# DIFFERENTIAL SELECTION PRESSURE AMONG DUPLICATED GENES IN TELEOSTS

by

Richard J. Nuckels, Jr., B.S, M.S.

A dissertation submitted to the Graduate Council of
Texas State University in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
with a Major in Aquatic Resources
August 2018

Committee Members:

Dana M. García, Chair

Jeffrey M. Gross

Karen A. Lewis

Noland Martin

Chris Nice

# FAIR USE AND AUTHOR'S PERMISSION STATEMENT

## Fair Use

## Duplication Permission

# ACKNOWLEDGEMENTS

Many people have contributed to the completion of this project either directly or indirectly. The research for this project officially began in 2010, although my time working in the research labs of Dr. David Parichy, Dr. Jeff Gross, and Dr. Dana García starting in 2002 aided in the foundation and development of this project. Each of these individuals maintained research programs that fostered unique experiences that were beneficial to my overall development as a scientist and an educator.

I would like to thank my committee members for assisting and guiding me to the point of completing this project and helping to prepare me for the next stage of my academic and professional career. Noland Martin supervised me as a genetics teaching assistant before my committee was formed. I am grateful for his guidance at that time as I further developed teaching skills and beyond that as he helped guide my project with helpful suggestions. Karen Lewis joined my committee later and as someone with a biochemistry background added insight from a different perspective which was much appreciated. I am grateful to Jeff Gross as his influence with my project began before the project was developed. I appreciate all the years working in his lab and the influence he has had on my life. Chris Nice had an instrumental role in assisting with the development of this project. He provided me with tools and encouraged me to experiment with those tools and he was there for guidance as needed. His assistance with this project is greatly appreciated. Dana García shared so much of her time and insight during the development of this project which is very much appreciated. The many hours

shared over coffee discussing the development and direction of my project, work and teaching experiences, along with the friendship that evolved during this time is greatly appreciated. In addition to helping me to develop as a scientist, she provided me with a great example of managing the challenges of research, work, teaching, and family.

I would like to extend a heartfelt thanks to my wife, Mary, for providing me with a much loved foundation and helping me to stay grounded. She reminded me that life is not all about data collection and analyzing that data and her love and support will always be treasured. To my children I am thankful for the joys they brought and continue to bring to my everyday life. Watching them grow, develop and mature has been a truly heartwarming experience.

This work was supported through funding by the National Science Foundation-Project Flowing Waters awarded to Drs. Julie Westerlund, Weston Nowlin and Tim Bonner (NSF grant # 742306) and the Department of Biology, Texas State University.

# TABLE OF CONTENTS

**Page**

CHAPTER

# LIST OF TABLES

**Table** **Page**

## LIST OF TABLES (continued)

**Table** **Page**

## LIST OF FIGURES

**Figure** **Page**

# ABSTRACT

Gene and genomic duplications provide organisms with new genetic material subject to selection. Using *myo5, rab11,* and *rab27* gene families as models, I examined the evolutionary rate differences among duplicated genes and whether selective forces (e.g. purifying selection or positive selection) could be identified in one or both duplicated gene clades. I used phylogenetic and syntenic analyses along with ancestral chromosomal mapping to identify each duplicate. I then analyzed the duplicates using tests for evolutionary selection at the molecular level. Using a branch site-random effects likelihood test, I found evolutionary rate values (ω) to fall into two or three rate classes along at least one branch for one duplicated gene clade for each of the gene trees created. One rate value ($\omega_1$) for a percentage of codon sites was close to zero, representing purifying selection. A second rate value ($\omega_2$) for a percentage of the codons was much greater than one ($\omega_2$ >>>1), signifying positive selection. Two rate classes were observed in the teleost *myo5bb* branch for the motor domain, two other rate classes were observed for the cargo binding domain for the teleost *myo5bb* branch, and two rate classes were present in teleost *myo5ba* for the cargo binding domain. Also, in teleosts, I found two rate classes for the *rab11a* branch, *rab11a1* branch, and *rab27bb*. Using sequences from 7-10 organisms that diverged from a common ancestor 140-440 million years ago, I found ω values between 0.01 and 0.24 for the whole coding sequence for duplicated genes ranging in size from 200-220 codons. For longer coding sequences (1915 codons), ω ranged from 0.26 to 0.41. Sequences with rates (ω values) that are near or below 0.1

represent more highly conserved sequences.  Conversely, sequences with more variation at the nucleotide and amino acid level  represent less well conserved sequences, and the possibility for new functioning proteins or domains increases.  I examined the percentage of invariant codons present in each of the gene clades and found the percentage of codons that were invariant to range from 6% for the highly variable neck region of *myo5* to more than 30% for some highly conserved duplicated *rab11* and *rab27* genes and some highly conserved regions for some *myo5* duplicates.  I identified low ω values for codons for amino acids that have previously been linked with the functionality of the Myo5 and Rab proteins.  These data lead me to infer that the duplicated genes remain functional and suggest they have some modified, acquired functionality that remains to be identified.

# I. INTRODUCTION

The project described in this dissertation demonstrates the usefulness of readily available DNA sequence information from numerous genomes when analyzing the evolutionary history of gene families. The amount of information that is freely accessible has grown tremendously over the last 15 years as the rates and costs of DNA sequencing have improved along with the available computing power to handle billions of bases of DNA sequences (Schatz and Langmead, 2013). The amount of information that has been acquired and the speed with which this information has been acquired will continue improving, enabling researchers to address questions that were previously too difficult to answer (Schatz and Langmead, 2013). Comparing sequences from closely and distantly related organisms that are available in these DNA databases can provide insight into the regions in a gene that are evolutionarily significant. This information can aid in understanding molecular evolution and the diversity of life, and additionally can provide insight into diseases associated with genetic, epigenetic, or environmental influences.

My dissertation work shows that understanding the evolutionary history of genes provides new insights into the selective pressures operating on duplicated genes and the possible functional consequences of gene and genome duplication events (See Chapters 2-4). I have been particularly interested in genes involved in pigmentation and the utility of fishes as model organisms in which to examine the evolutionary history of these genes.

## History of Vertebrate Genomes (Genome Evolution)

Based on a combination of isozyme analysis and the sizes of vertebrate genomes, Ohno (1970) hypothesized that genome duplications took place early in the evolutionary history of vertebrates and other organisms. Two duplications of the vertebrate genome are thought to have occurred approximately 500-550 million years ago, and the addition of new genetic material is thought to have contributed to the diversification of the vertebrate lineage (Ohno, 1970). Since many genes that are found in single copy in other vertebrates have duplicated orthologs in teleosts, it is thought that approximately 300-350 million years ago a genome duplication event occurred in a clade of ray-finned fish known as teleosts, a group of fish that account for approximately 96% of all bony fish ((Taylor *et al.*, 2001). A common fate for duplicated genes is that they are lost in evolutionary time as missing ohnologs (Catchen *et al.*, 2009), although alternative outcomes include becoming pseudogenes (Li,1980), serving as a backup copy of the original gene or evolving new or modified functions (Ohno, 1970; Force *et al.*, 1999). Expression studies to identify the presence of RNA's or proteins can be carried out to sort out the presence of these duplicates but these studies are time and labor intensive in addition to being costly. Additional studies may be utilized testing the functionality of proteins but these are also time intensive and costly. I propose an approach that may shed light on the fate of the duplicates and my approach may provide information that helps determine if further examination of specific duplicated genes seems warranted thus potentially saving research time and money.

