process. Most GO terms represented for molecular function pertained to binding, catalytic activity, transcription regulator activity, molecular transducer activity, transporter activity, enzyme regulator activity, structural molecule activity, and electron carrier activity. The majority of GO terms represented for cellular component pertained to the cell, organelle, macromolecular complex, membrane-enclosed lumen, extracellular region, and synapse.

Our annotations of the two (*S. goodei* and *S. saxicola*) transcriptomes were relatively similar across the major three divisions (Biological Process, Molecular Function, and Cellular Component) when levels 2 and 3 GO terms were compared. In most GO terms, *S. goodei* were slightly elevated, with 2,480 annotated contigs and for *S. saxicola* - 8,763 contigs. Although there were differences between the two sequencing methods, there were similarities in GO categories between the two transcriptomes. In addition, the two datasets showed 16% (*S. saxicola*) and 17% (*S. goodei*) of GO terms annotated to reproduction and 35% and 39% for developmental processes (respectively), which may provide an overview of reproductive processes within ovary tissues. In addition, the dual use of testes and ovary tissues from *S. goodei* contained similar GO terms between *S. saxicola* in which these two tissue types may contain similar GO functional traits.

3.3.2 Genes Under Positive Selection

Two hundred nine ortholog pairs contained a K_a/K_s less than 0.1, 726 pairs were between 0.1-1.0 (K_a/K_s) and 144 pairs that were found greater than 1 (positive selection, Figure 3.2). Seventy-one of these pairs were annotated with a majority of the sequences that were associated with macromolecule metabolic processes and regulation of biological processes based on the sequence distribution of Gene Ontologies (Table 3.3). Only a small fraction of the distribution of GOs was associated with reproductive process (11 orthologous pairs) and sexual reproduction (8 orthologous pairs). The average K_a/K_s value was 0.53 (s.d. = 0.62), and the average ortholog alignment length of 361.37 (s.d. = 132.45). There was no enrichment found between these two categories with a False Discovery Rate of (0.05).

3.3.3 PAML Analyses and Zona Pellucida Phylogeny Construction

From the 11 out of the 71 annotated genes found to be under positive selection in our first PAML dataset, the LRTs conducted showed that there was no significant difference between models M7 and M8 (Yang et al. 2000). From the second dataset, which contained our two rockfish species of interest as well as S. *caurinus* and *S. rastrelliger*, only two out of the four were identified to be under adaptive evolution (M8 was significantly different from M7 when using the LRTs). The two genes were FKB12 and TM50a, which contained 5 and 4 sites under positive selection respectively. In our third dataset, which was composed of five ZP genes from our two rockfish species, *Oreochromis niloticus* and *Oryzias latipes* did not demonstrate signatures of positive selection according to our LRTs analysis (Table 3.4).

In our construction of the gene family for ZP within rockfishes we first identified 18 and 26 ESTs that contained ZP annotations for *S. goodei* and *S. saxicola* respectively. Maximum likelihood (ML) trees were constructed with 143 ungapped aa sites (1075 total sites) and 92 ungapped aa sites (697 total sites) for the ZPAX and ZPB, and ZPC respectively (Figure 3.4 and 3.5). In our phylogenetic analysis, 7 ZPC, 2 ZPB, and 1 ZPAX homologs were identified (Figure 3.4 and 3.5). Some ESTs were excluded from this analysis (two ZPB fragments), because these fragments did not align with the majority of the remaining sequences, however were included in the K_a/K_s analysis. From the PAML analysis, 5 ZP ortholog groups were compared. This was based on the ortholog groups identified (*Sebastes* sequences, *Oryzias latipes*, and *Oreochromis niloticus*). The K_a/K_s comparison was conducted with ten ZP genes (six ZPC pairs, three ZPB pairs and one ZPAX pair), where only one pair (ZPB homolog) was identified under positive selection (Table 3.5).

3.3.4 UTR Divergence

Based on 1,079 pairwise comparisons between the two rockfish species the average K_a was 0.034 (s.d. = 0.053) and an average K_s value of 0.067 (s.d. = 0.077) by using the YN model. The untranslated region (UTR) divergence estimates between the two fishes were based on 311 and 192 pairwise comparisons for 5' and 3' UTRs respectively. The 5' UTR estimates with a Jukes-Cantor correction were 0.026 (s.d. = 0.025) and K_s values (Jukes-Cantor correction) from the corresponding coding sequences

was 0.063, (s.d. = 0.068). The 3' UTR average was 0.023 (s.d. = 0.024) from the 193 corresponding coding sequences and contained an average K_s value of 0.076 (s.d. = 0.089). Overall, the means for the UTRs were statistically less than the means from the K_s values and there were no clear relationships between UTRs and K_s values. In a pair-wise simple t-test and the Wilcoxon rank sum test there were only two comparisons that did not show any mean differences (5' UTR ends vs. 3' UTR ends, and 5' K_s from coding regions vs. 3' K_s from coding regions - Table 3.6).

3.4 Discussion

3.4.1 Natural Selection

In our study genes found under positive selection did not overlap with the genes found in a previous *Sebastes* comparative transcriptome study (Heras et al. 2011), which is not surprising given the differences in the tissue types. However, we did find similar functional traits (metabolism, immune function, and longevity) for genes under selection. In the earlier study, the transcriptomic analysis was conducted with brain, pituitary, kidney, and spleen tissues, which may differ in expression patterns from gonadal tissues. The complexity of expression among different tissue types is still currently being understood, in which genes once thought to be expressed in a tissue specific fashion have been identified in multiple tissues (von Schalburg et al. 2005). As more rockfish tissue-specific transcriptomic information becomes available, the determination of whether certain genes subject to positive selection belong to specific tissues can be determined.

Our scan for genes under positive selection also include genes with elevated K_a values (Ks = 0) that contained GO terms that were associated with adult life spans and gamete function/production. Genes with only nonsynonymous substitutions and assigned with a GO terms associated with gamete production/function were the t-complex protein 1 and lissencephaly-1 homolog. Jansa et al. (2003) has demonstrated that the t-complex protein 1 and zona pellucida – 3 (homologous to ZPC in fishes), and immune system protein $\beta_2 m$ in a group of closely related murine genus *Mus* contain sites under positive selection. T-complex protein 1 is expressed during spermatogenesis in murids (Jansa et al. 2003), but the specific function is still unknown. This gene is highly expressed within mouse testes and is suggested to maintain normal spermatogenesis. Lissencephaly-1 has been demonstrated to be conserved (Peterfy et al. 1994) when compared between mice and humans. This gene has been shown to demonstrate infertility when a homozygous mutant has been developed (Escalier 2006). Although only nonsynonymous substitutions were found, one likely scenario for these genes for elevated K_a values is because these are only fragments of the entire gene sequence. These genes would be interesting to examine at the population level within each respective species in order to determine whether there is variation found at both synonymous and nonsynonymous sites.

Ortholog pairs under positive selection with a $K_a/K_s > 1$ and GO terms associated with gamete production/function were deadenylating nuclease, DNA ligase III, DNA mismatch repair protein, eukaryotic translation initiation factor 2, and homolog subfamily a member 4 (Table 3.3). DNA repair mechanisms have a strong relationship with gametogenesis, where the genomes of gametic cells are subject to mutations such as recombination (Baarends et al. 2001). Within these gametic cells the repair mechanisms have to tolerate mutations that occur during gametogenesis which result in specialized functions (Baarends et al. 2001) that are possibly due to selective pressures. Deadenlyating nuclease has been suggested to silence maternal mRNA during oocyte formation (Körner et al. 1998), this is particularly interesting due to the transcript comparison between *S. goodei* testes and *S. saxicola* ovary tissue. Homolog subfamily a member 4 is known to be part of the DnaJ family, which is assigned to the structurally unrelated protein family of Heat Shock Proteins (HSPs; Vos et al. 2008). In humans, this gene is expressed in brain tissue, but many homologs within the family are associated

with sperm motility. García-Herrero et al. (2010) demonstrated differences in reproductive genes between infertile vs. fertile human males, in which DnaJ subfamily A was represented. Clearly, these genes need to be further investigated to understand the mechanistic and functional properties within rockfishes to understand how these genes are subjected to positive selection.

Within our scan for positively selected genes we identified genes associated with longevity. Although the two species have similar lifespans, there are extensive differences between life spans across species within the genus (Love et al. 2002), and genes associated with longevity were identified within our previous study (Heras et al. 2011). The congener closely related to *S. goodei* is *S. paucispinis* (Hyde and Vetter 2007), which can live up to 46 years (Munk, 2001). In juxtaposition, the nearest congener to *S. saxicola* is *S. semicintus*, which can live up to 15 years (Love et al. 1990). Genes we identified and associated with longevity were eukaryotic translation initiation factor 2 subunit 3, cytochrome c oxidase subunit, 40s ribosomal protein x isoform, and 60s ribosomal proteins L9 and L17, and protection of telomeres protein 1 (Chen et al. 2007; Deelen et al. 2013; Hansen et al. 2007). These genes associated with longevity are particularly interesting and hold the potential key for understanding how aging operates in this group of fishes. As more rockfish genomic information becomes available this will provide a clearer depiction of the patterns of longevity and how this may impact adaptation.

The genes that showed evidence of positive selection in our PAML analysis were 12-kDa FK506-binding protein (FKBP12) and transmembrane protein 50a (TM50a). FKB12 is known to be associated with various cellular functions that include apoptosis, cell-cycle progression, and calcium release (Somarelli and Herrera 2007). Genes that encode for the mechanisms of apoptosis have been suggested to be under positive selection (Ellegren, 2008). Speculation for why these genes are under selection is due to the genomic conflict that would occur as a result of apoptosis during spermatogenesis (Nielsen et al. 2005). As for TM50a, this gene encodes for a membrane protein and the function of this gene within fishes is unknown. There is more information needed to determine how these genes contribute to adaptation within *Sebastes*.

Currently, there is much debate over the assessment of natural selection at the molecular level. Hughes (2007) has suggested that current statistical methods of estimating K_a/K_s across an entire gene does not account for the relaxation of purifying selection, and/or the effects of population bottlenecks. In addition, Ellegren (2008) has suggested that an estimate of K_a/K_s is an overly conservative estimate of positive selection, because most of the protein is under a functional constraint and only a few amino acid sites would be subject to positive selection. However, within many comparative genomic studies there are genes that have been identified under positive selection which encode for proteins with immune or reproductive functions (Heger and Ponting, 2007; Kosiol et al. 2008). Although there may be difficulties detecting selection, there are reoccurring gene functions that are subject to positive selection. Within our study, we have identified certain genes under positive selection that encompassed a broad range of GO terms where a majority of terms include: cellular process, metabolic process, biological regulation, response to stimulus, multi-cellular organ process, cellular component organization, developmental process, localization,

signaling, and reproduction. The specifics about how these genes under positive selection contribute to adaptation within heterogeneous environments remains unknown, but provides a suite of candidates for understanding why these genes have evaluated nonsynonymous substitutions in comparison to the remainder of the transcriptome.

3.4.2 Zona Pellucida

Current evidence shows that there are six subfamilies of zona pellucida genes in vertebrates (ZPA/ZP2, ZPB/ZP4, ZPC/ZP3, ZPD, ZPAX, and ZP1: Goudet et al. 2008) and these are homologous with the ZP domain found within invertebrates (Jovine et al. 2005). Our phylogenetic construction of the ZP family suggests there is only one ZPAX gene, 2 (putatively 4) ZPB homologs, and 7 (ZPC/ZP3) homologs in our dataset. Most ZP genes within the rockfish genome grouped with *Oreochromis niloticus* and *Oryzias latipes*, which suggests these genes have arisen in a similar pattern from a recent common ancestor (Figures 3.4 & 3.5).

In our estimation of K_a/K_s of 10 ZP gene pairs most pairs contained a broad range of K_a/K_s values (Table 3.5) with only one ortholog pair that was subject to positive selection (ZPB homolog 1). Both ZPB homologs (1 and 2) were not displayed within our phylogenetic analyses because there was limited phylogenetic information due to short length. When added, these genes provided weak bootstrap support to the ZPAX and ZPB ML phylogenetic analysis and were removed. However, these genes are divergent from the remaining ZP homologs and an ortholog from one of the model teleost fish could not be detected. It is unknown if some of these ZP homologs are specific to the *Sebastes* lineage, where more information from species within this genus and closely related genera or families would be needed to make this assessment. Currently, there is no evidence of teleost ZP genes subject to positive selection (Meslin et al. 2012), however this was assessed with a select few model fishes (i.e. *Danio rerio*, *Oryzias latipes*, *Gasterosteus aculeatus*, *Tetraodon nigroviridis*, and *Takifugu rubripes*). This poses the question of whether there is enough evidence to show that ZP genes do not provide evidence for positive selection within teleost fishes or is there some other mechanism that would prompt reproductive barriers? These methods are more stringent at identifying selection and the addition of more taxa from *Sebastes* can provide insight on how these genes have contributed to the radiation within this group.

3.4.3 UTR Analysis

Untranslated regions (UTRs) provide a reference of divergence between species and can be utilized as a base for comparing synonymous substitutions within coding regions that are assumed to be evolving neutral. Our estimation of 3' and 5' UTR divergence is unprecedented within the genus *Sebastes* and was compared to the Ks values (from 5' and 3' sequences) between *S. goodei* and *S. saxicola*, which were not proportionally similar. In addition, the utilization of the cutoff mark ($K_s < 0.1$) is not an essential benchmark for the removal of aligned pairs as putative paralogs according to our UTR analysis (Figure 3.5). Elmer et al. (2010) demonstrated that between two recently diverged (~10,000 ya) crater lake cichlid species the K_s coding region and 3' UTR divergence contained rates of 0.0250 and 0.0252 (with a Jukes-Cantor correction)

respectively. Hurst (2002) stated that synonymous rates are relatively proportional to the neutral mutation rate, which suggests that the UTRs and K_s are relatively close to this rate. However with species that are more divergent, there are distinct differences between synonymous rates and UTR divergence. Divergence within closely related *Drosophila* species are distinct where 3' UTR and 5' UTR rates are lower than synonymous sites when comparing *D. melanogaster* and *D. simulans* (Andolfatto, 2005), which diverged ~2-3 mya (Russo et al. 1995). Our study did not have similar K_s and UTR rates as compared to the Elmer et al. 2010 study, which may be due to the amount of time since divergence (estimated 6 mya). Andolfatto (2005) suggests that lower UTR divergence in comparison to synonymous sites in *Drosophila* is likely to be subject to negative selection, which is consistent with our findings. This pattern of 3' UTRs subject to purifying selection has also been identified within chimpanzees and humans (Hellmann et al. 2003). More evidence will be required to demonstrate the impact of negative selection on the marine rockfish genome, which analyzing the UTRs from closely to distantly related congeners can provide insight on this evolutionary pattern.

The use of the UTRs has been a resourceful indicator for assigning the correct ortholog pairs as opposed to paralogs, in addition to the algorithms used in INPARANOID (O'brien et al. 2005). Depending on the function of the gene, UTRs can be highly conserved between orthologs and divergent between paralogs once a gene duplication event has occurred, which has been demonstrated between humans and mice (Coy et al. 1995). One of the many difficulties of identifying orthologous gene pairs within teleosts fishes is the proposed fish specific genome duplication (FSGD) event which occurred ~350 mya (Meyer and Van de Peer, 2005). This event provides a plethora of gene duplicates that may operate under different evolutionary pressures such as subfunctionalization, neofunctionalization, and pseudogenization. With this magnitude of gene duplicates, the assignment of orthologous gene pairs can be difficult because of the amount of duplicates that are closely related. In our study, we showed lower rates of divergence within the UTR region in comparison to the synonymous sites of these two species. If we constructed an alignment of UTRs from a pair of paralogs in which the paralogs arose due to the FSGD, then there would be an expected high degree of divergence as opposed to the divergence rate of true orthologs. However, exceptions may occur with recent gene duplications and/or concerted evolution permits for paralogs to be subjected to similar selective pressures. If we can detect novel genes within this genus we can gain a better perspective of the rate of divergence occurring within the UTR region. Understanding the importance and evolutionary patterns of novel genes is a promising avenue with the advent of next-generation sequencing.