**Pigmentation-related Genes**

Duplicated pigmentation-related genes seem to be maintained in the fish genomes at a higher rate than non-pigmentation related genes (Shartl et al., 2009a). Others have speculated that pigmentation genes may be maintained because of the role they play in determining the external appearance of the animal, enabling identification of conspecifics, for example, and providing features on which selective pressures operate

(Braasch et al. 2009a). Therefore, as candidates for study, I identified a group of genes (*myo5*, *myo7*, *mlph*, *myrip*, *rab27*) related to pigmentation and that had previously been studied in other vertebrates and in some cases in fish (See *The Function of Myosin and Rab11 and Rab27* below). The genes *myo5, mlph,* and *rab27* had previously been studied in mouse pigment melanocytes, and *myo7, myrip,* and *rab27* had previously been studied in the retinal pigmented epithelium of the eye. (See reviews by Trybus, 2009; Hammer and Wagner, 2013). However, neither the existence of nor the potential function of duplicates of these genes had been examined. My search of the Ensembl genomic database revealed duplicates for all these genes based on orthology, paralogy, percent identity and ultimately phylogenetics and syntenic analysis. The physiological role of these genes had been examined by others, and the results indicated that they would be useful to my effort to better understand the molecular evolution of duplicated, pigmentation-related genes since some duplicates were involved in pigmentation and some were not. For example, *myo5a* is involved in pigmentation (Wu *et al.*, 2002), but its duplicate *myo5b* is not involved in pigmentation (Li and Nebenführ, 2008), and it is unknown whether duplicates of these two myosin genes (*myo5ab* and myo5bb) are involved in pigmentation.

## The Function of Myosin and Rab11 and Rab27

Myosins are a diverse superfamily of proteins found in all lineages of eukaryotes and include more than twenty families (myosins I – XX) of motor proteins that travel along tracks formed from actin (See reviews by Trybus 2009; Hammer and Wagner, 2013). Myosin proteins form homodimers and contain an N-terminal motor domain (head), a neck region, and, in some subfamilies of myosin, a C-terminal cargo binding domain. The motor domain contains highly conserved sites for ATP- and actin-binding. The neck shows the least amount of conservation at both the nucleotide and amino acid levels. For the myosin V subfamily, different accessory proteins associate with the C-terminus, enabling them to interact with cargo (See reviews by Trybus 2009; Hammer and Wagner, 2013).

Within the *myosin V* (*myo5*) gene family, the gene products have been shown to be involved in numerous cellular motor functions, including organelle transport and membrane trafficking in several cell types such as epidermal pigment cells, intestinal epithelial cells, and neural cells (Rodriguez and Cheney, 2002; Swiateca-Urban *et al.*, 2007; Hammer and Wagner, 2013). In mammals, there are three types of myosin V proteins: a, b, and c. Myosin Va is involved in transporting organelles, including melanosomes, along actin tracks and is expressed in much of the central nervous system (Hammer and Wagner, 2013). Myosin Vb is involved in endosome recycling in airway epithelial cells (Swiateca-Urban *et al*., 2007), and it is expressed in the central nervous system (Hammer and Wagner, 2013). Myosin Vc is primarily expressed in epithelial cells (Rodriguez and Cheney, 2002). With the many roles that myosins play along with the many types of tissues where these proteins are active, there have been abundant opportunities for duplicated versions of these genes to take on new or specialized roles.

As mentioned earlier, myosin V interacts with cellular cargoes like pigment granules via linker proteins. Rab27a, melanophilin, and Myo5a have been shown to interact and bind with each other to transport melanosomes along actin cytoskeletal tracks (Hammer & Wu 2007). In this assembly, Myo5a functions as the motor, while Rab27a and melanophilin mediate its attachment to melanin-containing pigment granules, the former through a direct interaction with the Myo5a cargo-binding domain (Wu *et al.*,2002). The Rab proteins in general are members of a superfamily of Ras-related proteins which are all small GTPase proteins. Rab proteins make up the largest family of the five subtypes of Ras proteins (Stenmark and Olkkonen 2001). Rab11 interacts with Myo5b in human and mice (Pylypenko et al. 2013), and it has been shown to participate in regulating endosome recycling (Schafer et al. 2013).

**A Brief Summary of Findings**

Using phylogenetic and syntenic analyses along with ancestral chromosomal mapping, I identified duplicate pairs for *myo5, rab11,* and *rab27.* I used *myo5, rab11,* and *rab27* as model gene families in which to examine the evolutionary rate differences among duplicated genes and whether selective forces (*e.g.*, purifying selection or positive selection) could be identified in the duplicated gene clades. I found evolutionary rate values (ω) to vary along at least one branch for one duplicated gene clade for each of the gene trees created. The rate value indicated purifying and positive selection were occurring along particular branches of the trees tested. I found a surprisingly high percentage of codons that were invariant, ranging from 6% for the highly variable neck region of *myo5* to more than 30% for some highly conserved duplicated *rab11* and *rab27* genes and some highly conserved regions for some *myo5* duplicates. I infer from the extreme conservation that the duplicated genes remain functional or have some modified, acquired functionality that remains to be identified.

# II. DUPLICATED MYOSIN V GENES IN TELEOSTS SHOW EVOLUTIONARY RATE VARIATIONS AMONG THE MOTOR, NECK, AND CARGO BINDING DOMAINS.

## Abstract

I analyzed evolutionary rates of conserved, duplicated myosin V (*myo5*) genes in nine teleost species to examine the outcomes of duplication events. Syntenic analysis and ancestral chromosome mapping suggest one tandem gene duplication event leading to the appearance of *myo5a* and *myo5c*, two rounds of whole genome duplication for vertebrates, and an additional round of whole genome duplication for teleosts account for the presence and location of the *myo5* genes and their duplicates in teleosts and other vertebrates and the timing of the duplication events. Phylogenetic analyses reveal a previously unidentified *myo5* clade that I refer to now as *myo5bb.* Analysis using dN/dS rate comparisons revealed large regions within duplicated *myo5* genes that are highly conserved. Codons identified in other studies as encoding functionally important portions of the Myo5a and Myo5b proteins are shown to be highly conserved within the newly identified *myo5bb* clade and in other *myo5* duplicates. As much as 30% of 319 codons encoding the cargo binding domain in the *myo5aa* genes are conserved in all three codon positions in nine teleost species. For the *myo5bb* cargo binding domain, 6.6% of 336 codons have zero substitutions in all nine teleost species. Using molecular evolution assays, I identify the *myo5bb* branch as being subject to evolutionary rate variation with the cargo binding domain, having 20% of the sites under positive selection and the motor domain having 8% of its sites under positive selection. I suspect that the duplicated genes have acquired novel functions, especially the cargo binding domain of the Myo5 proteins.

## Introduction

In 1970, Ohno proposed that two rounds (2R) of genome duplication had occurred in the evolutionary history of vertebrates and suggested such duplication events could have contributed to the sudden radiation and diversity of vertebrates (Ohno, 1970). Since then support has grown for a 2R (two rounds of vertebrate genome duplication) hypothesis (Hughes, 1999) such that it is currently widely accepted. An additional genome duplication event is thought to have occurred in the teleost lineage around 300 million years ago (Taylor *et al.*, 2001) since many genes that are found in single copy in other vertebrates have duplicated orthologs in teleosts. A common fate for duplicated genes is that they become lost in evolutionary time as missing ohnologs (Catchen *et al.*, 2009), although alternative outcomes include becoming pseudogenes (Li,1980), acting as a backup copy of the original gene or evolving new or modified functions (Ohno, 1970; Force *et al.*, 1999). In teleosts, numerous genes related to pigmentation provide us with a model to study these gene duplication events (Braasch *et al.*, 2007).

It has been suggested that pigmentation related genes retain their duplicates in fish at a higher rate than other genes (Braasch *et al.*, 2009a). Although the total number of genes in fish is not much different than tetrapods, Braasch *et al.*, (2009a) found that there are approximately 30% more pigmentation related genes comparted to tetrapods. Duplicated genes related to pigmentation have provided new opportunities for phenotypic diversity among fishes (Braasch *et al.*, 2009b) in addition to opening the evolutionary door for neofunctionalization for one of the duplicated genes to acquire a non-pigmentation related function over time. For example, Mills *et al.* (2007) showed that the *kita* gene is expressed in specific populations of pigment cells, whereas Mellgren and

Johnson (2005) observed the *kitb* gene to be expressed in non-pigment related cell types including neurons. Together, the expression patterns of these two duplicated genes approximate the expression pattern of the non-duplicated *Kit* gene in mouse.

Among the pigmentation-related genes that seem to have retained functionality after duplication, the myosin genes are particularly interesting. Myosins are a diverse superfamily of proteins found in all lineages of eukaryotes and include more than 20 families (myosins I – XX) of motor proteins that travel along tracks formed from actin, including some unconventional myosins (See reviews by Trybus 2009; Hammer and Wagner, 2013.). Myosin proteins form homodimers and contain an N-terminal motor domain (head), a neck region, and, in some subfamilies of myosin, a C-terminal cargo binding domain. The motor domain contains sites for ATP- and actin-binding. The neck shows the least amount of conservation at the nucleotide and amino acid levels. For the myosin V subfamily, different accessory proteins associate with the myosin proteins, enabling them to interact with cargo. (See reviews by Trybus 2009; Hammer and Wagner, 2013.)

Within the *myosin V* (*myo5*) gene family, the gene products have been shown to be involved in numerous cellular motor functions, including organelle transport and membrane trafficking in several cell types such as epidermal pigment cells, intestinal epithelial cells, and neural cells (Rodriguez and Cheney, 2002; Swiateca-Urban *et al.*, 2007; Hammer and Wagner, 2013). In mammals, there are three types of myosin V proteins (a, b, and c). Myosin Va is involved in transporting organelles, including melanosomes, along actin tracks and is expressed in much of the central nervous system (Hammer and Wagner, 2013). Myosin Vb is involved in endosome recycling in epithelial cells (Swiateca-Urban *et al*., 2007), and it is expressed in the central nervous system (Hammer and Wagner, 2013).  Myosin Vc is primarily expressed in epithelial cells (Rodriguez and Cheney, 2002).   With the many roles that myosins play along with the many types of tissues where these proteins are active, there have been abundant opportunities for duplicated versions of these genes to take on new or specialized roles.

Acquisition of new roles is associated with differential evolutionary rates.  Muse and Gaut (1994) devised a model that determined an evolutionary rate (ω) based on a ratio of non-synonymous and synonymous substitutions in an alignment and this rate could vary from one branch to another in a phylogeny.  Nielsen and Yang (1998) developed a codon substitution model that allowed rates at each codon to vary but kept the rate among the branches constant.  With an increase in computational power, newer refined codon substitution models were developed to allow for different rates of codon site evolution to occur among codons and among branches (Yang and Nielson, 2002; Bielawski and Yang, 2003; Bielawski and Yang, 2004; Zhang *et al.*, 2005; Anisimova and Yang, 2007; Smith *et al.*, 2015). The quantification of evolutionary rates using these methods can provide insight into the fates of duplicated gene and elucidate the mechanisms by which novel functions might evolve.

Here, I characterize the *myo5* duplicates and their evolutionary history in vertebrates.  I identify a branch in a phylogeny of the myosin gene family for a duplicated gene (*myo5bb*) in teleosts and spotted gar.  I show that regions encoding the actin binding domains are highly conserved, including third codon positions, but there is more variability in third codon positions near the 3′ end of the gene where the cargo binding domain is encoded.  In addition to presenting data that supports previously described genome duplication events, namely the vertebrate R1/R2 and fish specific genome duplications, I identify a tandem gene duplication event for the *myo5a* and *myo5c* genes, and I propose a model for the evolution of the *myo5* gene family. In the proposed evolutionary model of the *myo5* gene family, I provide phylogenetic and syntenic data that supports the vestiges of two different *myo5b* clades that likely originated from one of the ancient R1/R2 vertebrate genome duplication events. With the analysis of codons, I identify extreme purifying selection present in 96 codons out of 319 codons (30.1%). These 96 codons are invariant and have zero nucleotide substitutions in the nine teleosts examined for the *myo5aa* 3′ end. In contrast, 46 codons out of 742 codons (6.2%) in the *myo5ab* neck region of the *myo5* gene are subject to extreme purifying selection for the nine teleosts examined.