3.4.4 Conclusion

This transcriptomic study between *S. goodei* and *S. saxicola* provides a template for understanding evolutionary processes at the molecular level within *Sebastes*. We identified a series of candidate genes that are resourceful for the assessment of the critical genes that diverged and are responsible for the radiation within this group. Genes that pertain to longevity hold potential for understanding the molecular mechanisms that have contributed to the radiation within this genus. The establishment of genes under positive

selection from this study can be insightful and utilized to assess whether these positively selected genes are under selection across the entire genus *Sebastes*. If these genes are under positive selection across the entire genus, this will provide new clues about how natural selection is contributing to speciation by reproductive isolation within this group. This study was intended to further advance the field of evolutionary biology by providing support of which functional genes are important for adaptation and sexual selection. With transcriptomic data from multiple species within *Sebastes*, we can identify the repeated patterns of adaptive evolution and elucidate our understanding of how adaptation and the speciation processes occurred across the entire genus of *Sebastes*.

3.5 Acknowledgments

We would like to thank D. Ardell for assistance with PERL scripts. We thank Eddy Rubin, the DoE Joint Genome Institute Director, for providing the sequencing support needed to perform this work in the context of a course co-taught with JGI scientists at UC Merced. This work was supported by a grant from the NSF (DEB-0719475) to A.A., a UC Merced GRC fellowship to J.H. We would also like to thank M. N Dawson, M. Medina, and D. Ardell for providing constructive comments on the scientific merit of this manuscript. A. Winek and D. Elizondo who reviewed this manuscript for grammatical errors and clarity of concepts. Lastly, J. Liberto for assistance in organizing data files prior to analyses.

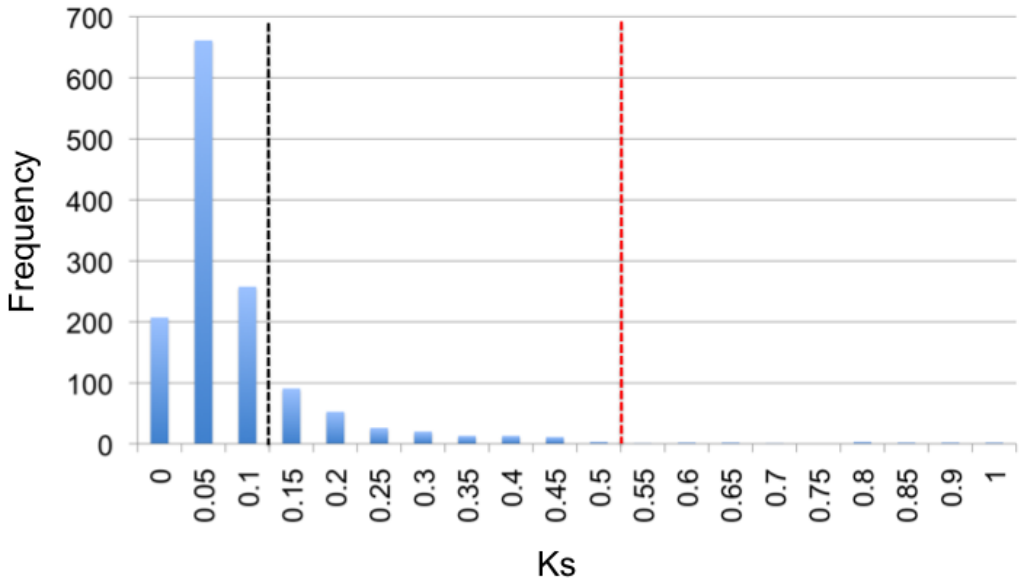

Figure 3.1: Frequency of ortholog pairs with synonymous substitution estimates. The black dotted line indicates the traditional cut off line and the red dotted line indicates our new threshold cut-off.

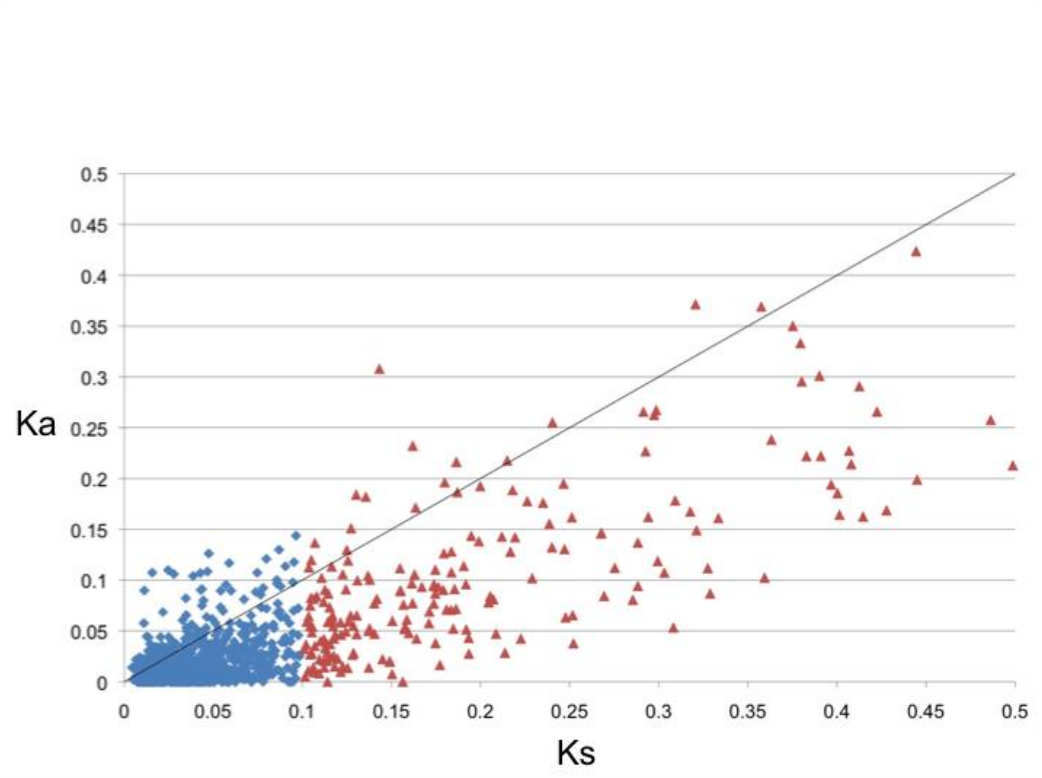

Figure 3.2: Plot of (K_a) nonsynonymous vs. (K_s) synonymous substitutions. Blue diamonds indicate values with a Ks < 0.1, whereas red triangles indicate Ks values greater than 0.1 but less than 0.5. The black line suggests neutrality, values above the line are subject to positive selection and values below are subject to purifying selection.

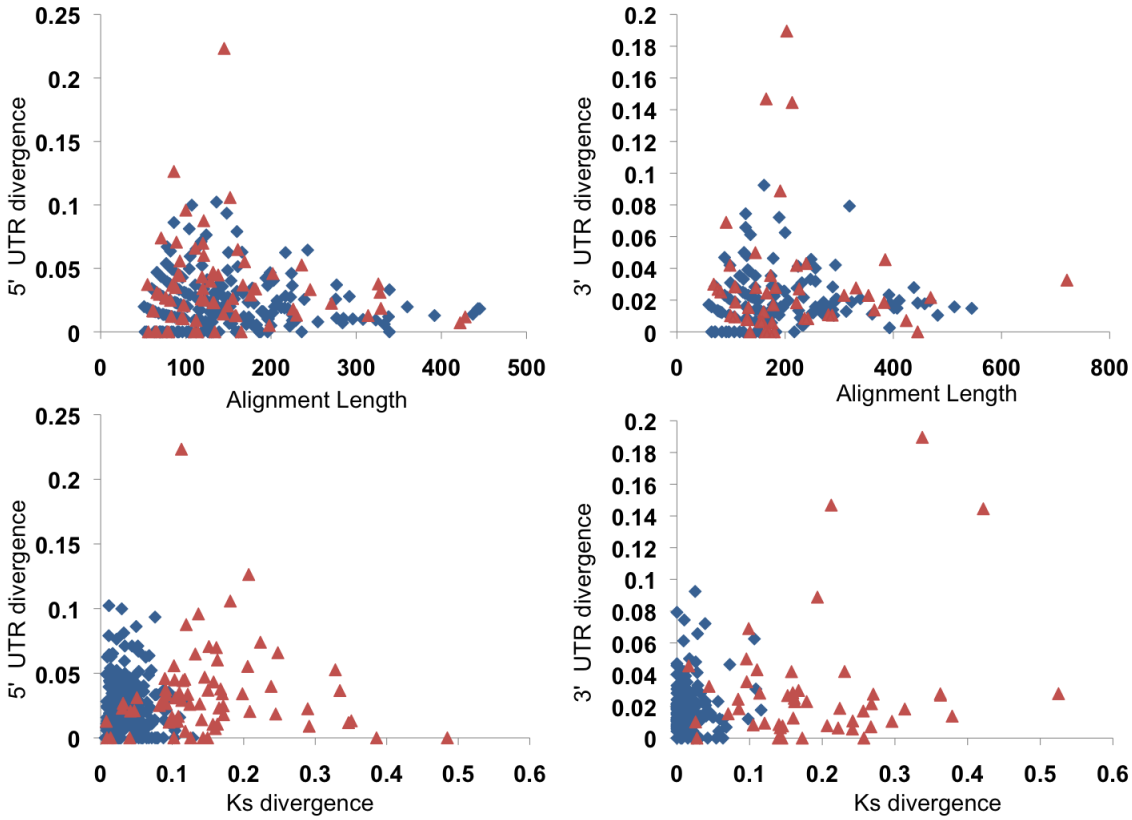

Figure 3.3: Comparison of UTR divergence with alignment length and K_s divergence. Blue diamonds indicate ortholog pairs with a $K_s > 0.1$, whereas red triangles indicate K_s values that are greater than 0.1 and less than 0.5.

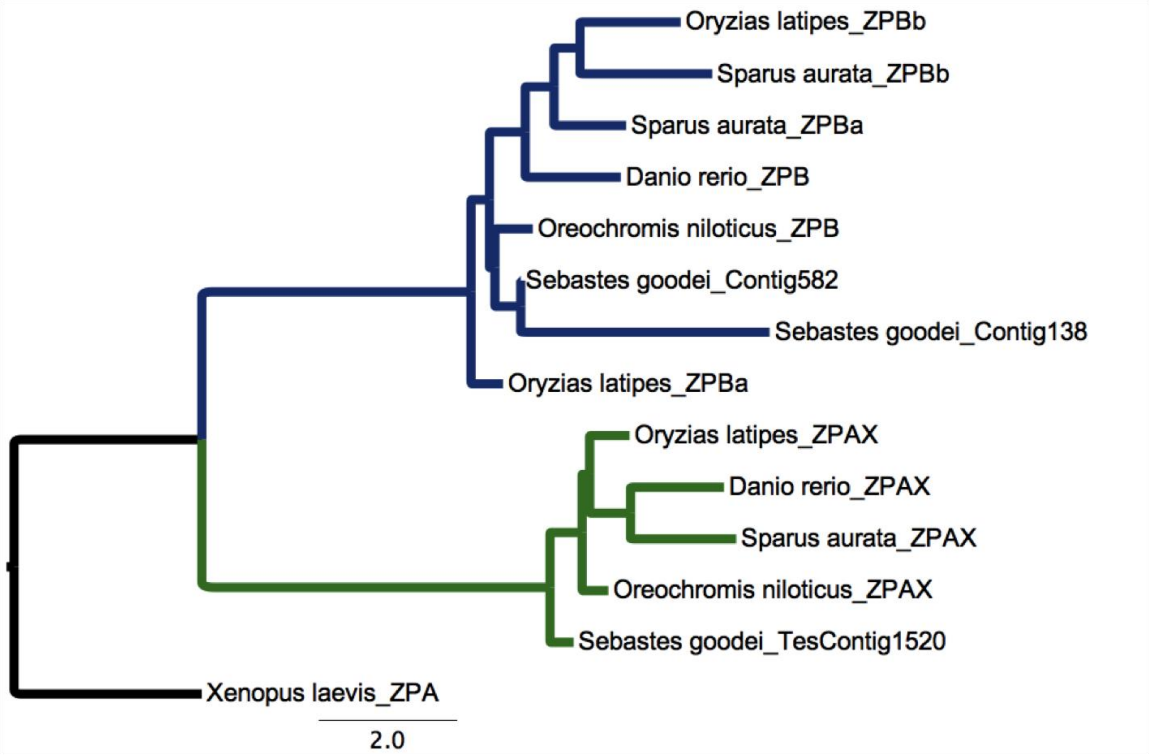

Figure 3.4: ML tree generated for ZPAX and ZPB genes found within *S. goodei* and *S. saxicola* with 1,000 bootstrap replicates. Additional teleost species were used to construct this phylogeny, and bootstrap values greater than 70 are displayed.

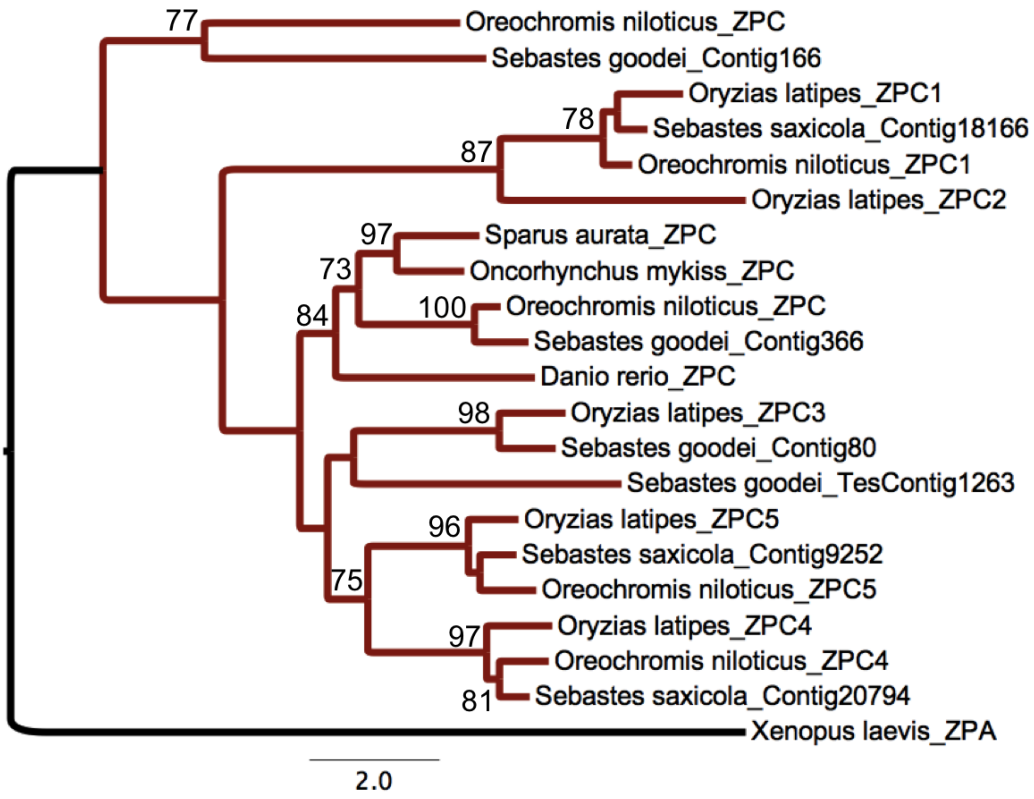

Figure 3.5: ML tree generated for ZPC genes found within *S. goodei* and *S. saxicola* with 1,000 bootstrap replicates. Additional teleost species were used to construct this phylogeny, and bootstrap values greater than 70 are displayed.