## Materials and Methods

### Sequence acquisition

I collected *myosin 5* sequences using the Ensembl genomic database (Ensembl Release 86), Genbank release 221.0, and the Japanese Lamprey Genome Project (Table 1). The following species and genomic assemblies were used for *myo5* sequence downloads: nine teleost species (cavefish, *Astyanax mexicanus,* AstMex102; cod, *Gadus morhua,* gadMor1; fugu, *Takifugu rubripes,* FUGU 4.0; medaka, *Oryzias latipes,* HdrR; platyfish, *Xiphophorus maculatus,* Xipmac4.4.2; stickleback, *Gasterosteus aculeatus,* BROAD S1; tetraodon, *Tetraodon nigroviridis,* TETRAODON 8.0; tilapia, *Oreochromis niloticus,* Orenil1.0; zebrafish, *Danio rerio,* GRCz10), one holostean fish (spotted gar, *Lepisosteus oculatus,* LepOcu1), one lobe finned fish (coelacanth, *Latimeria chalumnae,* LatCha1), one amphibian (western clawed frog, *Xenopus tropicalis,* JGI 4.2), five sauropsids (chicken, *Gallus gallus,* Gallus_gallus-5.0; turkey, *Meleagris gallopavo* , Turkey_2.01 ; duck, *Anas platyrhynchos*, BGI_duck_1.0; Chinese soft shell turtle, *Pelodiscus sinensis*, PelSin_1.0;green anole lizard, *Anolis carolinensis,* AnoCar2.0), two mammals (human, *Homo sapiens,* GRCh38.p7; mouse, *Mus musculus,* GRCm38.p5), one cartilaginous fish (elephant shark, *Callorhinchus milii*, Genbank assembly-GCA_000165045.2 ) two jawless vertebrates (sea lamprey, *Petromyzon marinus,* Pmarinus_7.0; Japanese lamprey, *Lethenteron japonicum*, Japanese lamprey genome project- APJL00000000), and two urochordates (sea squirts, *Ciona intestinalis,* KH; *Ciona savignyi,* CSAV 2.0).

**Syntenic analysis**

Using Biomart in the Ensembl database, genes located within 1.5 megabases of each *myo5* gene were identified. Synteny maps were constructed based on conserved patterns of gene locations for each of the *myo5* gene families, and results are presented in Figure 1. Construction of syntenic regions used zebrafish and tetraodon genomes as an initial source to identify genes within 1.5 megabases for each *myo5* gene family. After downloading genes from BioMart within the previously specified regions, I found 39 genes from zebrafish and 125 genes from tetraodon for the *myo5aa* gene family, 89 genes from zebrafish and 176 genes from tetraodon for the *myo5ab* gene family, 117 genes from zebrafish and 70 genes from tetraodon for the *myo5ba* gene family, and 74 genes from zebrafish and 137 genes from tetraodon for the *myo5bb* gene family. For the *myo5c* gene family, I used the same set of genes as in the *myo5ab* gene family since *myo5c* and *myo5ab* are directly next to each other on the chromosome for most of the teleosts tested, and *myo5a* and *myo5c* are directly next to each other on the chromosome for other vertebrates that have those two genes. The number of genes I found within 1.5 megabases of any *myo5* gene was between 39 and 176. In making a more concise syntenic map presented in Figure 1 I used approximately 30 genes total and about ten genes in each *myo5* gene neighborhood. Each gene neighborhood generally contained genes within 200,000 bases of each *myo5* gene.

**Ancestral chromosome mapping**

I used ancestral chromosomal reconstructions from Nakatani *et al.* (2007) and Bian *et al.* (2015) to determine the timing of the *myo5* gene duplication events relative to the major genome duplication events. Nakatani *et al.* provide chromosomal maps for syntenic blocks of genes for the genomes of human, chicken, and medaka and relate these syntenic blocks back to one of ten ancestral chromosomes designated A-J. Bian *et al.* provide chromosomal maps for syntenic blocks of genes for medaka, zebrafish, arowana and spotted gar and relate these syntenic blocks back to one of thirteen ancestral chromosomes present before the teleost and non-teleost fish (including spotted gar) split. Utilizing these two sets of chromosomal mapping data, I was able to identify whether the genes of interest split after the vertebrate first or second whole genome duplication or if the genes of interest were a result of the fish specific genome duplication (Figure 2).

**Alignment and phylogenetics**

Eighty-seven sequences were aligned using ClustalW and Geneious Pro 6.0 (Biomatters Ltd). Sequences were virtually translated and verified to contain open reading frames. The ends of the aligned sequences were trimmed and smaller alignments from three regions (motor domain, neck, cargo binding domain) within the *myo5* gene were obtained from the full length coding sequence alignment. Model testing was performed for each of the four alignments, and the model with the best AICc value was chosen for the generation of the phylogenetic trees using Geneious 6.0. Using Mr.Bayes 3.1 and a GTR+I+G model of evolution, trees were generated for the full length coding sequence (6870 bp) of *myo5*, the motor domain, the neck, and the cargo binding domain. The parameters used in the Mr. Bayes-generated trees were as follows: three gamma categories were used with unconstrained branch lengths. Markov Chain Monte Carlo methods were used for 1,100,000 steps with thinning every 200 steps, four heated chains, and a preheated chain temperature of 0.2. A burn-in length of 500 steps was used. Alternative models were tested using maximum likelihood and parsimony methods, and these provided similar topologies. Figures 3-6 show the final trees generated for each alignment.

For the four alignments I generated, I removed sequences that did not have at least 50% coverage. For example, the duck *myo5c* sequence only had sequence coverage in the motor domain and in the cargo binding domain, so it was only included in those alignments and phylogenetic analyses and not in the neck or full sequence alignments. Similarly, other sequences were missing sequence data for more than 50% of the alignment. These sequences were not included in those specific alignments (Figure 7).

**dN/dS rates and identification of invariant codons**

I determined the evolutionary rate (dN/dS) using MEGA6. “dN” is defined as the ratio of non-synonymous substitutions per non-synonymous site; “dS” is defined as the ratio of synonymous substitutions per synonymous site. Maximum likelihood reconstructions of ancestral states were generated using a Muse-Gaut model (Muse and Gaut 1994) of codon substitution and a general time reversible model (Nei and Kumar

2000) for nucleotide substitution. I used MEGA6 to determine the dN and dS values for each codon in the alignment for a specific clade which generally consisted of 8-10 teleost sequences for a specific *myo5* duplicate. Summing the dN and dS values for all the codons in the alignment and then dividing dN by dS allowed us to determine the dN/dS ratio for each alignment. To quantify the percentage of codons that are invariant and experiencing extreme purifying selection, I counted the number of codons in each of the original four alignments (whole gene, motor domain, neck, and cargo binding domain) that had dN and dS values of zero and divided this by the total number of codons in the alignment to determine the percentage of codons that are invariant and experiencing extreme purifying selection (Tables 2-4).