Table 3.1: EST assembly statistics for *S. goodei*

Library	*Raw Sequences*	*High Quality ESTs*	*Mean Length (bp)*	*STD EST Length*	*Singletons*	*Contigs*	*Uniseqs*	*Redundancy (%)*	*ESTs Used For Analyses*
***S. goodei* ovary**	2,370	1,814	655.9	121.5	664	630	1,294	29	combined 5,336
***S. goodei* testes**	13,824	11,657	614	118.4	2,849	1,996	4,845	58	combined 5,336

Table 3.2: *S. saxicola* Assembly Summary Statistics

	Primary Assembly (MIRA)	*Secondary Assembly (MIRA + CAP3)*	*> 300 bp*
Singletons	85,431	41,174	
Primary Contigs	51,310	14,090	3,112
Secondary Contigs	N/A	23,475	15,393
Reads Assembled into contigs	209,683	113,266	-
Total Contigs	51,310	37,565	18,505
Mean Length Primary (bp)	257	241	378
Mean Length secondary contigs (bp)	-	406	505
Total Mean Length (bp)	257	344	484
Range	41-1,046	40-2,751	300-2,751

Table 3.3: *S. goodei* and *S. saxicola* ortholog pairs that were identified as positive selection

Annotation	**K_a**	**K_s**	**K_a/K_s**	**Length**	***S. goodei* Hit ACC**	***S. goodei* E-value**	***S. saxicola* Hit ACC**	***S. saxicola* E-value**
12 kda fk506-binding protein	0.058	0.011	5.262	327	P48375	4.55E-36	P48375	2.16E-38
40s ribosomal protein x isoform	0.014	0.004	3.209	594	Q642H9	7.20E-144	N/A	N/A
60s ribosomal protein l17	0.232	0.162	1.434	153	P18621	9.36E-62	N/A	N/A
60s ribosomal protein l9	0.019	0.011	1.735	300	Q90YW0	2.53E-92	Q90YW0	6.04E-44
atp synthase mitochondrial f1 complex assembly factor 1 flags: precusor	0.032	0.020	1.583	210	Q1L987	2.27E-33	Q1L987	3.14E-81
bone morphogenetic protein 7 flags: precursor	0.011	0.010	1.124	276	P23359	1.60E-30	P23359	4.32E-45

chitobiosyldiphosphodolichol beta-mannosyltransferase	0.021	0.019	1.104	249	Q9BT22	1.46E-52	Q9BT22	5.16E-29
choline transporter-like protein 4 solute carrier family 44 member	0.029	0.021	1.368	372	Q7T2B0	3.20E-60	Q7T2B0	8.71E-80
cytochrome c oxidase subunit mitochondrial flags: precursor	0.010	0.009	1.070	423	P00426	3.68E-55	B0VYX4	5.39E-56
cytochrome p450 26a1	0.117	0.059	1.979	168	P79739	8.97E-18	P79739	9.14E-29
disintegrin and metalloproteinase domain-containing protein 9 flags: precursor	0.095	0.057	1.666	378	Q61072	2.15E-18	Q61072	1.42E-08
dna ligase 3	0.040	0.027	1.480	345	P49916	1.25E-39	N/A	N/A
dna mismatch repair protein mlh1	0.018	0.017	1.052	291	P40692	1.03E-79	P40692	3.94E-35

dna primase large subunit	**0.126**	**0.048**	**2.651**	**354**	**O89044**	**1.47E-57**	**O89044**	**8.29E-25**
double-strand-break repair protein rad21 homolog	0.130	0.087	1.493	249	O93310	4.30E-13	N/A	N/A
eukaryotic translation initiation factor 2 subunit 3	0.098	0.095	1.031	429	Q2KHU8	1.72E-110	Q2KHU8	1.35E-59
f-box only protein 11	0.081	0.067	1.204	288	Q86XK2	1.97E-49	Q86XK2	2.69E-23
f-box only protein 43 endogenous meiotic inhibitor 2	0.031	0.022	1.426	753	Q4G163	1.06E-23	Q8AXF4	2.07E-08
glioma tumor suppressor candidate region gene 2 protein	0.017	0.008	2.081	351	Q9NZM5	8.21E-33	Q9NZM5	6.12E-28
growth factor receptor-bound protein 10	0.113	0.104	1.085	216	Q13322	2.80E-47	N/A	N/A

gtpase mitochondrial	0.130	0.125	1.037	297	B5X2B8	2.46E-27	B5X2B8	1.21E-07
guanine nucleotide-binding protein g subunit alpha-2	0.090	0.060	1.491	333	P04897	3.39E-82	P04897	2.16E-43
h-2 class i histocompatibility q10 alpha chain flags: precursor	0.104	0.039	2.681	522	P01898	8.97E-44	P15979	2.66E-32
histidine triad nucleotide-binding protein 3	0.094	0.088	1.069	414	Q28BZ2	1.82E-39	Q28BZ2	7.95E-34
homolog subfamily a member 4 flags: precursor	0.037	0.029	1.289	231	Q8WW22	2.56E-49	Q8WW22	2.16E-25
importin subunit alpha-1	0.114	0.091	1.254	789	P52170	1.64E-93	P52170	2.67E-72
inositol-3-phosphate synthase 1-a	0.053	0.046	1.152	465	Q7ZXY0	3.82E-41	Q7ZXY0	5.38E-37

kelch domain-containing protein 1	0.022	0.011	2.034	318	Q8N7A1	3.47E-34	Q8N7A1	1.02E-18
lag1 longevity assurance homolog 2	0.022	0.014	1.623	258	Q96G23	2.83E-59	Q3ZBF8	6.32E-42
lamina-associated polypeptide isoform beta	0.027	0.015	1.803	432	Q62733	4.90E-10	Q62733	9.94E-09
lipid phosphate phosphohydrolase 3	0.090	0.043	2.096	201	Q3SZE3	3.77E-25	Q3SZE3	1.14E-50
map3k12-binding inhibitory protein 1	0.009	0.008	1.081	450	Q99LQ1	1.89E-38	N/A	N/A
mif4g domain-containing protein a	0.036	0.034	1.034	174	B0UXU6	2.02E-28	B0UXU6	3.33E-18
n-acetylneuraminate lyase	0.045	0.038	1.184	477	Q5RDY1	3.28E-54	Q6NYR8	8.41E-72

nad-dependent deacetylase sirtuin-5 flags: precursor	0.029	0.027	1.111	354	Q8K2C6	4.35E-65	Q3ZBQ0	2.41E-32
nuclear pore complex protein nup54	0.075	0.044	1.714	492	P70582	3.74E-33	N/A	N/A
p43 5s rna-binding protein	0.056	0.038	1.485	249	P25066	9.42E-14	P25066	4.10E-08
pentatricopeptide repeat-containing protein 2	0.012	0.011	1.073	585	Q566X6	1.63E-44	Q566X6	7.01E-97
peptidyl-prolyl cis-trans isomerase-like 2	0.018	0.008	2.303	471	Q13356	4.11E-74	Q13356	2.29E-71
poly-specific ribonuclease parn	0.019	0.016	1.200	375	O95453	4.25E-71	O95453	1.25E-61
pq-loop repeat-containing protein 2	0.029	0.018	1.605	402	Q8C4N4	2.03E-42	Q8C4N4	4.36E-41

proteasome subunit alpha type-2	0.045	0.013	3.550	423	O73672	8.15E-122	O73672	1.73E-64
protection of telomeres protein 1	0.022	0.007	3.335	528	Q9NUX5	4.88E-20	Q9NUX5	1.89E-19
protein b4	0.027	0.027	1.001	492	P15308	2.41E-12	P15308	1.47E-13
protein cwc15 homolog	0.012	0.010	1.193	516	Q6IQU4	2.95E-27	Q6IQU4	1.63E-18
protein lin-9 homolog	0.031	0.018	1.706	372	Q5RHQ8	1.12E-79	Q5RHQ8	2.81E-36
protein lsm14 homolog b	0.029	0.014	2.063	501	Q566L7	1.29E-46	Q566L7	3.57E-34
protein serac1	0.050	0.035	1.403	411	Q5SNQ7	4.94E-60	Q5SNQ7	1.60E-44

ras-related protein rab-11a flags: precursor	0.106	0.028	3.786	246	Q5ZJN2	1.06E-65	Q5ZJN2	5.37E-25
selenoprotein t1a flags: precursor	0.033	0.016	2.056	288	Q802F2	1.95E-80	Q802F2	1.06E-35
synaptotagmin-like protein 2 exophilin-4	0.064	0.046	1.374	246	Q99N50	5.69E-34	Q99N50	1.06E-19
tfiia-alpha and beta-like factor	0.034	0.023	1.444	339	Q9UNN4	1.80E-32	Q9UNN4	2.87E-23
tho complex subunit 1	0.033	0.027	1.210	291	Q96FV9	2.55E-68	Q96FV9	1.49E-42
torsin-1b	**0.090**	**0.011**	**8.023**	**258**	**O14657**	**4.26E-38**	**O14657**	**7.91E-07**
transcription initiation factor tfiid subunit 12	0.013	0.010	1.259	501	Q3T174	3.39E-61	Q3T174	1.90E-29

translin	0.008	0.008	1.068	486	Q62348	2.08E-78	Q62348	9.20E-58
transmembrane protein 106b	0.019	0.011	1.817	360	Q1LWC2	1.23E-18	Q1LWC2	1.28E-25
transmembrane protein 50a	0.053	0.035	1.518	279	O95807	1.58E-65	O95807	3.27E-37
trna guanosine-2 -o-methyltransferase trm11 homolog	0.036	0.021	1.704	285	Q05B63	2.25E-57	Q7TNK6	7.43E-34
tumor necrosis factor ligand superfamily member 10	0.151	0.127	1.186	309	P50591	2.46E-15	P50591	1.43E-08
tumor necrosis factor receptor superfamily member	0.120	0.105	1.140	309	Q92956	2.24E-28	Q92956	6.81E-13
ubiquilin-4	0.042	0.034	1.234	384	Q99NB8	8.99E-31	Q5R684	3.98E-27

ubiquitin fusion degradation protein 1 homolog	0.011	0.010	1.050	378	Q9ES53	2.56E-98	Q9ES53	3.51E-78
ubiquitin-conjugating enzyme e2 n	0.182	0.136	1.342	171	Q9EQX9	8.30E-36	Q9EQX9	3.51E-07
ubiquitin-like modifier-activating enzyme atg7	0.031	0.013	2.269	276	O95352	4.22E-63	Q5ZKY2	6.44E-33
upf0420 protein c16orf58	0.034	0.029	1.188	435	Q96GQ5	3.11E-40	Q499P8	5.73E-29
vacuolar protein sorting-associated protein 16 homolog	0.016	0.012	1.312	477	Q5E9L7	2.96E-110	Q5E9L7	1.89E-91
wd repeat-containing protein 5	0.121	0.080	1.513	288	Q2KIG2	2.23E-27	Q2KIG2	3.23E-45
zinc finger cchc domain-containing protein 4	0.044	0.020	2.155	450	Q66IH9	2.42E-67	Q6DCD7	1.19E-33

zinc finger hit domain-containing protein 3	0.010	0.009	1.194	408	Q9CQK1	4.83E-24	Q15649	8.10E-24
zona pellucida sperm-binding protein 4 flags: precursor	0.048	0.042	1.136	402	Q12836	1.35E-22	Q12836	1.78E-18

*Bold face indicates a significant Fisher's exact test (p-value <0.05)

**This table includes orthologs with a Ks 0.1 and 0.5 cut-off.

Table 3.4: PAML Analyses of Candidate and ZP genes with M7 & M8 Models

Gene ID	K_a/K_s	EST Length	M7 vs. M8	Sites Under Selection
fkb12	1.342	201	14.733	22 (0.997**), 45 (0.952*), 48(0.971*), 53 (0.997**), and 67 (0.992**)
r19	1.432	300	Not-Significant	N/A
taf12	0.372	231	Not-Significant	N/A
tm50a	2.163	279	20.773	90 (0.996*), 91 (0.977*), 92 (0.977*), and 93 (0.958*)
cox5a	0.040	297	Not-Significant	N/A
cp058	0.252	294	Not-Significant	N/A
cwc15	0.067	336	Not-Significant	N/A
if2g	0.116	420	Not-Significant	N/A
ino1a	0.142	456	Not-Significant	N/A
ls14b	3.620	396	Not-Significant	N/A
pri2	0.376	333	Not-Significant	N/A
sirt5	0.174	297	Not-Significant	N/A
tm50a	0.214	231	Not-Significant	50 (0.994**)
tsn	0.129	237	Not-Significant	N/A
znhi3	0.222	282	Not-Significant	N/A

zpax	0.295	477	Not-Significant	N/A
zpb	0.248	663	Not-Significant	N/A
zpc1	0.368	567	Not-Significant	N/A
zpc4	0.437	315	Not-Significant	N/A
zpc5	0.275	483	Not-Significant	N/A

Note: K_a/K_s values were averaged between models M7 & M8. M7 & M8 were compared for the likelihood ratio test and significant Likelihood Ratio Test values are listed. Sites that were found under positive selection are presented with only the Bayes Empirical Bayes (BEB) analyses. Blue text indicates a PAML analyses with four rockfish species (*S. goodei*, *S. saxicola*, *S. caurinus*, and *S. rastrelliger*). Red text indicates a PAML analyses with *S. goodei*, *S. saxicola*, *Oryzias latipes*, and *Oreochromis niloticus* with candidate genes under positive selection. Green text indicates a PAML analyses with *S. goodei*, *S. saxicola*, *Oryzias latipes*, and *Oreochromis niloticus* with ZP genes.

Table 3.5: Pairwise K_a/K_s estimates for ZP ortholog pairs

ZP ID	*S. goodei* EST ID	*S. saxicola* EST ID	Method	K_a	K_s	K_a/K_s	Nuc. Length
ZPAX	TesSgooContig1520	Contig7124	YN	0.008	0.028	0.304	468
ZPB	SgooContig582	Contig2319	YN	0.007	0.015	0.435	663
ZPB homolog 1	**TesSgooContig1769**	**Contig9672**	**YN**	**0.048**	**0.042**	**1.136**	**402**
ZPB homolog 2	SgooContig184	Contig10146	YN	0.003	0.030	0.110	402
ZPC homolog 1	SgooContig366	Saxicola_C47406	YN	0.011	0.061	0.186	342
ZPC homolog 2	SgooContig166	Contig6798	YN	0.009	0.034	0.274	957
ZPC1	SgooContig100	Contig18166	YN	0.002	0.013	0.187	558
ZPC3	SgooContig80	Contig9633	YN	0.005	0.021	0.253	525
ZPC4	SgooContig309	Contig20794	YN	0.005	0.012	0.381	300
ZPC5	SgooContig179	Contig9252	YN	0.006	0.026	0.223	471

* Bold face indicates an ortholog pair which is found under positive selection

Table 3.6: Pairwise Analyses of Sequence Divergence

Analysis	T test P-value	Wilcoxon Rank sum test P-value
Ks 3prime vs. UTR 3prime	***9.25E-14***	***< 2.2E-16***
Ks 3prime vs. UTR 5prime	***8.48E-13***	***< 2.2E-16***
Ks 3prime vs. Ks 5prime	0.104	0.505
UTR 5prime vs. UTR 3prime	0.207	0.145
Ks 5prime vs. UTR 3prime	***9.27E-20***	***< 2.2E-16***
Ks 5prime vs. UTR 5prime	***2.57E-18***	***< 2.2E-16***

* Bold face indicates a significant P-value

Chapter 4

Analysis of multiple transcriptomes to identify adaptive evolution in rockfishes (*Sebastes*) subgenus *Pteropodus*

Joseph Heras[1] and Andres Aguilar[2]

1- School of Natural Sciences and Graduate Group in Quantitative and Systems Biology, University of California Merced, 5200 N. Lake Rd., Merced, CA USA 95344

2- Department of Biological Sciences, California State University, 5151 State University Drive, Los Angeles, CA USA 90032

* - Author for correspondence: Joseph Heras, School of Natural Sciences, University of California, Merced, Merced, USA, (209) 228-4508, jheras@ucmerced.edu

Key Words: Next-generation sequencing, rockfishes, functionalization, adaptation, and speciation

Abstract

Our understanding of molecular evolution has greatly increased in the genomic era. The transcriptome provides insight about gene expression within developmental stages of an organism and the opportunity to identify the genetic mechanisms that contribute to adaptation and speciation. We selected a speciose group of marine rockfishes (genus: *Sebastes*) to study evolutionary patterns of five brain transcriptomes from *S. carnatus*, *S. nebulosus*, *S. maliger*, *S. mystinus*, and *S. serranoides*. *De novo* assemblies from these five transcriptomes were used to identify 3,867 orthologous clusters and 866 genes under positive selection based on Likelihood Ratio Tests (LRTs). When comparing datasets from two additional species (*S. caurinus* and *S. rastrelliger),* there were 36 orthologous clusters identified with seven genes found under positive selection. We did not identify the same genes under positive selection as identified in previous adaptive evolution studies on marine rockfishes. Genes under positive selection belonged to a variety of gene functions that included sensory perception, growth, and metabolism. In addition we identified 10 sequences under positive selection that were part of the phosphatidylinositol signaling system pathway. The use of next generation sequencing technology has allowed us to increase the amount of information about the rockfish genome and elucidate patterns of adaptive evolution.