**Selection Tests**

I used the Datamonkey server and the HyPhy software package (Delport *et al.*, 2010; Kosakovsky Pond *et al.*, 2005) to test for purifying selection, positive selection, and episodic selection at the codon level and the branch level among the phylogenies I generated. Trees that were generated as described previously using the Geneious Software package were saved as Nexus files and uploaded to the Datamonkey Server to run the selection tests. I used BUSTED (Branch site Unrestricted Statistical Test for Episodic Diversification) to assess whether episodic diversification occurs on at least one branch and at least at one site in the phylogeny. The BUSTED test allows for varying rates of evolution (ω) applied to a constrained model of selection (null model) and an unconstrained model of selection (alternative model) using a Likelihood Ratio Test. I then tested the alignments using MEME (Mixed Effects Model of Evolution), BS-REL (Branch Site Random Effects Likelihood), aBS-REL (adaptive BS-REL), and SLAC

(Single Likelihood Ancestor Counting). MEME identifies the number of sites (codons) showing episodic diversifying selection using a maximum likelihood approach. Different evolutionary rates are allowed for each codon within an alignment. The aBS-REL test determined which branches in the phylogeny showed evidence of diversifying selection using a likelihood ratio test and providing statistical support with $p \leq 0.05$. Methods for the tests I used in the analyses are further described in Nielsen and Yang (1998; REL), Murrell *et al.* (2012; MEME), Kosakovsky Pond and Frost (2005; SLAC), Kosakovsky Pond *et al.* (2011; BS-REL), Murrell *et al*. (2015; BUSTED), Smith *et al*. (2015; aBS-REL). I used 8-10 teleost sequences from the alignments to test for selection among the duplicated *myo5* genes for the MEME, REL, and SLAC tests. I did this for each teleost duplicated *myo5* gene clade and for the smaller regions within the gene. For example, I used the 5′ end motor domain alignment of nine teleost sequences for the *myo5aa* teleost gene clade and ran the MEME, REL, and SLAC selection tests. Similarly, I tested the neck and cargo binding domain (CBD) for the *myo5aa* teleost clade, and I ran these same selection tests using the comparable domains for the teleost clades which included *myo5ab, myo5ba,* and *myo5bb* genes (Table 5).

## Results

### Syntenic analysis

To determine whether *myo5* duplicates arose through duplication of individual genes, chromosomes or their segments, or entire genomes, I performed syntenic analysis. I found the chromosomal locations for *myo5aa* and *myo5ab* in zebrafish on chromosomes 18 and 25, respectively, and the locus for *myo5c* is directly downstream of *myo5ab*. This arrangement with *myo5aa* and *myo5ab* on separate chromosomes and *myo5c* on the same chromosome as *myo5ab* was observed in all teleosts examined; furthermore, *myo5c* was observed directly downstream of *myo5a* in non-teleost vertebrates (Figure 1). Initial phylogenetic analyses revealed a new *myo5* clade (the *myo5bb* clade), and syntenic analyses provided further support of the presence of this clade and neighboring genes in teleosts, spotted gar, chicken, duck, turkey, turtle, coelacanth, and shark. This clade appears to be absent in mammals, anole, and *Xenopus*. Figure 1 shows genes that are syntenic with *myo5a* as rectangles, genes syntenic with *myo5ba* as ovals, and genes syntenic with *myo5bb* as triangles.

I traced the origin of extant *myo*5 sequences to ancestral vertebrate and teleost chromosomes to further test the findings from the syntenic analysis (Figure 2). All *myo5* sequences traced back to an ancestral vertebrate chromosome A. Nakatani *et al*. (2007) identified six chromosomes or linkage groups numbered A0-A5 resulting from two whole genome duplication events (R1 and R2) and a fission event. The *myo5a* and *myo5c* tandem duplicated genes are linked with the A4 fragment (Figure 2A). The *myo5ba* genes are linked with the A0 fragment and the *myo5bb* genes are linked with the A1 fragment (Figure 2A). Co-duplicated genes exist, for example, *mbd1* near *myo5ba* and *mbd3* near *myo5bb*. Additional co-duplicated genes were identified with *mapk4* found near *myo5ba* and *mapk6* found near *myo5a-myo5c*. The *onecut3* gene was found near *myo5bb,* and *onecut6* was found near *myo5a-myo5c*. Teleost *myo5* genes were traced back to three of thirteen ancestral teleost chromosomes. *myo5ba* was traced back to ancestral teleost chromosome i, *myo5aa* and *myo5ab-myo5c* were traced back to ancestral teleost chromosome j, and *myo5bb* was traced back to ancestral teleost chromosome m (Figure 2B).

I identified two partial *myo5* sequences for each lamprey species tested. Figure 1 shows the syntenic arrangement of genes around the *myo5* sequences in both lamprey species and Figures 4a and 4b show the alignment of lamprey sequences in relation to the whole *myo5* genes and the smaller regions of the genes used in this study. Using BLAST to compare 400,000 bases of Japanese lamprey DNA around the Japanese lamprey *myo5* sequence against the sea lamprey genomic database in Ensembl, I found the *pigo* and *ensab* genes on one side of the *myo5* genes, and I found *mapk4*, *cfap53*, and *atp8b* on the other side of the *myo5* genes.

**Phylogenetic and dN/dS analyses**

To understand the molecular evolution of the *myo5* gene family, phylogenetic analysis was performed using 87 genes from 24 different species. (See Table 1 for names and genomic database identifiers.) Using ClustalW, I produced a final alignment of 6468 base pairs per gene. Four phylogenetic trees were generated representing the full-length coding sequence (Figure 3), the portion encoding the cargo-binding domain at the 3′ end of the *myo5* gene (Figure 4), the 5′ end of the gene which encodes the motor domain with its highly conserved ATP-binding domain (Figure 5), and the more variable portion of the *myo5* gene which encodes the neck and tail regions (Figure 6). Figure 4A shows where the smaller alignments fit with the full-length alignment. The *myo5aa* teleost sequences form a monophyletic clade, and the *myo5ab* teleost sequences form a monophyletic clade (Figures 3, 4 and 6). Separate clades form for the *myo5ba* teleost sequences, the *myo5bb* sequences, and the *myo5c* sequences (Figures 3, 4 and 6). Tetrapod *myo5a* was monophyletic with the teleost *myo5aa* and *myo5ab* (Figures 3, 4 and 6). Similarly, tetrapod *myo5c* was monophyletic with teleost *myo5c* (Figures 3, 4 and 6); however, tetrapod *myo5b* was for the most part monophyletic with teleost *myo5ba*, but not with teleost *myo5bb* (Figure 3, 4 and 6). These topologies were less evident in the phylogenetic trees generated for the sequences encoding the motor domain due to the higher degree of conservation. (See "Codon specific analysis" below and Figure 5.)

To assess the likelihood of the newly identified *myo5bb* clade having followed an evolutionary path leading to neo-functionalization or one leading to their becoming pseudogenes, I determined dN/dS values for each clade and each region of the *myo5* gene family (see Figure 8 and Table 2). The dN/dS ratios for the *myo5ba* and *myo5bb* were higher than the dN/dS ratios for *myo5aa* and *myo5ab* (Figure 8). The percentage differences are summarized in Table 3. For the *myo5a* duplicates (*myo5aa* and *myo5ab*), the percentage increase is higher for the dN/dS values for the motor domain and the cargo binding domain with the largest amount of dN/dS change taking place in the cargo binding domain for the *myo5ab* clade. This increased dN/dS could reflect the *myo5ab* clade's having evolved numerous different functions among teleosts for the cargo binding domain or it could be a non-functional or sub-functional domain. For the *myo5bb* clade I see a much smaller increase in the dN/dS rates for the cargo binding domain with a 68% increase compared to the 250% increase seen in the *myo5ab* clade. This smaller increase in dN/dS rates supports the idea that this region is more likely to have retained function compared to the cargo binding domain for the *myo5ab* clade. The dN/dS for the motor domain has also increased a relatively small amount (33%) f0or the *myo5bb* clade compared to the motor domain of the *myo5ab* clade (140%).

In addition to calculating the dN/dS ratios for each of the whole genes and for specific regions within the *myo5* genes, codon specific values for dN and dS for each alignment tested were calculated. For a given region of a *myo5* gene, *e.g.* the 5′ end, dN and dS were calculated by comparing sequences from at least eight teleost species. The 3′ end where the cargo-binding domain is encoded evinced far fewer invariant sites in the *myo5ab* clades compared to the *myo5aa* clades. I identified 30.1% of the codons for the teleost *myo5aa* clade to be subject to extreme purifying selection but only 7.5% of codons in the *myo5ab* clade showed extreme purifying selection (dN=dS=0). No substitutions were identified in any of the three positions for codons in these invariant sites among the eight to nine teleosts analyzed. The *myo5ba* clade has 23.8% of codons invariant in the diverse teleost sequences tested, but only 6.6% of the codons for the *myo5bb* clade are invariant. The *myo5ab* neck region also showed fewer invariant sites than the *myo5aa* neck region (Table 4). For other regions of the *myo5* genes, the percentages of invariant codons were similar among the different paralogs. For example, the 5′ end of the *myo5* genes, where the actin- and ATP-binding domains are encoded, shows similar percentages for each clade ranging from 11.5% to 13.4%, suggesting the motor domain is similarly conserved between homologous clades and may be functional for all the *myo5* duplicates in teleosts.

**Codon-specific analysis**

After identifying an unexpectedly high percentage of invariant codons, I compared the dN/dS ratios of codons that encode amino acids that are known to play a functional role in MYO5 proteins in mammals. Amino acids linked with functionality in mammals are highly conserved in teleosts in the 5′ region for *myo5a* and myo5*b* duplicates (Table 10), suggesting these duplicates retain the motor functions related to ATP- and actin-binding. However, there is a significant difference in the cargo binding domain when looking at the codon sites linked with functionally important amino acids for MYO5 proteins (Table 10).  I examined the ten sites that are linked with RAB11a binding to MYO5b in mammals (Pylypenko, 2013). I found that the dN/dS values for these ten sites are either zero or mathematically undefined, because the value of the denominator (dS) equaled zero, highlighting the high conservation for these sites in *myo5ba* in teleosts (Table 10).  These same sites are not as well conserved in the *myo5bb* duplicates.

Out of the 217 codons in the 5′ region of the *myo5* genes, I specifically selected 21 codons that code for amino acids linked with the functional myosin motor activity for further analysis.  The dN/dS rate for these 21 codons was lower compared to the dN/dS rate of the entire 5′ region.  The average dN/dS for codons in the 5′ regions for teleost *myo5* genes was 0.08 (Table 2), but the average dN/dS value for the 21 codons linked with functionality was only 0.02.  The increase in conservation for these 21 codons was seen for all five *myo5* genes in teleosts for the 5′ region which included the part encoding the ATP-binding domain for the Myo5 proteins (Table 2).

In the myosin head, the aspartate at position 134 (D134) for human sequences is an example of an amino acid that was conserved in all *myo5* sequences analyzed for all species, with the following exceptions: The inferred amino acid sequence from the single *myo5* gene in *Ciona* manifests a conservative D→E change.  In *Tetraodon*, there is also a D→E change for *myo5ab*.  Cavefish and platyfish show sequence variation in the 5′ end of the *myo5aa* gene such that the D134 amino acid is not present (Figure 7).  The cavefish *myo5aa* gene has a premature stop codon which truncates the protein before the cargo-binding domain is translated, so cavefish may not have a functional *myo5aa* gene.

Another feature of the myosin head ATP-binding domain is the p-loop, a region of the protein that interacts with the terminal phosphate on ATP (amino acids 163-170 in human sequences; Coureux *et al.*, 2003). Nearly all the amino acid residues in the *myo5* p-loops are highly conserved, yielding a consensus sequence of GESG**A**GKT. The only variation in this region in the fish sequences is found in the *myo5bb* teleost clade, in which all sequences replace the alanine at position 167 with a serine, yielding the consensus sequence GESG**S**GKT.

The 742 codons in the neck region show the largest amount of sequence variation with dN/dS rates ranging from 0.23 for *myo5aa* to 0.39 for *myo5ab* (Table 2). When comparing the duplicates for this region, *myo5ab* has a larger dN/dS value (0.39) than the paralogous *myo5aa* genes (0.23). The 23 codons that code for amino acids linked with actin binding have much more conserved sequences compared to the neck domain except for *myo5aa* (Tables 2, 7 and 11). For *myo5aa*, the dN/dS value for the 23 codons associated with actin binding is 0.23 but for the other four teleost genes the dN/dS range is 0.04 to 0.14.

For the 319 codons in the 3′end of the myo5 genes, which include the cargo binding domain, the *myo5ab* and *myo5bb* genes have the highest dN/dS values at 0.34 and 0.37 respectively. The *myo5aa* and *myo5ba* genes are much more conserved in this region with dN/dS rates of 0.10 and 0.17, respectively. When looking at the 10-13 codons linked with cargo binding, *myo5ba* has a dN/dS rate of zero and *myo5aa* has a dN/dS rate of 0.08 (Tables 9 and 10).

## Selection Test Results

I carried out several selection tests (Table 5) accessed from the Datamonkey server and utilizing the HyPhy software package. I used BUSTED to test for selection across the phylogeny and this test revealed that episodic diversifying selection was occurring somewhere in the full length phylogeny ($p < 0.05$). I specifically selected *myo5b* branches to test as foreground branches, and the remaining branches were considered background branches. Three rate classes ($\omega_1, \omega_2, \omega_3$) were determined for the test branches and background branches for a constrained model (null model) and an unconstrained model of selection. For the *myo5b* test branches, episodic diversifying selection was occurring on at least one site with a $\omega_1 = 0.01$ for 74.5% of the sites, $\omega_2 =0.60$ for 23.63% of the sites, and $\omega_3=248.95$ for 1.87% of the sites. To more specifically address on which branch(es) and at which sites selection was taking place, I used MEME (Mixed Effects Model of Evolution). The results from the MEME test showed many sites with episodic diversifying selection in the neck region of the *myo5* gene, which is the least conserved region of the *myo5* genes. The functional domains are in the motor domain and in the cargo binding domain. In the cargo binding domain, I see more episodic diversifying selection in the *myo5bb* clade of teleosts versus the *myo5ba* clade of teleosts. I also see large variations between these two clades when comparing the number of codons experiencing positive selection versus purifying selection using REL (Random Effects Likelihood). The REL test shows the number of sites (codons) experiencing positive (REL +) or negative/purifying (REL -) selection. The REL test computes two Bayes factors such that one will test for $dN < dS$, suggesting purifying selection, and the other Bayes factor will test for $dN > dS$, suggesting positive selection at specific codons (Nielsen and Yang, 1998; Kosakovsky Pond and Frost, 2005). The results from the REL test showed that five sites in the cargo binding domain of *myo5bb* were subject to positive selection and 247 sites were subject to purifying selection. For the *myo5ba* duplicate there were zero sites subject to positive selection and 78 sites subject to purifying selection. For the *myo5aa* clade and the cargo binding domain there were two sites under positive selection and 132 sites subject to purifying selection. For the *myo5ab* duplicate, there was one site subject to positive selection for the cargo binding domain and 103 sites subject to purifying selection.