4.1 Introduction

The advent of next-generation sequencing technology has provided a vast amount of genomic information that is being used to advance our understanding of the molecular evolutionary processes that operate within living organisms. As genomic information is being revealed from a multitude of species, this provides the opportunity for evolutionary patterns to be elucidated within coding and non-coding regions, and methylation and regulatory sites of genes within an organism (Alexander et al. 2010; Laird, 2010; and Grabherr et al. 2011). Transcriptome sequencing (RNA-Seq) offers the opportunity to understand which genes are expressed within a given developmental stage of an organism and to obtain protein coding sequence information from organisms (Yang and Smith, 2013). This information can provide a better understanding of which genes are under selective pressures and can provide a clearer image of how this can contribute to the processes of speciation and adaptation. Traditionally, the identification of positive selection, an indicator of the occurrence of adaptive evolution, has been estimated through an elevated nonsynonymous (K_a) to synonymous (K_s) substitution rate ($\omega > 1$). This assessment of selection gives an indication of whether positive or purifying selection is operating within the coding region of genes. In conjunction with RNA-Seq data and the advancement of computational processing power, a large amount of genetic information can be readily available for non-model organisms and the estimation of positive selection can be assessed throughout the genome (Wang et al. 2009). Aside from estimating ω ratios, novel methods for identifying signatures of positive selection can be determined through the identification of parallel amino-acid substitutions, the increase in the accumulation of indels, accelerated rate of gene loss, and elevated gene expression noise (Zhang, 2010). In addition, novel genes have been considered to undergo adaptive evolution based on a study of 12 closely related *Drosophila* species (Chen et al. 2010). These novel genes are known to be essential and are subject to either relaxed selection or positive selection (Chen et al. 2010). As more genomic content is available and compared across extended taxa, this can provide insight on the function of novel genes and how they can contribute to adaptation and speciation. With multiple methods to make stronger inferences about genes under positive selection, this information can be used to depict a broader view of natural selection within species that permit them to thrive within a set of environmental conditions.

The fish specific genome duplication (FSGD) has been estimated to have occurred around ~350 million years ago (Meyer and Van de Peer, 2005). With a duplicated genome, this provides the opportunity for novel gene functions to occur via neofunctionlization or subfunctionalization (Wolfe, 2001). With novel functions, these duplicated genes can also be favorable and can become fixed within a population as a result of positive selection. There are about ~25,000 different species of teleost fishes presently known, which is more than any other vertebrate group (Meyer and Van de Peer, 2005). The diversity of this speciose group is likely due to the FSGD, which has allowed for adaptive evolution to occur within novel genes in response to a variety of novel habitats (Meyer and Van de Peer, 2005). Therefore with the presence of gene duplications, the identification of true orthologs, genes found in different species that evolved from a common ancestral gene due to speciation, across a suite of species is vital for any inferences to be made about the function of the genes. If paralogous genes, genes

that are related due to a duplication event, are being compared and assumed to be orthologous across different species can lead to misleading conclusions about function and mutation rates. In addition to ortholog identification, the assessment of nonsynonymous and synonymous rates across taxa provides insight on the type of selection operating within a given gene. Although there are many facets that contribute to adaptation and speciation, some of the basic understanding of the evolutionary patterns within organisms can be assessed with access to genomic information such as the identification of positive selection within the transcriptome and the identification of orthologous and paralogous genes.

Sebastes is considered a prime system for marine adaptive radiations. There are about 105 different species within the genus *Sebastes* (Hyde and Vetter, 2007). Species within this genus occupy a variety of different habitats and depths, which are predominantly found in the Northeast Pacific. Most species are distributed from Baja California to the Gulf of Alaska (Love et al. 2002). There are 22 subgenera recognized within *Sebastes* (Kendall, 2000), which includes *Pteropodus* (Eigenmann and Beeson, 1893), the subgenus of interest in this study. Aside from being a diverse group of marine fishes, the genus *Sebastes* is composed of species important to fisheries and commercial aquaculture (Love et al. 2002). All members of *Pteropodus* reside in the Northeast Pacific and contain common characteristics such as mottled color patterns and strong head spines (Li et al. 2006). In addition, this subgenus generally occupies nearshore and shallow shelf habitats and is found sympatric among each other along most of the coast of California (Li et al. 2006). Based on geological and genetic data, *Pteropodus* has been suggested to have evolved around 3 million years ago whereas the genus *Sebastes* arose close to 8 million years ago (Hyde and Vetter, 2007). Current phylogenies constructed for species within *Sebastes* (Hyde and Vetter, 2007; Li et al. 2006) show that *Pteropodus* contains the following species: *S. atrovirens* (kelp), *S. auriculatus* (brown), *S. caurinus* (copper), *S. chrysomelas* (black and yellow), *S. carnatus* (gopher), *S. dalli* (calico), *S. maliger* (quillback), *S. nebulosus* (China), and *S. rastrelliger* (grass). This subgenus (*Pteropodus*) was selected for this dissertation because this group is nested within a diverse genus, and yet contains species that consistently group together in current published phylogenies (Li et al. 2006; Hyde and Vetter, 2007). Hyde and Vetter (2007) have suggested that future analyses of species flocks within rockfish species should focus on monophyletic and geographically constrained clades (subgenera) such as *Acutomentum, Pteropodus, Sebastocles,* or *Sebastomus,* because of the invariable speciation rates found within this genus. In congruence with Hyde and Vetter's (2007) suggestions, this study will provide insight on the evolution within this species-diverse genus *Sebastes* by focusing on the subgenus *Pteropodus.*

To understand molecular evolution within *Sebastes,* we identified loci that are under adaptive evolution within selected species from *Pteropodus* and with the inclusion of two species, *S. mystinus* and *S. serranoides* from the subgenus *Sebastosomus,* which were used as an outgroup for this study. In an earlier study found genes under positive selection by scanning Expressed Sequence Tags (ESTs) between *S. caurinus* and *S. rastrelliger* species within *Pteropodus* in the Heras et al. 2011 study. Three additional transcriptomes were used for this study (*S. carnatus*, *S. nebulosus,* and *S. maliger*), which can further validate whether genes found under positive Darwinian selection within this

subgenus. These candidate genes will provide a better understanding of how this subgenus radiated and adapted to their respective habitats within the Northeast Pacific. We used genetic sequence data generated from Illumina (San Diego, CA) sequence technology to conduct a multi-transcriptomic approach to identify loci under positive selection. The genes found with sites under positive selection were annotated to a diverse set of gene ontologies that gives a better understanding of which genes within the brain transcriptome are under positive selection. This study was developed as a source to ground truth whether the genes found under positive selection within the Heras et al. 2011 study are found under positive selection across the entire subgenus. As patterns within the subgenera of *Sebastes* are identified, we can make stronger inferences about adaptation and speciation.

4.2 Materials and Methods

4.2.1 Collecting Samples

Individuals from the species *S. carnatus* and *S. nebulosus* were collected off the coasts of Santa Cruz and Morro Bay, California respectively. *S. maliger* was collected off the coast of Newport, Oregon by W. Wagman from the Oregon Department of Fish and Wildlife. Lastly, individuals from *S. mystinus*, and *S. serranoides* within the subgenus *Sebastosomus* were collected near Monterey Bay, California. Both of the species from the subgenus *Sebastosomus* were used as an outgroup for this study (Figure 4.1, Table 4.1). All samples with the exception of *S. maliger* were collected as salvage and with hook and line methods, and with a scientific collecting permit from the California Department of Fish and Game (SC-11594).

4.2.2 Library Preparation

Brain tissue samples were preserved in RNA later (Qiagen, Valencia, CA) and stored at room temperature for 24 hours before kept in long term storage at -20°C. Total RNA was extracted from all samples by using a standard Trizol extraction protocol (Invitrogen, Carlsbad, CA). Next, the quality of the samples was checked on an Agilent 2100 Bioanalyzer with a RNA 6000 nano LabChip (Agilent, Palo Alto, CA) to assess the quality of total RNA from each sample. Messenger RNA (mRNA) was targeted with a poly-A extraction kit and was synthesized into double stranded cDNA. The cDNA libraries had a range of 250-550bp, and the insert range was 120-420bp with Illumina adapter lengths of 126 bp (both ends 2 x 63bp). Samples were multiplexed with five different index adapters developed by Illumina, San Diego, California (Table 4.2). cDNA samples were sequenced on a single lane with a HiSeq 2000 instrument (Illumina, San Diego, CA) at the Vincent J. Coates Genomics Sequence Laboratory at the University of California, Berkeley. In addition to the sequences generated for this study, expressed sequence tags (ESTs) from *S. caurinus* and *S. rastrelliger* from the Heras et al. 2011 study were also included in this study.

4.2.3 Assembly of Sequence Reads

Sequences were downloaded from the Vincent J. Coates Genomics Sequencing Laboratory FTP site. Prior to downloading, sequences were demultiplexed and divided into files based on size (a maximum of 4 million sequences were contained in each file). Due to memory limitations each file was first trimmed and assembled (*de novo*). Reads from each species were cleaned based on Phred quality scores and the adaptors were removed with TRIMMOMATIC v0.32 (Lohse et al. 2012) with the following parameters: ILLUMINACLIP:2:30:10, LEADING:3, TRAILING:30, SLIDINGWINDOW:4:15 and MINLEN:37 in order to make certain that trailing bases have a Phred score of a minimum of 30. Phred quality scores give an indication of the probability that a base was called incorrectly, where a Phred score of 30 is an indication that there is a 1 in a 1,000 chance the base was called improperly. Sequence read quality was checked with FASTQC v0.10.1 (Andrews, 2010) and afterwards, the clean and trimmed sequence reads from each file

were assembled with OASES v0.2.08 (Schulz et al. 2012) in conjunction with VELVET v1.2.08 (Zerbino and Birney, 2008) with multiple kmer sizes of19, 23, 27, and 31. For each divided sequence file, the four output files of each kmer size (19, 23, 27, and 31) were concatenated as suggested by Schutz et al. (2012) and subsequently processed with CD-HIT-EST (CD-HIT V4.5.4 package: Li and Godzik, 2006) in order to form clusters using the following parameters: -c 0.98 –n 10 –r 1 (Yang and Smith 2013). Next, all divided sequence files from each species were concatenated and assembled with CAP3 (Huang and Madan, 1999) with the following parameters: –o 200 –p 99. Samples were processed through CD-HIT-EST a second time to obtain any redundant clusters.

4.2.4 Identification of Orthologs, Estimation of Positive Selection with PAML

Assembled sequences were masked for repetitive elements with REPEATMASKER v4.0.5 (Smit et al. 2010) that included the rmblastn version 2.2.27+ and the query species was listed as teleost fish. The open reading frame was indentified with the ORFPREDICTOR web server (Min et al. 2005) and putative protein sequences were used to identify orthologous pairs through INPARANOID v.4.0 (O'Brien et al. 2005) with all pairwise comparisons from the five rockfish species (10 possible pairwise comparisons). All ortholog pairs identified through INPARANOID v.4.0 were used to identify ortholog clusters from all five species through QUICKPARANOID (http://pl.postech.ac.kr/QuickParanoid/). Following this, perl scripts were developed to obtain sets of orthologs that only contain one sequence from each species and were then aligned with the use of MUSCLE v3.7 (Edgar, 2004) and PAL2NAL 12.2 (Suyama et al. 2006) based on coding and protein sequences. Next, a perl script was generated to process multiple aligned ortholog clusters into PAML v4.7a (Yang, 1997) in order to estimate positive selection (dataset 1). We evaluated models M0 (one ω), M1a (nearly neutral), M2a (positive), M7 (neutral), and M8 (positive selection) in PAML v4.7a.

Models M1a (nearly neutral) and M2a (positive) were compared as well as models M7 (neutral) and M8 (positive selection), in which the likelihood values were used for a Likelihood Ratio Test (LRTs) to detect positive selection. The LRT values were used to compare with a χ^2 distribution. We used an α level of significance at 0.05 and 0.01 to determine when comparing two models (either M7 and 8 or M1a and M2a). A Bonferroni correction was calculated (Uitenbroek, 1997) for all LRT values and the same α levels were used as thresholds. Q-values (an analog of P-values) and the False Discovery Rate (FDR) were calculated with the q-value package (Storey, 2002) in R.

A second dataset was generated (dataset 2) that included the five species from dataset 1 and ESTs from two additional species (*S. caurinus* and *S. rastrelliger* - Heras et al. 2011). This second set was processed through INPARANOID v.4.0, which amounted to 21 pairwise comparisons. Ortholog clusters were aligned and positive selection was estimated in PAML v4.7a. The reasoning for two separate analyses is due to the difference in sequencing technology to generate these samples, where the EST datasets from *S. caurinus* and *S. rastrelliger* were generated from Sanger sequencing and also the minute number of ortholog clusters identified when using all seven species as compared to the five species datasets generated with Illumina sequencing technology (see Results). Ortholog clusters were identified from all seven species and then processed through PAML

v4.7a (Yang, 1997) in order to identify genes that are under positive selection across all species within this analysis with the same models used in dataset 1.

4.2.5 Annotation Process

Only orthologous sequences from dataset 1 were annotated with BLAST2GO V.2.7.1 (Conesa et al. 2005). *S. carnatus* was used as a reference dataset to represent the ortholog clusters identified from all five species. We used 3,867 orthologs to BLASTX against the Swissprot database with an e-value cutoff of $1e^{-10}$. Gene Ontologies (GOs) were mapped and then annotated with an e-value-hit-filter of $1.0e^{-6}$, annotation cutoff of 55, and GO weight of 5. Level 2 GO annotations were determined for the 3,876 sequences and compared to genes that were identified under positive selection (866 genes based on the LRT values at an α value of 0.05). We also loaded the Kyoto Encyclopedia of Genes and Genomes pathway maps into BLAST2GO in order to identify the interactions of the enzymes that are encoded by genes under positive selection and to determine whether changes in metabolic pathways can provide clarity on patterns of adaptive evolution.

4.2.6 Pairwise Estimation of Positive Selection

Estimation of positive selection in a pairwise fashion was made for dataset 1. In order for alignments to be comparable across all pairwise comparisons, the ends were trimmed from all the multiple alignments that were conducted with MUSCLE v3.7 with TRIMAL v1.2 (Capella-Gutiérrez, 2009), which allowed for all pairwise comparisons to be of equal length and provide unbiased estimates of K_a and K_s. There was a reduction in the comparisons made due to trimming each multiple sequence alignment and 3,863 ortholog clusters were used for the pairwise comparison. After all sequence ortholog clusters were trimmed, they were processed through the ORFPREDICTOR web server and then through PAL2NAL 12.2. Following this, positive selection was estimated with KAKS_CALCULATOR v1.2 (Zhang et al. 2006). Pairs with a K_s value greater than 0.5 were removed and/or pairs that did not contain a K_a or K_s value "na" was removed from further analyses.

4.3 Results

4.3.1 Sequence Assembly and Repeat Masker

From *S. carnatus*, *S. maliger*, *S. nebulosus*, *S. serranoides*, and *S. mystinus* the number raw reads obtained from the Illumina sequencing were 43,776,426; 43,912,259; 38,580,056; 35,315,365; and 34,489,637 respectively (Table 4.2). After sequences were trimmed and assembled, we obtained 35,245, 38,888, 31,166, 29,438; and 36,008 contigs after de novo assembling and clustering. From REPEATMASKER v4.0.5, all sequences from dataset 1 were masked for repetitive elements with the range of GC levels were from 50.97% to 51.36%. The number of Short Interspersed Elements (SINEs) ranged from 31-56 and the amount of Long Interspersed Elements (LINEs) ranged from 93 to 137. The amount of DNA transposons ranged from 508 to 730 with 2.86% to 3.01% of bases masked (Table 4.3).

4.3.2 Annotation

S. carnatus was selected as a representative of the ortholog clusters identified from QUICKPARANOID, in which 3,867 sequences were used, where 17 gene ontology categories were identified when using Biological Processes level 2. There were 12-gene ontology categories identified for molecular function and 10 categories for cellular components (Figure 4.2). Genes that were identified under positive selection from the PAML analyses based on the LRT values and our alpha cutoff of 0.05, we annotated 866 genes and identified 16 gene categories for biological processes and 8 gene categories were found for both molecular function and cellular component (Figure 4.3). The Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathways were downloaded with BLAST2GO based on genes that were found under positive selection. The highest number of sequences under positive selection within a single KEGG map was 10 sequences (8 enzymes), which also contained 12 sequences (5 enzymes) that were not under positive selection. These were found in the phosphatidylinositol signaling system pathway (Figure 4.5; Table 4.7, 4.8).