SLAC (Single Likelihood Ancestor Counting) testing uses a statistical approach under the assumption that all codon sites in a provided alignment evolve at the same rate. This is a more conservative approach compared to the other tests employed. One of the outcomes of this test is a dN/dS rate is given for each codon in the alignment. Another outcome is the number of sites subject to positive or negative selection. Using this conservative approach, I see fewer sites in the alignments subject to positive selection compared to the REL and MEME tests (Table 5). The SLAC test also shows that the highest proportion of sites under purifying selection are in the motor domain (41%; 88 out of 217 codons for *myo5aa*-motor) with the cargo binding domain following (27%; 94 out of 343 codons for *myo5aa*-cbd), and the neck region having the fewest sites (18%; 153 out of 830 codons for *myo5aa*-neck) under purifying selection.

A BS-REL (Branch Site-Random Effects Likelihood) test was used on the phylogenies I generated to test for episodic or diversifying selection along branches. I identified episodic selection taking place along the *myo5bb* branch leading up to the ray-finned fish lineage (Table 6). On this branch, 20% of the sites in the cargo binding domain are under positive selection, 26% of the sites are under neutral selection, and 54% of the sites are under purifying selection. Two other branches that showed signs of episodic diversifying selection in the cargo binding domain were branches that led to the *myo5ba* teleost clade and the *myo5b* clade as a whole. However, both of those branches had a much higher percentage of sites under purifying selection and many fewer sites subject to positive selection.

An aBS-REL (adaptive Branch Site-Random Effects Likelihood) test was used on all the branches in the cargo binding domain (CBD). Out of 147 branches tested in the CBD, 78 branches were subject to a single rate class, ω (dN/dS). The remaining 69 branches were modeled using two rate classes $\omega_1$ and $\omega_2$. Of these 69 branches that were subject to two rate classes, five branches showed evidence of diversifying selection with statistical significance ($p < 0.004$). Four of the five branches were for single genes for a single species (*myo5ba*-spotted gar, *myo5bb*-spotted gar, *myo5*-sea lamprey, *myo5bb*-coelacanth). The fifth branch that showed evidence of diversifying selection was the branch at the base of teleost *myo5bb* ($p = 0.0003$). On this branch leading to the CBD for the *myo5bb* teleost clade, there were two rate classes identified $\omega_1 = 0.316$ for 76% of the sites and $\omega_2 =80.1$ for 24% of the sites.

## Discussion

I investigated gene duplications in the myo5 family to provide insight into the mechanisms that constrain and promote the evolution of novel gene functions. In fish, *myo5* and other myosin genes have been examined but an analysis of the duplicated genes has not been done. Sonal *et al.* (2014) described *myo5b* expression in fish but did not identify or examine *myo5bb*. Similarly, Sittaramane and Chandrasekhar (2008) described *myo5a* expression along with other myosin genes in zebrafish but did not examine the duplicated versions. Hodel *et al.* (2014) studied Myo7a in fish using Western blots, fluorescent immunohistochemistry and immunogold labeling and proposed that the antibody used is likely recognizing Myo7a1 since it was raised against zebrafish Myo7a1. Hodel *et al.* also explain that the antibody used by Lin-Jones *et al.* (2009) likely did not differentiate between Myo7a1 and Myo7a2 since their antibody was raised against the human sequence. No further details on the epitope sequences are provided. Here, I demonstrate the usefulness of analyzing genes duplicated in teleosts to provide insight into molecular evolutionary processes.

**Syntenic analysis**

Our syntenic analysis supports a model in which numerous events in the evolutionary history of teleosts and non-teleost chordates contributed to *myo5* gene duplications and gene losses. Four gene or genome duplication events could account for the five *myo5* genes present in teleosts, four *myo5* genes present in spotted gar, and three to four *myo5* genes present in the lobe finned fish lineage. Three of these duplicated *myo5* genes (*myo5a, myo5ba*, and *myo5bb*) appear to result from the vertebrate genome duplication events, R1 and R2 (Figure 1). One of these *myo5* duplications may be the result of a tandem gene duplication (TGD) event which preceded the divergence of jawed vertebrates; the resulting paralogs are currently referred to as *myo5a* and *myo5c*. The fourth duplication event I identified is specific to teleosts and is likely the result of the teleost- or fish-specific genome duplication event (R3); this event led to the *myo5aa* and *myo5ab* genes in fish. As four genes would be expected from the two genome duplication events (R1 and R2), I suspect a gene loss took place after the R2 duplication event. The syntenic data and ancestral chromosome mapping support these interpretations on the placement of duplication events in the evolutionary history of the *myo5* gene family. The newly identified *myo5bb* clade present in birds, turtle, shark, coelacanth, spotted gar and teleosts seems to represent a case of hidden paralogy. Hidden paralogy (Kuraku, 2010) is a term used to highlight the misidentification of orthologs and paralogs due to depauperate data or inadequate analyses of existing genomic data. Hidden paralogy can be identified when more data is made available to properly place duplicated genes in a phylogeny. For example, a gene for a species is initially labeled as an "A" version only because there are no other versions identified for that gene and for that species and no "B" versions have been identified in other organisms yet. However, once a "B" version of the gene is identified in another organism and then this sequence data is included in a new phylogenetic analysis, then the misidentification of the original "A" version is observed (hidden paralogy) and it is determined that the previously labeled "A" version should be a "B" version of the gene (Qiu *et al.,*2011 and Kuraku, 2013). In the Ensembl genomic database, several of these genes are identified as *myo5b* for non-teleosts or not identified at all for teleosts. These *myo5bb* genes are more closely related to teleost *myo5bb* than they are to human or mouse *myo5b.* For example, chicken *myo5b* should not be assumed to be more closely related to human *myo5b* even though they have the same name. My results show that the chicken *myo5b* gene is a *myo5bb* gene, and it should be seen as more closely related to fish and other vertebrate *myo5bb* genes (hidden paralogy).