4.3.3 Ortholog Clusters Identified from INPARANOID and QUICKPARANOID

When identifying ortholog pairs from all pairwise comparisons (10 pairwise comparisons) from dataset 1, the range of identified ortholog pairs were 10,022 to 15,234 (Table 4.4). Ortholog pairs identified only between *Pteropodus* species had a range of 14,570-15,234, whereas any other ortholog pairs from a *Pteropodus* species to a *S. mystinus*/*S. serranoides* was from 10,022 to 12,004 pairs. From the identification of ortholog clusters in QUICKPARANOID, there were 3,867 and 36 orthologous clusters from dataset 1 and dataset 2, respectively.

4.3.4 Pairwise Comparisons of Dataset 1

We used 3,863 genes from dataset 1 for the pairwise comparison for the estimation of positive selection. We removed ortholog pairs with a K_s value greater than 0.5, which has been set as a benchmark selected in our previous study, (Table 4.5). In

addition, the genes found under positive selection among *Pteropodus* species (78-80 ortholog pairs), were less than any other pairwise comparison that includes *S. mystinus* and *S. serranoides* (153 or greater). The average K_s values for each pairwise comparison was around ~0.042. When ortholog pairs compared among all three *Pteropodus* species (Figure 4.4), there was a reduction in the number of genes under positive selection identified (3 genes).

4.3.5 PAML Analyses

Based on our LRT values when comparing models M1a (nearly neutral) and M2a (positive selection) we identified 851 and 703 genes under positive selection with an α significance cutoff of 0.05 and 0.01 respectively (Table 4.6). With a Bonferroni correction of 336 and 293 ortholog clusters with an α significance cutoff of 0.05 and 0.01 respectively. Lastly, the FDR was 708 and 572 with a 0.05 and 0.01 cutoff, respectively. When comparing models M7 (neutral) and M8 (positive selection) from our PAML analyses, the LRT values compared from a chi-squared distribution, we identified 866 and 708 genes under positive selection with a significance cutoff of 0.05 and 0.01 were used. In addition, from a Bonferroni correction we identified 350 and 308 gene pairs with a 0.05 and 0.01 cut off respectively. Lastly, we identified 712 and 581 (0.05 and 0.01 cut-off) with a false discovery rate (Table 4.6). When comparing all seven rockfish species, there were 36 ortholog clusters with 7 and 5 genes under positive selection with an α value of 0.05 and 0.01 respectively. When using a Bonferroni correction there were 3 genes under positive selection for both α values (0.05 and 0.01). When using the 0.05 cutoff with the χ^2 distribution, there were only four that had annotations out of the seven genes found under positive selection, which were: ribosome biogenesis protein NSA2 homolog, Selenium-binding protein 1, Alpha-N-acetylgalactosaminide alpha-2,6-sialyltransferase 6, and Mitochondrial import inner membrane translocase subunit Tim23.

4.4 Discussion

4.4.1 The Identification of Ortholog Pairs, Clusters of the Brain Transcriptome

There is a general consensus that brain tissue represents a diverse set of transcripts, as indicated in a comparative transcriptomic (brain tissue) study on 10 bird species (Künstner et al. 2010). Künstner et al. (2010) successfully identified 6,499 different genes across all 10 species, but they were only to obtain 55 genes that were sequenced across all species, which included zebra finch that was used as a reference genome. In addition, although the methodology from the Künstner et al. (2010) study was different from our study, the maximum K_s value reported was 0.47 between bird species, whereas within our study the maximum K_s value between rockfish species was estimated to be 0.047 (Table 4.5). The divergence between most avian groups is estimated to be 30 mya or greater, which provides plausible reason for the extensive K_s values found between avian species in comparison to our comparison of rockfish species with the maximum divergence estimated to be around ~6 mya (Figure 4.1). Künstner et al. (2010) demonstrated that genes found on larger chromosomes had elevated ω values which is likely due to the assumption that selection is more efficient in chromosomes with high recombination rates.

Currently, there is not a published reference genome for *Sebastes* that would provide necessary information about the location of genes and quantity of genes within the rockfish genome. Therefore, the fraction of genes expressed and identified in our study as compared to the entirety of the rockfish genome remains unknown.

We were able to identify 3,867 ortholog clusters from dataset 1 (five species). This number of ortholog clusters is considerably less than the number of orthologs identified through pairwise analysis of species (10,022 ortholog pairs and greater, Table 4.4). This gives an indication that all species may express a broad set of genes that partially overlap in each pairwise comparison or that the entire brain transcriptome was not sequenced from each species. There is an increase in the amount of orthologs identified between *Pteropodus* species as compared to any species comparison with to blue and olive rockfishes (Table 4.4). This increase is mostly likely due to how closely related these three species are in comparison to *S. mystinus* and *S. serranoides*. As more genomic information become available, determining whether genes are novel within *Pteropodus* can provide more information about adaptation and speciation. If these genes are common within *Pteropodus* and not indentified within other rockfish species, then these genes pose an interesting question of whether novel genes become quickly essential, which has been identified within *Drosophila* species (Chen et al. 2010). Chen et al. 2010 demonstrated that genes within recently diverged species of *Drosophila* species are essential for development at larval stages.

4.4.2 Genes Identified Under Positive Selection from Dataset 1: PTEN and Other Genes Under Positive Selection Associated with Longevity

Species within *Sebastes* vary in lifespans where the shortest-lived rockfish species is calico rockfish (*S. dalli*) at 12 years and the longest-lived rockfish is rougheye (*S.*

aluetianus) which have a maximum lifespan of 205 years (Love et al. 2002). The extensive lifespan difference among closely related species provides an opportunity to understand the genetic mechanisms behind longevity and how this may have contributed to the adaptive radiation within this group. Studies have shown genes that encode for proteins associated with translation impact the lifespan of an organism (Tavernarakis, 2007; Tavernarakis, 2008). From our PAML analyses, we identified an extensive number of genes under positive selection (866 gene pairs), which include eukaryotic translation initiation factor, and ribosomal subunits, which has been demonstrated to be associated with longevity within other organisms (Tavernarakis, 2008).

Although not part of the mRNA translation process, the metabolic pathway with the most identified genes under positive selection was phosphatidylinositol signaling system, which includes the PTEN gene (Figure 4.5). With genes under positive selection found in the signaling pathway is interesting because this pathway has been shown to be associated with Parkinson's disease (Yang et al. 2005). This information about the phosphatase and tensin homolog (PTEN) gene within marine rockfishes is unprecedented and there are studies that have demonstrated that PTEN regulates longevity (Solari et al. 2005; Ortega-Molina et al. 2012). The function of PTEN acts as a tumor suppressor gene in which the function is involved in the cell cycle that prevents cells from growing and dividing rapidly. This is of great interests in terms of rockfish life history, because the differences in lifespans across species within *Sebastes* even within subgenera are extensive.

Within the genus *Sebastes*, rougheye rockfish (*S. aleutianus*) has a lifespan of 205 years while the shortest-lived rockfish is 12 years (*S. dalli*; Cailliet, 2001). With genomic information available, provides an opportunity to understand the genetic mechanisms for longevity. Since there is no selection on post-reproduction longevity, which contributes to the fitness of the offspring (Rose, 1991), mortalities rates are expected to increase. However, there are enough age-independent alleles in which are beneficial at earlier stages in life and also at later stages in life, this can extend the lifespan of a population, which is described as protagonist plieotropy (de Grey, 2007).

In our study we identified genes that are associated with longevity such as ribosomal protein rl19, eukaryotic translation initiation factor 3 subunit 3, eukaryotic translation initiation factor 2 subunit 1; 39s, 40s and 60s Ribosomal subunits in which we have found other subunits under selection within our pairwise comparative studies of *S. caurinus* and *S. rastrelliger* (Heras et al. 2011) and Heras et al. unpublished study. How substitutions differ depending the age of the rockfish species and the interactions of these genes is promising to understanding how extended lifespans within *Sebastes* is a key adaptation for certain species.

The second highest amount of sequences found under positive selection within the KEGG metabolic pathways was the inositol phosphate metabolism pathway. Inositol phosphate has a multipurpose signaling pathway, which is known to be associated with cell growth, cell differentiation, and gene expression (Shen et al. 2003). Shen et al. (2003) demonstrated that mutations within genes that encode for inositol polyphosphate kinases impair transcription. Within invertebrates the regulation of longevity has been demonstrated with inositol phosphate signaling (Wolkow, 2003). It appears that there are multiple metabolic pathways that are associated with longevity and the variance that is

identified within rockfishes may be more complex than a couple of genes associated with translation. Other metabolic pathways that contained a high amount of genes under positive selection were associated with purine and pyruvate metabolism. It is not clear how genes associated with these pathways would have elevated nonsynonymous substitutions because changes within these genes would appear to be deleterious, as supported by studies where defects in purine metabolism as clinical studies (Duran et al. 1997), and pyruvate is essential in many metabolic pathways.

From the pairwise estimate of positive selection there was a large reduction in the amount of genes found under positive selection among *Pteropodus* species as compared to any other pairwise comparison (i.e. *Pteropodus* species to *S. mystinus* or *S. serranoides,* and between *S. mystinus* and *S. serranoides*; Table 4.5). The overlap of genes found under positive selection between all three *Pteropodus* species that was under positive selection was 3, where only two were annotated, h-2 class ii histocompatibility e-d alpha chain and ubiquitin-protein ligase. MHC proteins such as h-2 class ii histocompatibility are known to be under positive selection because of their association with co-evolutionary process of pathogens (Hughes, 2007). Ubiquitin ligases have also been found under positive selection due to their putative mechanisms for foreign protein degradation (Thomas et al. 2006).

4.4.3 Caveats of Genes Identified Under Positive Selection

There is a dilemma when it comes to the alignment of multiple sequences. An inaccurate alignment can still provide phylogenetic resolution, but may alter estimates of K_a and K_s greatly, because the estimates of selection are dependent on aligning homologous sites (Markova-Raina and Petrov, 2011). Markova-Raina and Petrov (2011) have demonstrated that different alignment methods generate a great increase in the amount of false positives of genes assumed to be under positive selection. This poses an issue when handling next generation sequencing datasets because of the amount of identified sequences processed through automated pipelines to make evolutionary inferences. Markova-Raina and Petrov (2011) suggested that alignments of large datasets can still provide optimal phylogenetic information, but are weary of making inferences about adaptive evolution because an incorrect alignment can give misleading results about the estimation of positive selection. Within our dataset, this information is pertinent, in which alignment of our candidate genes under positive genes with different aligners can provide stronger inference on whether these genes are under positive selection as opposed to type I error (false positives).

4.4.4 Genes Under Positive Selection from Dataset 2

When we included the two additional rockfish datasets from *S. caurinus* and *S. rastrelliger* that were generated via Sanger sequencing, we were only able to identify 36 ortholog clusters as compared to the 3,867 identified with the five-Illumina datasets, which is a great reduction in the identification of orthologs. This difference is likely supported because of the use of two different sequencing technologies, library preparation, sample collection, and different developmental stages when these species were caught. Other plausible reasons for the reductions is because of the library preparation differences or degradation total RNA in the Sanger sequencing method for *S.*

caurinus and *S. rastrelliger* sample collection which could reduce the number of transcripts identified. In addition, the genes identified under positive selection with our PAML analyses from all seven species were not the same genes identified under positive selection as identified in the Heras et al. (2011) study.

There were only 7 genes that were under positive selection across all seven species of marine rockfishes, with only four that were annotated in BLAST2GO. The nsa2_human gene, which encodes for a ribosome biogenesis protein nsa2 homolog was identified under positive selection. The function of nsa2 is involved in the biogenesis of the 60S subunit. The 60S subunits of a ribosome have been associated with the extension of the lifespan of an organism, where the interference with mRNA translation impacts longevity (Tavernakis, 2008). Longevity across the species within this study is extensive (Table 4.1), where *S. rastrelliger* lives up 23 years and *S. maliger* lives up 90 years. Genes that impact longevity are prime candidates to study across all species within *Sebastes* because the maximum age varies from each species. Another gene identified under positive selection was tim23_danre, which is the mitochondrial import inner membrane translocase subunit tim23. This gene is involved in embryonic development, mutations that occur within this gene within zebrafish are known to provide developmental defects such as smaller head, reduced tail circulation, and reduced eye and body formation. This appears to be contradictory if this gene is pertinent for development, then mutations during development would be lethal. Selenium-binding protein 1 (sbp1_danre) was another gene found under positive selection, which is involved in identifying xenobiotics within the cytoplasm. This type of gene would be strikingly interesting due to the amount of pollutants found within marine systems due to anthropogenic and non-anthropogenic means, but the differences in pollutants or toxins among different depths would need to be determined. Lastly, sia7f_human was identified under positive selection, which is Alpha-N-acetylgalactosaminide alpha-2,6-sialyltransferase which is known to be associated with invasivity of metastasis of tumors in vitro and the inhibition in vivo tumor development. Genes associated with tumor suppression have been found under positive selection (Mikkelsen et al. 2005). This would be interesting to determine whether tumor suppression and longevity is correlated within *Sebastes*.

4.4.5 How Do These Genes Contribute to Speciation and Adaptation?

There is strong support that parapatry is the likely the mechanism which allowed speciation to occur within the genus *Sebastes* (Ingram 2011; Hyde et al. 2008). Ingram 2011 suggested that parapatry is the plausible mechanism for speciation within this genus, and that allopatry would be an unlikely mechanism for speciation because potential for gene flow within marine systems. This was supported by Ingram's (2011) phylogenetic analyses where divergence is found by depth occupied by the species and depth associated-morphology that supported by parapatry models. Most patterns of speciation within marine systems would be likely due to divergent or disruptive selection, such as mechanisms of reproduction isolation (Puebla, 2009). Since species within *Pteropodus* are found in shallow shelf habitats and sympatric among each other along most of the coast of California (Li et al. 2006), reproductive isolation mechanisms and

depth preference would be the plausible reason for the diversity within this subgenus. Genes found under positive selection from the first dataset were from a broad scope of gene ontologies, but the majority of categories were associated with biological regulation (10%), cellular process (11%), developmental process (8%), metabolic process (9%), multicellular organismal process (8%), response to stimulus (8%), and single-organism process (11%; Figure 4.3). The genes that pertain to these GO categories would be worth investigating further. Genes associated with metabolic pathways provide an opportunity to understand if these genes are associated with adaptation to differences in depth and longevity. In addition, developmental processes would be interesting to determine whether all *Pteropodus* species share similar developmental patterns in comparison to the outgroup members (*S. mystinus* and *S. serranoides*). In addition, genes associated with sensory would be prime candidates for functional studies. With the diversity of coloration in rockfishes (Love et al. 2002), recognition of mates appears to play a large role in rockfish reproduction. Where Sivasundar and Palumbi 2010 have indicated that rhodopsin genes are under parallel evolution due to depth differences, which would be pertinent for recognizing mates. These genes from these categories are prime candidates for future functional studies. Based on Ingram's (2011) study that supports parapatry as the primary mode of speciation within rockfishes, identifying genes associated with depth (i.e. genes that encode for sensory, metabolism, and developmental processes) can provide a stronger support for this hypothesis.

As more genomic information becomes available, a more systems approach can be applicable to identifying patterns of adaptation and speciation. Where all genes that encode for enzymes would be mapped out and the assessment of positive selection can determined throughout all metabolic pathways. Then we can determine whether certain pathways have a higher prevalence of genes under positive selection. This information can be used to understand how these changes in these pathways impact the life history of an organism within model organisms (i.e. *Danio rerio* and *Oryzias latipes*). An additional step would be to identify differences in *cis*-regulatory elements and presence of novel genes and how these genomic mechanisms contribute to adaptation and speciation (Hughes, 2007; Chen et al. 2010).

4.4.7 Conclusions

Our objectives were to utilize species from the subgenus *Pteropodus* and *Sebastosomus* as a template for understanding evolutionary processes at the molecular level. The establishment of the genes we identified under Positive selection from the two subgenera can be insightful and utilized to assess whether these positively selected genes are under selection across the entire genus *Sebastes*. If these genes are under positive Darwinian selection across the entire genus, this will provide robust evidence about how natural selection has contributed to speciation within this group. This will advance our understanding about how natural selection, adaptive radiations, and speciation operate within marine ecosystems. In addition, our methodologies serve to advance the field of comparative genomics. With the increase of next generation sequencing, protocols to process and analyze sequence data will be crucial. With this increase of uncovering the rockfish genome, this will pave the foundation for understanding how substitutions

impact metabolic pathways and enzymes. To advance this field of identification of adaptive evolution, a clear description of how these genes operate within the genome and the environmental factors, which impact these changes.

Many species within the genus *Sebastes* are overfished, in which a few have been listed as threatened species (canary rockfish - *S. pinniger*, and yelloweye rockfish – *S. ruberrimus*) and as endangered (Boccaccio - *S. paucispinis*; Magnuson-Ford et al. 2009). The advancement in knowledge of the rockfish genome can elucidate our understanding of risk factors for genetic disorders, variation in immune function, specialization (i.e. dietary tolerances), and reproduction. These factors can provide insight about physiology and guide conservation management by means of improving wildlife health conditions and plans for intervention of population viability (Ryder, 2005). With transcriptomic data from multiple species within *Sebastes*, we obtained a suite of candidate genes under positive Darwinian selection, which can be used to assess whether these genes are responsible for the radiation and how adaptation and speciation across the entire genus of *Sebastes*.

4.5 Acknowledgements

We would like to thank N.C. Hall, J.A. Heras, A. and S. Martinez, B. Morrow, O. Ureta, V. Landa, and P. Heras for their assistance in collecting rockfish samples. Wolfe Wagman also provided samples from ODFW, Marine Resources Program. J. Liberto for assisting with bioinformatic scripts analyses. In addition, E. Gibb for providing computational assistance and review of perl scripts which were used to analyze our Illumina datasets. In addition, we would like to thank M.N. Dawson, D. Ardell, M. Medina, and M. Barlow for feedback and suggestions on obtaining next generation sequence data. In addition, we would like to thank the Quantitative and Systems Biology graduate group at the University of California, Merced for providing partial support for sequencing reagents for this study.

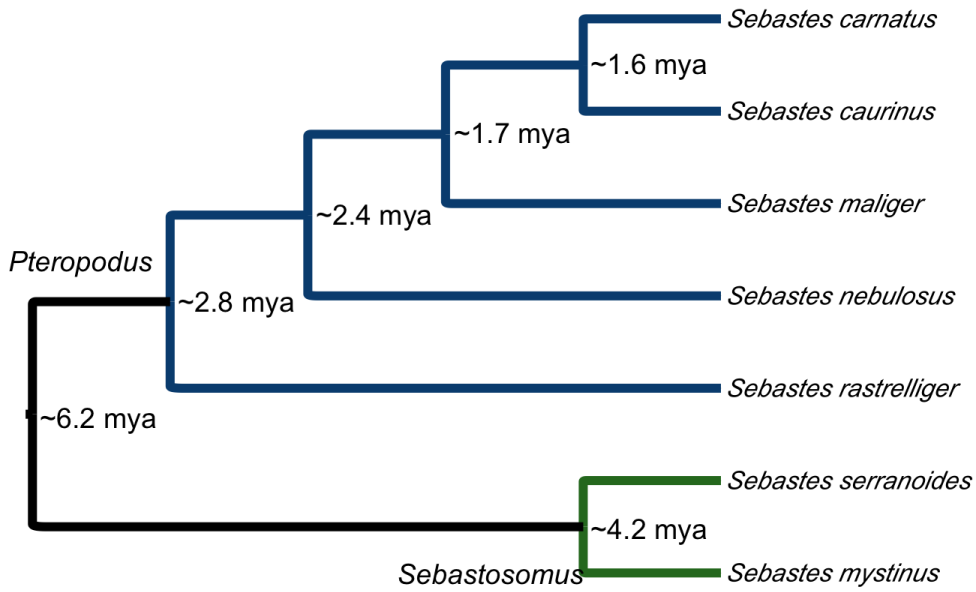

Figure 4.1: A cladogram that depicts the evolutionary relationships of the species used in this study. Located at the nodes are the estimated time since the divergence of the most recent common ancestor based on the Hyde and Vetter (2007) study. The blue clade represents species from the subgenus *Pteropodus*. The green clade represents species from the subgenus *Sebastosomus*.

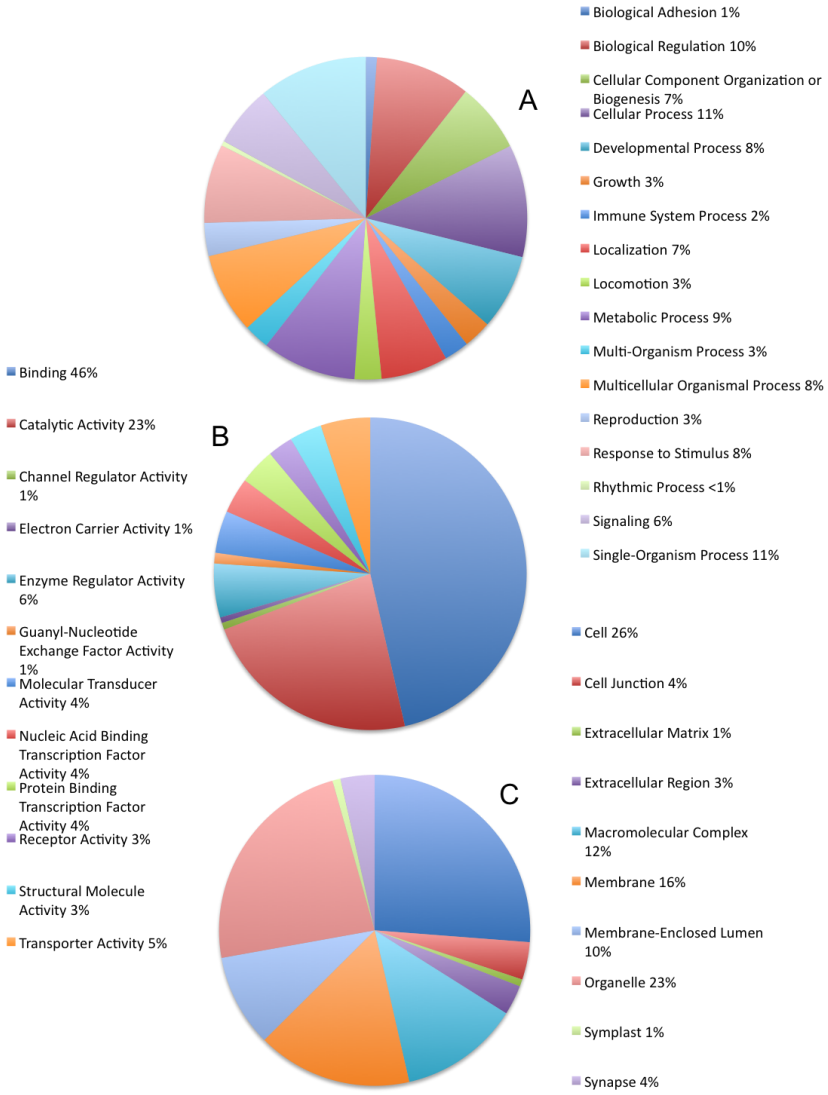

Figure 4.2: Gene Ontology Annotations. A: Level 2 Annotation for Biological Process; B: Level 2 Molecular Function; and C: Level 2 Cellular Component.

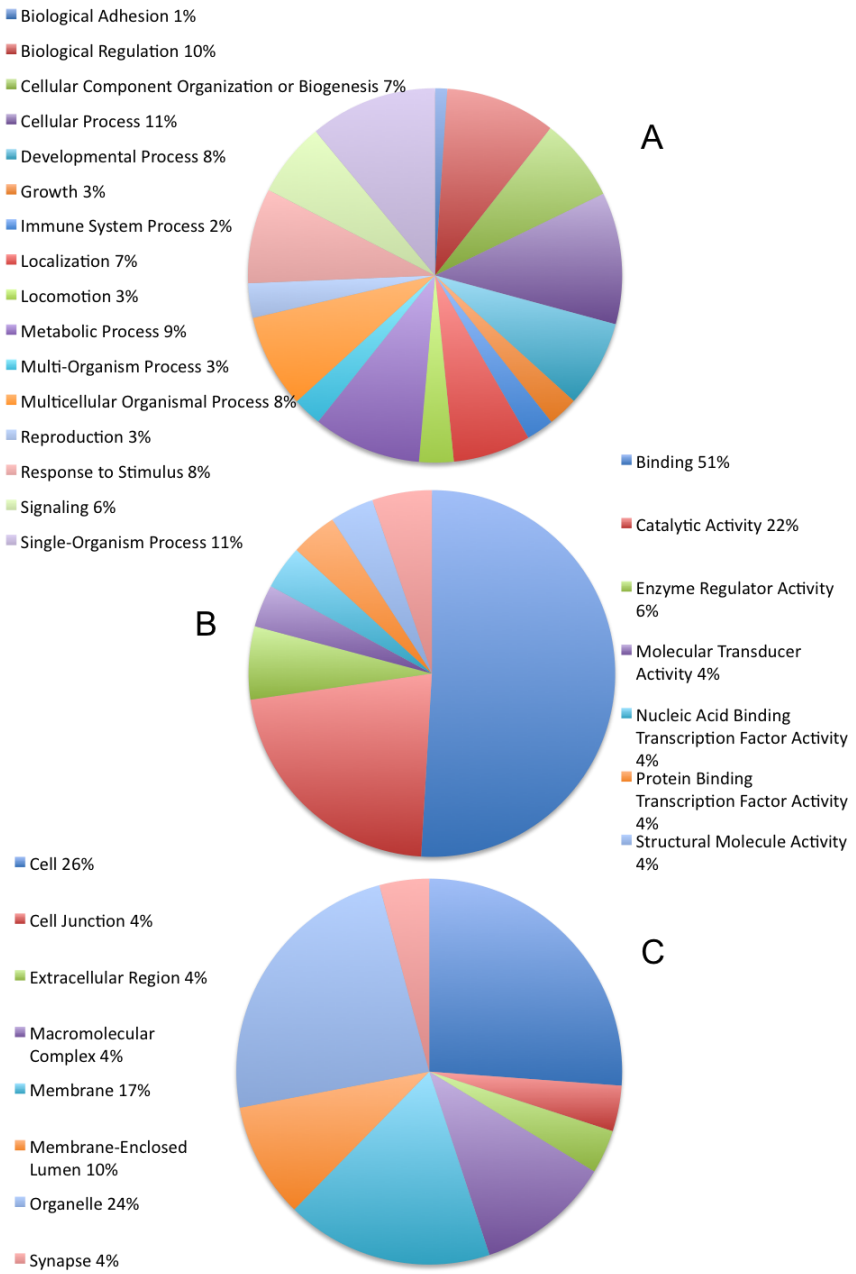

Figure 4.3: Gene Ontology Annotations for genes under positive selection. A: Level 2 Annotation for Biological Process; B: Level 2 Molecular Function; and C: Level 2 Cellular Component.

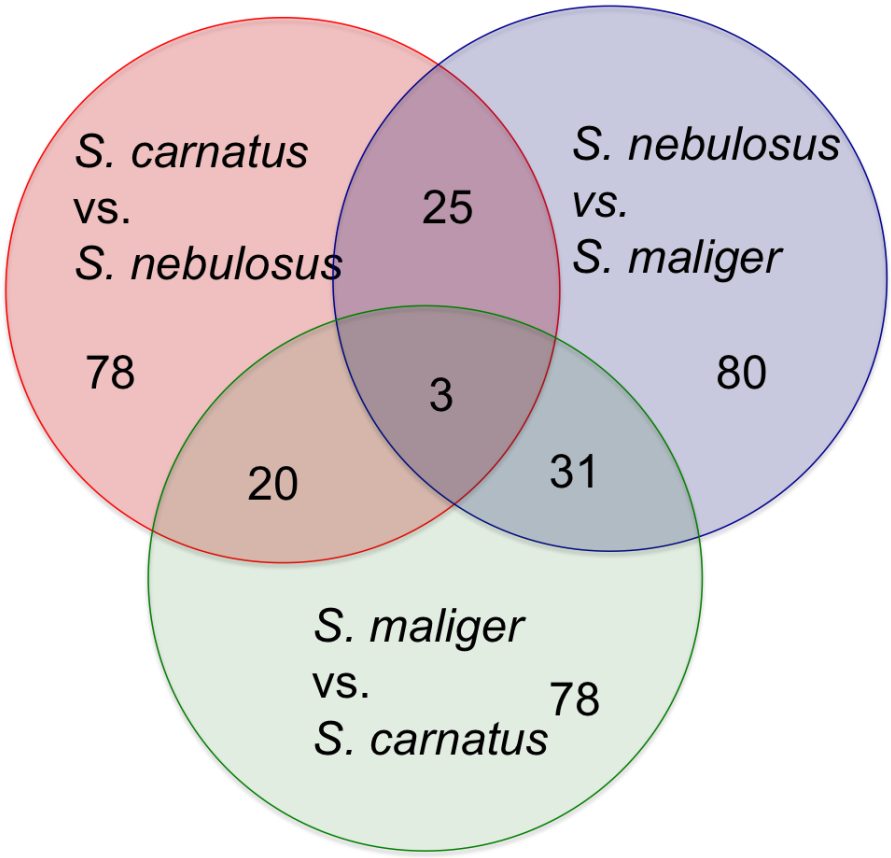

Figure 4.4: Venn diagram of the genes identified under positive selection with a pairwise comparison for species within the subgenus *Pteropodus*.

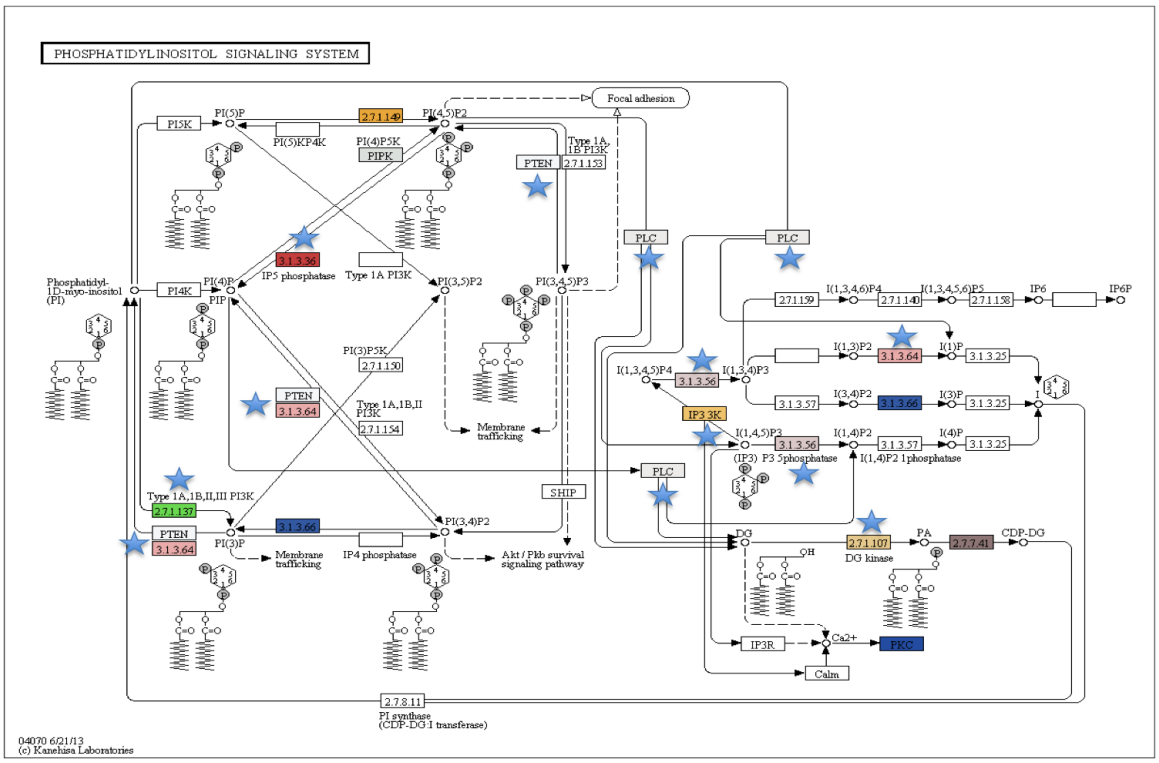

Figure 4.5: A KEGG metabolic map of Phosphatidylinositol Signaling System, colored boxes indicate a sequence identified in our analyses in BLAST2GO (Conesa et al. 2005). Stars indicate sequences that were identified under positive selection with our PAML (Yang, 1997) analysis.

Table 4.1: Description of each species in this study

Species	Maximum Age	Depth Range	Larval Time Release	Seq. Technique	Subgenus
S. carnatus	30 years	intertidal to 80 m	January to July	HiSeq 2000	*Pteropodus*
S. nebulosus	79 years	3 to 128 m	January to June	HiSeq 2000	*Pteropodus*
S. maliger	90 years	41 to 60 m	March to July*	HiSeq 2000	*Pteropodus*
S. caurinus	50 years	subtidal to 183 m	January to May*	Sanger	*Pteropodus*
S. rastrelliger	23 years	intertidal to 46 m	January to March	Sanger	*Pteropodus*
S. mystinus	44 years	~90 m	October to March	HiSeq 2000	*Sebastosomus*
S. serranoides	30 years	subtidal to 172 m	December to March	HiSeq 2000	*Sebastosomus*

*Dependent on location, the time of release of larvae varies within interval

Data taken from Love et al. 2002

Table 4.2: Illumina Multiplexed Samples

Species	Index Sequence	Raw Reads	Post Trimmomatic	Post Oases and CD-HIT-EST	CAP3	CD-HIT-EST (2nd time)
S. mystinus	index2: CGATGT	34,489,637	30,047,675	1,160,786	47,194	36,008
S. nebulosus	index4: TGACCA	38,580,056	33,910,082	977,209	38,577	31,166
S. serranoides	index5: ACAGTG	35,315,365	30,814,519	900,280	37,360	29,438
S. carnatus	index6: GCCAAT	43,776,426	38,352,769	1,170,764	44,952	35,245
S. maliger	index7: CAGATC	43,912,259	38,438,058	1,502,738	50,913	38,888

Table 4.3: Repetitive Elements Identified Within Each Dataset

	Sequences	**GC level**	**SINE**	**LINE**	**LTR elements**	**DNA transposons**	**Bases Masked**
S. carnatus	35,245	50.97%	40	137	127	640	417,263 bp (3.01%)
S. nebulosus	31,166	51.17%	40	109	120	646	378,300 bp (2.86%)
S. maliger	38,888	51.21%	56	147	151	730	463,391 bp (2.87%)
S. mystinus	36,008	51.29%	52	150	153	669	425,954 bp (3.08%)
S. serranoides	29,438	51.36%	31	93	108	508	341,533 bp (2.87%)

Table 4.4: Pairwise Identification of Orthologs with INPARANOID (O'brien et al. 2005) for the five species in this analysis

	S. mystinus	*S. nebulosus*	*S. serranoides*	*S. carnatus*	*S. maliger*
S. mystinus	-				
S. nebulosus	10,367	-			
S. serranoides	10,022	11,475	-		
S. carnatus	10,731	15,173	12,004	-	
S. maliger	11,423	14,570	11,665	15,234	-

*Blue text indicates species comparisons from the subgenus *Pteropodus*

Table 4.5: Pairwise Estimation of Positive Selection

Pairwise Comparison	Ortholog Pairs Analyzed	Removal of pairs*	K_a/K_s 0.5-1.0	# with Fisher P-value of 0.05	Ka/Ks > 1	# with Fisher P-value of 0.05	Avg. K_s values / Stdev.
S. mystinus vs. *S. nebulosus*	3,863	1646	323	3	166	6	0.0455 / 0.0624
S. mystinus vs. *S. serranoides*	3,863	1572	317	5	172	5	0.043 / 0.0591
S. mystinus vs. *S. carnatus*	3,863	1659	307	2	163	5	0.0453 / 0.0607
S. mystinus vs. *S. maliger*	3,863	1697	309	3	197	4	0.0467 / 0.065
S. nebulosus vs. S. serranoides	3,863	1596	301	4	153	3	0.0419 / 0.0582
S. nebulosus vs. S. carnatus	3,863	642	145	0	78	1	0.0377 / 0.0683
S. nebulosus vs. S. maliger	3,863	696	168	5	80	1	0.0402 / 0.0702
S. serranoides vs. S. carnatus	3,863	1640	275	2	162	2	0.0402 / 0.0545
S. serranoides vs. S. maliger	3,863	1633	280	1	186	2	0.0417 / 0.556
S. carnatus vs. S. maliger	3,863	626	151	3	78	1	0.0385 / 0.0678

*Pairs were removed if Ks value was greater than 0.5 and/or contained N/A

Table 4.6: Identification of Genes under Positive Selection with PAML analyses. The significance cutoffs are at 0.05 and 0.01 when comparing an LRT with a $\chi 2$ distribution, Bonferroni correction, and a False Discovery Rate.

	Chi-Squared		**Bonferroni Correction**		**FDR**	
Model Comparison	0.05	0.01	0.05	0.01	0.05	0.01
Model 8 vs 7	866	708	350	308	712	581
Model 2 vs 1	851	703	336	293	708	572

Table 4.7: Sequences Identified in the Phosphatidylinositol Signaling System

Enzyme ID	Sequence ID	Omega Value(s) from PAML
2.7.1.107	JH4Contig34955, JH4Contig42163, JH4Contig31417	0.305, 0.627, 0.326
2.7.11.13	JH4Contig28530	0.0001
3.1.3.67	JH4Contig24445, JH4Contig13726	0.0001, 2.0377
3.1.3.66	JH4Contig14076	0.0001
3.1.3.64	JH4Contig35178, JH4Contig24445, JH4Contig13449, JH4Contig13726, JH4Contig6793	0.0001, 0.0001, 0.4095, 2.0377, 999.000
3.1.3.56	JH4Contig39599, JH4Contig8975, JH4Contig33462	0.0254, 0.3276, 0.0971
3.1.3.36	JH4Contig39599, JH4Contig8975	0.0254, 0.3276
3.1.4.11	JH4Contig16316, JH4Contig21900, JH4Contig15995, JH4Contig11961	0.0917, 0.0327, 0.1211, 0.9616
2.7.1.149	JH4Contig31272	0.0001
2.7.1.68	JH4Contig31272	0.0001
2.7.1.137	JH4Contig37545, JH4Contig5075	0.0001, 0.116
2.7.7.41	JH4Contig34989	0.0001
2.7.1.127	JH4Contig38605	0.301

Note: Columns highlighted in blue are sequences found under positive selection

Table 4.8: KEGG Pathways from genes under Positive Selection

Pathway	Number of Seqs.	Number of Enzymes
Phosphatidylinositol Signaling System	10	8
Purine Metabolism	8	7
Inositol Phosphate Metabolism	8	8
Pyruvate Metabolism	7	6
Oxidative Phosphorylation	5	2
Citrate Cycle (TCA cycle)	5	5
Glycerophospholipid Metabolism	5	3
Glycerolipid Metabolism	5	3
Carbon Fixation Pathways in Prokaryotes	5	6

References

Aagaard JE, Vacquier VD, MacCoss MJ, Swanson WJ. 2010. ZP domain proteins in the abalone egg coat include a paralog of VERL under positive selection that binds lysin and 18-kDa sperm proteins. Mol Biol Evol. 27(1): 193-203.

Abascal F, Zardoya R, Posada D. 2005. PROTTEST: Selection of best-fit models of protein evolution. Bioinformatics. 21(9): 2104-2105.

Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB. 2010. Annotating non-coding regions of the genome. Nat Rev Genet. 11: 559-571.

Alkan C, Coe BP, Eichler, EE. 2011. Genome structural variation discovery and genotyping. Nat Rev Gen. 12(5): 363-376.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. J. Mol Biol. 215(3): 403-410.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. Nature. 437(7062): 1149-1152.

Andrés JA, et al. 2013. Patterns of transcriptome divergence in the male accessory gland of two closely related species of field crickets. Genet. 193(2): 501-513.

Andrews S. 2010. FASTQC. A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Baarends WM, van der Laan R, Grootegoed JA. 2001. DNA repair mechanisms and gametogenesis. Reproduction. 121(1): 31-39.

Bairoch A, Apweiler R. 1997. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic Acids Res. 25(1): 31-36.

Barker MS, et al. 2010. EVOPIPES.NET: Bioinformatic tools for ecological and evolutionary genomics. Evol Bioinform. 6: 143-149.

Barreto FS, Moy GW, Burton RS. 2011. Interpopulation patterns of divergence and selection across the transcriptome of the copepod *Tigriopus californicus*. Mol Ecol. 20(3): 560-572.

Berg LS. 1965. Freshwater fishes of the USSR and adjacent countries. Vol. III. 4th ed. Israel Program for Scientific Translation, Jerusalem.

Berlin S, Smith NGC. 2005. Testing for adaptive evolution of the female reproductive protein ZPC in mammals, birds and fishes reveals problems with the M7-M8 likelihood ratio test. BMC Evol Biol. 5(1): 65.

Bierne N, Bonhomme F, David P. 2003. Habitat preference and the marine-speciation paradox. Proc. R. Soc. Lond. B. 270(1522): 1399-1406.

Blanca J, Chevreux B. 2010. SFF_EXTRACT. COMAV Institute, Universidad Politécnica, Valencia, Spain.

Bluthgen N, et al. 2004. Biological profiling of gene groups utilizing gene ontology. Genome Inform. 16: 106-115.

Buonaccorsi VP. et al. 2005. Limited realized dispersal and introgressive hybridization influence genetic structure and conservation strategies for brown rockfish, *Sebastes auriculatus*. Conserv Genet. 6(5): 697-713.

Burford MO, Bernardi G. 2008. Incipient speciation within a subgenus of rockfish (*Sebastosomus*) provides evidence of recent radiations within an ancient species flock. Mar Biol. 154(4): 701-717.

Bustamante CD, et al. 2005. Natural selection on protein-coding genes in the human genome. Nature. 437(7062): 1153-1157.

Chen D, Pan KZ, Palter JE, Kapahi P. 2007. Longevity determined by developmental arrest genes in *Caenorhabditis elegans*. Aging Cell. 6(4): 525–533.

Chen S. et al. 2010. New genes in *Drosophila* quickly become essential. Science. 330(6011): 1682-1685.

Chevreux B, et al. 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res. 14(6): 1147-1159.

Chou HH, Holmes MH. 2001. DNA sequence quality trimming and vector removal. Bioinformatics. 17(12): 1093-1104.

Clark AG, et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature. 450(7167): 203–218.

Clark, NL, Findlay GD, Yi X, MacCoss MJ, and Swanson WJ. 2007. Duplication and selection on abalone sperm lysin in an allopatric population. Mol Biol Evol. 24(9): 2081-2090.

Conesa A, et al. 2005. BLAST2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 21(18): 3674-3676.

Coy JF, et al. 1995. Highly conserved 3' UTR and expression pattern of FXR1 points to a divergent gene regulation of FXR1 and FMR1. Hum. Mol. Genet. 4(12): 2209-2218.

Deelen J, et al. 2013. Gene set analysis of GWAS data for human longevity highlights the relevance of the insulin/IGF-1 signaling and telomere maintenance pathways. Age. 35(1): 235-249.

Dlugosch KM, Rieseberg LH. SNOWHITE: A pipeline for aggressive cleaning of next-generation sequence reads. In prep.

Dolinski K, Botstein D. 2007. Orthology and functional conservation in eukaryotes. Annu. Rev. Genet. 41: 465-507.

Downing T, Cormican P, O'Farrelly C, Bradley DG, and Lloyd AT. 2009. Evidence of the adaptive evolution of immune genes in chicken. BMC Res Notes. 2(1): 254.

Dumont JN, Brummet AR. 1980. The vitelline envelope, chorion and micropyle of *Fundulus heteroclitus* eggs. Gamete Res. 3(1): 25–44.

Eastman JT, and McCune AR. 2000. Fishes of the Antarctic continental shelf: evolution of a marine species flock? J Fish Biol. 57(A): 84–102.

Eberhard, WG. 1996. Female Control: Sexual Selection by Cryptic Female Choice. (Princeton Univ. Press, New Jersey).

Edgar RC. 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5): 1792–1797.

Eigenmann CH, and Beeson CH. 1893. Preliminary note on the relationship of the species usually united under the generic name *Sebastodes*. Am Nat. 27: 668-671.

Eigenmann CH, Eigenmann RS. 1890. Description of a new species of *Sebastodes*. CA Acad. Sci. Proc., Ser. 2, Vol. III, p.13.

Ellegren H. 2008. Comparative genomics and the study of evolution by natural selection. Mol Eco. 17(21): 4586-4596.

Elmer KR, et al. 2010. Rapid evolution and selection inferred from transcriptomes of sympatric crater lake cichlid fishes. Mol Ecol. 19(1): 197-211.

Endler JA, Basolo, AL. 1998. Sensory ecology, receiver biases and sexual selection. Trends Ecol Evol. 13(10): 415-420.

Escalier D. 2006. Knockout mouse models of sperm flagellum anomalies. Hum Reprod Update. 12(4): 449-461.

Evans JP. 2000. Getting sperm and egg together: things conserved and things diverged. Biology of Reproduction. 63(2): 355-360.

Ewing B, and Green P. 1998. Base-calling of automated sequencer traces using PHRED. II. Error probabilities. Genome Res. 8(3): 186-194.

Ewing B, Hillier L, Wendl MC, and Green P. 1998. Base-calling of automated sequencer traces using PHRED. I. accuracy assessment. Genome Res. 8(3): 175-185.

Forment J, et al. 2008. EST2uni: an open, parallel tool for automated EST analysis and database creation, with a data mining web interface and microarray expression data integration. BMC Bioinformatics. 9(1): 5.

García-Herrero S, Garrido N, Martínez-Conejero JA, Remohí J, Pellicer A et al. 2010. Ontological evaluation of transcriptional differences between sperm of infertile males and fertile donors using microarray analysis. J Assist Reprod Genet. 27(2-3):111-120.

Gavrilets S, and Vose A. 2005. Dynamic patterns of adaptive radiation. P Natl Acad Sci USA. 102(50): 18040-18045.

Gilbert CH. 1890. A preliminary report on the fishes collected by the steamer Albatross on the Pacific coast of North America during the year 1889, with descriptions of twelve new genera and ninety-two new species. Proc US Natl Mus. 13:49–126.

Gill T. 1864. Critical remarks on the genera *Sebastes* and *Sebastodes* of Ayres. P Acad Nat Sci Philad. 16: 145–147.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 11(5): 725-736.

Goudet G, Mugnier S, Callebaut I, Monget P. 2008. Phylogenetic analysis and identification of pseudogenes reveal a progressive loss of zona pellucida genes during evolution of vertebrates. Biol Reprod. 78(5): 796-806.

Grabherr MG, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotech. 29(7): 644-654.

Green P. 2010. Cross-match. Available at http://www.phrap.org.

Greenwood PH. 1981. The haplochromine fishes of the east African lakes. Cornell Univ. Press, Ithaca, NY.

Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PHYML 3.0. Syst Biol. 59(3): 307-321.

Hansen M, et al. 2007. Lifespan extension by conditions that inhibit translation in *Caenorhabditis elegans*. Aging Cell. 6(1): 95–110.

Heger A, and Ponting CP. 2007. Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes. Genome Res. 17(12): 1873-1849.

Hellmann I. et al. 2003. Selection on human genes as revealed by comparisons to chimpanzee cDNA. Genome Res. 13(5): 831-837.

Heras J, Koop BF, Aguilar A. 2011. A transcriptomic scan for positively selected genes in two closely related marine fishes: *Sebastes caurinus* and *S. rastrelliger*. Mar Genomics. 4(2): 93-98.

Holmskov U, Malhotra R, Sim RB, Jensenius JC. 1994. Collectins: collagenous C-type lectins of the innate immune defense system. Immunol Today. 15(2): 67-74.

Huang X, Madan A. 1999. CAP3: A DNA sequence assembly program. Genome Res. 9(9): 868-877.

Hubbard T, et al. 2005. ENSEMBL 2005. Nucleic Acids Res. 33(1): D447-D453.

Hughes AL. 1999. Adaptive evolution of genes and genomes. New York, NY: Oxford University Press, Inc.

Hughes AL. 2007. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. Heredity. 99(4): 364-373.

Humphries JM, and Miller RR. 1981. A remarkable species flock of pupfishes, genus Cyprinodon, from Yucatan, Mexico. Copeia. 52-64.

Hurst LD. 2002. The K_a/K_s ratio: diagnosing the form of sequence evolution. Trends Genet. 18(9): 486-487.

Hyde JR, Vetter RD. 2007. The origin, evolution, and diversification of rockfishes of the genus *Sebastes* (Cuvier). Mol Phylgenet Evol. 44(2):790-811.

Hyde JR, Kimbrell C, Robertson L, Clifford K, Lynn E, Vetter R. 2008. Multiple paternity and maintenance of genetic diversity in the live-bearing rockfishes *Sebastes* spp. 357: 245-253.

Ingram T. 2011. Speciation along a depth gradient in a marine adaptive radiation. P Roy Soc B-Biol Sci. 278(1705):613-618.

Jacobson R. 2013. http://www.gpo.gov/fdsys/pkg/FR-2013-08-06/pdf/2013-18832.pdf

Jansa SA, Lundrigan BL, Tucker PK. 2003. Tests for positive selection on immune and reproductive genes in closely related species of the murine genus *Mus*. J Mol Evol. 56(3): 294–307.

Jiang J, Zhang Y-B, Li S, Yu F-F, Sun F, Gui J-F. 2009. Expression regulation and functional characterization of a novel interferon inducible gene Gig2 and its promoter. Molecular Immunology. 46(15): 3131-3140.

Johansson ML, and Banks MA. 2011. Olfactory receptor related to class A, type 2 (V1r-Like Ora2) genes are conserved between distantly related rockfishes (Genus *Sebastes*). J Hered. 102(1): 113-117.

Johns GC, Avise JC. 1998. Tests for ancient species flocks based on molecular phylogenetic appraisals of *Sebastes* rockfishes and other marine fishes. Evolution. 52: 1135-1146.

Johnson TC, et al. 1996. Late Pleistocene desiccation of Lake Victoria and rapid evolution of cichlid fishes. Science. 273(5278): 1091-1093.

Jordan DS. 1921. The fish fauna of the California tertiary. Stanford Univ. Publ., Palo Alto, CA.

Jovine, Luca, et al. 2005. Zona pellucida domain proteins. Annu. Rev. Biochem. 74: 83-114.

Kendall AWJr. 2000. An historical review of *Sebastes* taxonomy and systematics. Mar Fish Rev. 62(2): 1-23.

Kimura M. 1991. The neutral theory of molecular evolution: a review of recent evidence. Jpn J Genet. 66(4): 367.

Koop BF, et al. 2008. A salmonid EST genomic study: genes, duplications, phylogeny and microarrays. BMC Genomics. 9(1): 545.

Körner CG, et al. 1998. The deadenylating nuclease (DAN) is involved in poly(A) tail removal during the meiotic maturation of *Xenopus* oocytes. 17(18): 5427-5437.

Kosiol C, et al. 2008. Patterns of Positive Selection in Six Mammalian Genomes. PLoS Genet. 4(8): 1-17.

Kunster A, et al. 2010. Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species. Mol Ecol. 19(1): 266-276.

Lack D. 1983. Darwin's finches. Cambridge, UK: Cambridge Univ. Press.

Laird PW. 2010. Principles and challenges of genome-wide DNA methylation analysis. Nat Rev Gen. 11(3): 191-203.

Larkin MA, et al. 2007. CLUSTAL W and CLUSTAL X version 2.0. Bioinformatics. 23(21): 2947-2948.

Lassmann T, Hayashizaki Y, Daub CO. 2009. TAGDUST - A program to eliminate artifacts from next generation sequencing data. Bioinformatics. 25(21): 2839-2840.

Lee Y-H, Vacquier VD. 1995. Evolution and systematics in Haliotidae (Mollusca: Gastropoda): inferences from DNA sequences of sperm lysin. Mar Biol. 124(2): 267–268.

Levitan DR, Ferrell DL. 2006. Selection on gamete recognition proteins depends on sex, density, and genotype frequency. Science. 312(5771): 267-269.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 22(13): 1658-1659.

Li Z, Gray AK, Love MS, Asahida T, and Gharrett AJ. 2006. Phylogeny of members of the rockfish (*Sebastes*) subgenus *Pteropodus* and their relatives. Can J Zool. 84: 527-536.

Lohse M, et al. 2012. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nuc Acids Res. 40: W622–W627.

Love MS, Morris P, McCrae M. 1990. Life history aspects of 19 rockfish species (Scorpaenidae: Sebastes) from the southern California bight. NOAA Tech Rep. NMFS 87.

Love MS, Yoklavich M, Thorsteinson L. 2002. The rockfishes of the Northeast Pacific. University of California Press, Berkeley.

Maan, ME, Seehausen, O. 2011. Ecology, sexual selection and speciation. Ecology Letters. 14(6): 591-602.

Machnes Z, Avtalion R, Shirak A, Trombka D, Wides R, Fellous M, Don J. 2008. Male-specific protein (MSP): A new gene linked to sexual behavior and aggressiveness of tilapia males. Horm Behav. 54(3): 442-449.

Magnuson-Ford K, Ingram T, Redding DW, Mooers AO. 2009. Rockfish (*Sebastes*) that are evolutionarily isolated are also large, morphologically distinctive and vulnerable to overfishing. Biol Cons. 142(8): 1787-1796.

Masta SE, Maddison WP. 2002. Sexual selection driving diversification in jumping spiders. Proc. Natl. Acad. Sci. USA 99(7): 4442–4447.

Mayr E. 1942. Systematics and the origin of species. Columbia Press, New York.

McCartney MA, Acevdo J, Heredia C, Rico C, Quenouille B, Bermingham E, McMillan WO. 2003. Genetic mosaic in a marine species flock. Mol Ecol. 12(11): 2963-2973.

McDonald JH, and Kreitman M. 1991 Adaptive protein evolution at the Adh locus in *Drosophila*. Nature. 351(6328): 652-654.

Meslin, Camille, et al. 2012. Evolution of Genes Involved in Gamete Interaction: Evidence for Positive Selection, Duplications and Losses in Vertebrates. PloS one. 7(9): e44548.

Meyer A, Van de Peer, Y. 2005. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). BioEssays. 27(9): 937-945.

Min XJ, Butler G, Storms R, Tsang A. 2005. TARGETIDENTIFIER: a web for identifying full-length cDNAs from EST sequences. Nucleic Acids Res. 33(2): W669-W672.

Min XJ, Butler G, Storms R, Tsang A. 2005. ORFPREDICTOR: predicting protein-coding regions in EST-derived sequences. Nucleic Acids Res. 33(2): W677-680.

Miyata T, Miyazawa S, Yasunaga T. 1979. Two types of amino acid substitutions in protein evolution. J Mol Evol. 12(3): 219-236.

Modig C, Westerlund L, Olsson P-E. 2007. Chapter 5: Oocyte zona pellucida proteins. Pg. 113–139. In Babin P. J., Cerda, J., and Lubzens, E. 2007. The Fish Oocyte: From Basic Studies to Biotechnological Applications. Springer.

Moore JES. 1903. The Tanganyika problem. Hurst Blackett, London.

Munk K. 2001. Maximum ages of groundfishes in waters off Alaska and British Columbia and considerations of age determination. Alaska Fish Res Bull. 8(1): 12-21.

Muse SV, Gaut. BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Molec Biol Evol. 11(5): 715-724.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3(5): 418-426.

Nielsen R, et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. Public Library of Science, Biology. 3(6): 976–985.

O'Brien KP, Remm M, Sonnhammer ELL. 2005. INPARANOID: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res. 33(1): D476-D480.
Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. Nature. 246(5428): 96-98.

Oleksiak MF, Churchill GA, Crawford DL. 2002. Variation in gene expression within and among natural populations. Nat Genet. 32(2): 261-266.

Palumbi SR. 1994. Genetic divergence, reproductive isolation, and marine speciation. Annu Rev Ecol Syst. 25: 547-572.

Palumbi SR. 2009. Speciation and the evolution of gamete recognition genes: pattern and process. Heredity. 102(1): 66-76.

Péterfy M, Gyuris T, Basu R, Takács L. 1994. Lissencephaly-1 is one of the most conserved proteins between mouse and human: a single amino-acid difference in 410 residues. Gene. 150(2): 415-416.

Podolsky RD. 2002. Fertilization ecology of egg coats: physical versus chemical contributions to fertilization success of free-spawned eggs. J Exp Biol. 205(11): 1657–1668.

Pond SLK, et al. 2011. A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol. 28(11): 3033-3034.

Pond SLK, Frost SDW. 2005. DATAMONKEY: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21(10): 2531-2533.

Pop M, Salzberg SL. 2008. Bioinformatics challenges of new sequencing technology. Trends in Genet. 24(3): 142-149.

Puebla O. 2009. Ecological speciation in marine v. freshwater fishes. J Fish Biol. 75(5): 960-996.

Pujolar JM, Pogson GH. 2011. Positive Darwinian selection in gamete recognition proteins of *Strongylocentrotus* sea urchins. Mol Ecol. 20(23): 4968-4982.

Rise ML, et al. 2004. Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics. Genome Res. 14(3): 478-490.

Rocha-Olivares A, Kimbrell CA, Eitner BJ, Vetter RD. 1999. Evolution of the mitochondrial cytochrome b gene sequence in the species-rich genus *Sebastes* (Teleostei: Scorpaenidae) and its utility in testing the monophyly of the subgenus *Sebastomus*. Mol Phylogenet Evol. 11(3): 426-440.

Russo CAM, Takezaki N, Nei M. 1995. Molecular Phylogeny and Divergence Times of Drosophilid Species. Mol Biol Evol. 12(3): 391-404.

Ryder OA. 2005. Conservation genomics: applying whole genome studies to species conservation efforts. Cytogenet Genome Res. 108(1-3): 6-15.

Santini F. et al. 2009. Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes. BMC evol biol 9(1): 194.

Sarropoulou E, Fernandes JMO. 2011. Comparative genomics in teleost species: knowledge transfer by linking the genomes of model and non-model fish species. Comp Biochem Physiol Part D, Genomics and Proteomics. 6(1): 92-102.

Schartl M. et al. 2013. The genome of the platyfish, *Xiphophorus maculatus*, provides insights into evolutionary adaptation and several complex traits. Nat genet. 45(5): 567-572.

Schneider CJ. 2008. Exploiting genomic resources in studies of speciation and adaptive radiation of lizards in the genus *Anolis*. Integ Comp Bio. 48(4): 520-526.

Schranz EM, Song B-H, Windsor AJ, and Mitchell-Olds T. 2007. Comparative genomics in the Brassicaceae: a family-wide perspective. Cur Opinion Plant Biol. 10(2): 168-175.

Schulz MH, Zerbino DR, Vingron M, Birney E. 2012. OASES: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 28(8): 1086–1092.

Seibel BA, Drazen JC. 2007. The rate of metabolism in marine animals: environmental constraints, ecological demands and energetic opportunities. Philos Trans R Soc Lond B: Biol Sci. 362(1487): 2061-2078.

Shapiro, MD. et al. 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature.* 428(6984): 717-723.

Shinomiya A, Ezaki O. 1991. Mating habits of the rockfish *Sebastes inermis*. Environ Biol Fish. 30: 15-22.

Sivasundar A, Palumbi SR. 2010. Parallel amino acid replacements in the rhodopsins of the rockfishes (Sebastes spp.) associated with shifts in habitat depth. J Evol Biol. 23(6): 1159-1169.

Smit AFA, Hubley R, Green P. 2010. REPEATMASKER Open-3.0. http://www.repeatmasker.org.

Smith J, Paton IR, Hughes DC, Burt DW. 2005. Isolation and mapping the chicken zona pellucida genes: an insight into the evolution of orthologous genes in different species. Mol Reprod Dev. 70(2): 133–145.

Sogard SM, Gilbert-Horvath E, Anderson EC, Fisher R, Berkeley SA, Garza JC. 2008. Multiple paternity in viviparous kelp rockfish, *Sebastes atrovirens*. Environ Biol Fish. 81(1): 7-13.

Somarelli, JA, Herrera RJ. 2007. Evolution of the 12 kDa FK506 binding protein gene. Biol of the Cell. 99(6): 311-321.

Spargo SC, Hope RM. 2003. Evolution and nomenclature of the zona pellucida gene family. Biol Reprod. 68(2): 358-362.

Stern A, Doron-Faigenboim A, Erez E, Martz E, Bacharach E, Pupko T. 2007. SELECTON 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. Nucleic Acids Research. 35(2): W506-511.

Storey J. 2002. A direct approach to false discovery rates. J R Stat Soc Series B Stat Methodol. 64(3): 479-498.

Sugawara T, Terai Y, Okada N. 2002. Natural selection of the rhodopsin gene during the adaptive radiation of East African Great Lakes cichlids fishes. Mol Biol Evol. 19(10):1807-1811.

Sunagawa S, et al. 2009. Generation and analysis of transcriptomic resources for a model system on the rise: the sea anemone *Aiptasia pallida* and its dinoflagellate endosymbiont. BMC genomics. 10(1): 258.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34(2): W609-W612.

Swanson WJ, Vacquier VD. 2002. The rapid evolution of reproductive proteins. Nat Rev Genet. 3(2):137-144.

Swanson WJ. 2003. Adaptive evolution of genes and gene families. Curr Opin Genet Dev. 13(6): 617-622.

Tavernarakis N. 2007. Protein synthesis and aging. Cell Cycle 6(10): 1168-1171.

Taylor MS, Hellberg ME. 2005. Marine radiations at small geographical scales: speciation in neotropical reef gobies (*Elacatinus*). Evolution. 59(2): 374-385.

Taylor JS, et al. 2003. Genome duplication, a trait shared by 22,000 species of ray-finned fish. Genome Res. 13:382–390.

Teufel A, Maass T, Galle P, Malik N. 2009. The longevity assurance homologue of yeast lag1 (Lass) gene family (Review). Int J Mol Med. 23(2): 135-140.

Thomas, JH. 2006. Adaptive evolution in two large families of ubiquitin-ligase adapters in nematodes and plants. Genome Res. 16(8): 1017-1030.

Tian X, Pascal G, Fouchécourt S, Pontarotti P, Monget P. 2009. Gene birth, death, and divergence: The different scenarios of reproduction-related gene evolution. Biol Reprod. 80(4): 616–621.

Trapnell C. et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat biotech. 28(5): 511-515.

Turner LM, Hoekstra HE. 2006. Adaptive evolution of fertilization proteins within a genus: variation in ZP2 and ZP3 in deer mice (*Peromyscus*). Mol Biol Evol. 23(9):1656-1669.

Uitenbroek DG. 1997. Simple interactive statistical analysis bonferroni calculator. http://www.quantitativeskills.com/sisa/calculations/bonfer.htm.

Via S. 2009. Natural selection in action during speciation. P Natl Acad Sci 106(1): 9939-9946.

von Schalburg KR, Rise ML, Brown GD, Davidson WS, Koop BF. 2005. A comprehensive survey of the genes involved in maturation and development of the rainbow trout ovary. Biol Reprod. 72: 687-699.

Voolstra CR, et al. 2009. Evolutionary analysis of orthologous cDNA sequences from cultured and symbiotic dinoflagellate symbionts of reef-building corals (Dinophyceae: Symbiodinium). Comp Biochem Physiol Part D: Genomics Proteomics 4(2): 67-74.

Vos MJ, Hageman J, Carra S, Kampinga HH. 2008. Structural and Functional Diversities between members of the human HSPB, HSPH, HSPA, and DNAJ chaperone families. Biochemistry. 47(27): 7001-7011.

Walsh MR, and Reznick DN. 2008. Interactions between the direct and indirect effects of predators determine life history evolution in a killifish. P Natl Acad Sci. 105(2): 594-599.

Warren WC, Hillier LW, Graves JAM. 2008. Birney E, Ponting CP, Grutzner F, Belov K, Miller W, Clarke L, Chinwalla AT. et al. 2008. Genome analysis of the platypus reveals unique signatures of evolution. Nature. 453(7192): 175-183.

Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet. 2(5): 333-341.

Yang Y, and Smith SA. 2013. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. BMC Genomics. 14:328.

Yang Z, and Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 15(12): 496-503.

Yang Z, and Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 17(1): 32-43.

Yang Z, Nielsen R, Goldman N, and Pedersen A-MK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 155(1): 431-449.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 17(1): 32-43.

Yang Z, Swanson WJ, and Vacquier VD. 2000. Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. Molec Biol Evol. 17(10): 1446-1455.

Yang Z, Wong WSW, and Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Molec Biol Evol. 22(4): 1107-1118.

Yang Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. Comp Appl Biosci. 13(5): 555-556.

Yeates SE, et al. 2009. Atlantic salmon eggs favour sperm in competition that have similar major histocompatibility alleles. Proc. R. Soc. B. 276(1656): 559-566.

Zerbino DR, Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18(5): 821–829.

Zhang Z, Li J, Zhao X–Q, Wang J, Wong G.K–S, and Yu J. 2007. KAKS_CALCULATOR: Calculating Ka and Ks through model selection and model averaging. Genomics, Proteomics and Bioinformatics. 4(4): 259-263.

Zhang, J. 2010. Positive Darwinian selection in gene evolution. Darwin's Heritage Today.