For each lamprey species, I found a gene that aligns with the 5′ end of my alignments and a second gene that aligns with the 3′ end of my alignments. However, I suspect that one of three scenarios accounts for this finding.  One possibility is that there is an error in the assembly of the contigs in Ensembl for the sea lamprey. For the sea lamprey, there are approximately 80,000 "N" nucleotides in between the *ensab* gene and the *myo5* CBD where the *myo5* motor domain should be located.  I identified the *myo5* motor domain on an independent small scaffold without any genes around it.  I suspect that this scaffold, which includes the sea lamprey *myo5* motor domain, is misplaced and that it should be part of the 80,000 "N" nucleotides which occur between the *ensab* gene and the *myo5* CBD.  Although two separate Japanese lamprey contigs were identified (one with the motor domain and a second with the CBD), both of these contigs are on the same scaffold, and results from using the surrounding sequences as query sequences for BLAST searches and comparing syntenic regions suggest that the two Japanese lamprey sequences are part of the same, contiguous *myo5* gene.  Additionally, the sizes of the exons and introns for Japanese lamprey sequences are compatible to those of sea lamprey.

A second possibility is that the presence of two genes for each species reflects a fracturing event. A fracturing event could have occurred early in the lamprey's evolutionary history such that one of the duplicated genes fragmented into two genes.  If fragmentation took place before the divergence of the sea lamprey and Japanese lamprey, then these events would have only happened once in the ancestral lamprey.  A third possibility is that two ancestral *myo5* genes in lamprey could have gradually lost part of each gene and over time these became shortened.  If this were the case then my phylogenetic analysis should have placed the CBD for lamprey in a different *myo5* clade than the lamprey *myo5* motor domains.  Interestingly, I see a couple of other examples in my study where there are truncated *myo5* genes. One of the cavefish genes, *myo5aa*, seems to have only a short sequence covering the motor domain. I identified a short tetraodon *myo5c* sequence that contains the first 3000 bp in the 5′ region of the *myo5* gene and is missing the cargo binding domain. I also identify the *C. intestinalis* sequence to be a short sequence of 2,982 bp, missing the cargo binding domain.

In addition to providing a model for the evolutionary history of the *myo5* genes, my syntenic analysis (along with the phylogenetic analysis) helped validate the nomenclature of the duplicated teleost genes. In some cases (namely the *myo5bb* genes) a *myo5* name has not been assigned to one of the *myo5* genes in Ensembl or other genomic data depositories (Table 1).

**dN/dS analyses**

As branch lengths represent the number of substitutions per site, I thought the long branches evident in the *myo5bb* lineage might reflect a large amount of substitutions resulting in amino acid changes (Figures 3-6). Were that the case, my examination of the amino acid sequences encoded by the *myo5bb* genes would be expected to reveal an increase in the dN/dS ratio, reflecting a faster rate of evolution. A faster rate of evolution, in turn, could reflect a release from selective constraints, perhaps consequent to the duplicate becoming a pseudogene. However, I observed strong purifying selection in the region of the gene encoding the myosin head, reflected in a surprising amount of invariance in select codons (Table 3).

Some of the invariant codon sites in teleosts code for amino acids that have been shown to be functionally important in the orthologous MYO5a and MYO5b proteins in mammals (Pylypenko *et al.*, 2013). The *myo5* codons orthologous to those in human

*MYO5A* or *MYO5B* linked with a functional role in the Myo5 protein such as motor activity, ATP-binding, actin-binding, or cargo-binding (Pylypenko *et al.*, 2013) had smaller dN/dS values than other codons in the *myo5* genes, indicating these codons were among the most conserved codons in all nine teleost species and among the duplicated genes (Tables 7-10). For many of the sites in the *myo5* genes (Tables 7-10), dN/dS values equal zero as a result of having zero nonsynonymous nucleotide changes at that codon.

The more functionally constrained and therefore more conserved parts of myosin 5 proteins include the motor domain or ATP-binding region, the actin binding domain, and the cargo-binding domain (Figure 5). The largest percentage difference of invariant codons between two paralogous clades exists between the cargo-binding domain of the *myo5aa* (30.1%) and *myo5ab* (7.5%) clades and between the cargo binding domains of the *myo5ba* (23.8%) and *myo5bb* (6.6%) clades (Table 3). The cargo-binding domain has previously been characterized as playing a role in lightly or non-pigmented (*dilute*) phenotypes (Nascimento *et al.,*1997), and in this domain sequence conservation mostly persists. The serine residue at position 1650 (in human Myo5b) has been shown to be a site for phosphorylation by which release of melanosomes from the myosin motor is regulated (Karcher *et al.*,2000; Pylypenko *et al.*, 2013). S1650 is present in all Myo5 sequences analyzed with the exception of Myo5ab from *Tetraodon nigroviridis*, in which the serine residue is replaced with an X. The basic residues K1706 and K1779 were identified in Li et al (2008) as having an important role in regulating motor activity by binding to the acidic motor domain sites D134 and D136. With few exceptions, these four residues are conserved in all sequences analyzed.

One exception to D134 not being conserved is found in the cavefish *myo5aa* gene. The cave-dwelling, non-pigmented cavefish have a premature stop codon in the *myo5aa* gene (Figure 7D), precluding translation of the cargo binding domain, therefore likely preventing the Myo5aa protein from transporting any melanin cargo. Since the sequence data is based on the cave-dwelling cavefish, a comparison with the closely related surface-dwelling form which has pigment could provide additional insight into the significance of these changes in the duplicated gene. It is possible that the surface-dwelling cavefish utilize the Myo5ab protein to transport melanosomes or the surface-dwellers might have a fully functional Myo5aa protein that contains a cargo-binding domain.

Slightly C-terminal to the D134 site is the p-loop of the motor domain. This highly conserved region has an alanine to serine (A→S) change in the *myo5bb* clade. This change could render the *myo5bb* gene products non-functional by compromising ATP binding, or this could be a regulatory change as serine residues are known to be sites of phosphorylation. Ramakrishnan et al. (2002) summarized the numerous variants for this conserved sequence with a general motif of GXXXXGKT being present in 92 identified variations of this region. Although this A to S substitution has been identified in two other proteins (phosphoenolpyruvate carboxykinase and dioxygenase), it had not been previously identified in any of the myosin proteins.

**Selection Tests**

Our tests for selection using MEME showed that there were more evolutionary changes taking place in the neck region of the *myo5* genes compared to other regions of the *myo5* genes. In comparing the cargo binding domain, there were many more sites in the *myo5bb* clade subject to positive selection (5 with p < 0.05) than in the *myo5ba* clade (0 with p < 0.05). Using the BS-REL selection test, I found evidence of episodic diversifying selection along the *myo5b* clades, including the whole *myo5b* clade and the *myo5ba* clade, and the *myo5bb* clade. Most of the diversity here came along the *myo5bb* branch, supporting the idea that this branch and the *myo5bb* cargo binding domain has experienced more evolutionary changes than other clades, increasing the likelihood for the neofunctionalization or subfunctionalization of this clade. This inference may be supported by the observation that the sites associated with binding Rab11a are not as well conserved in the Myo5bb duplicates, suggesting that Myo5bb binds to something other than Rab11a or that there are different regions within the Myo5bb cargo-binding domain that have not been previously identified and that are involved in binding to cargo. Also, because there is significant variation among teleosts for the *myo5bb* clade, there could be different cargoes or functions associated with this Myo5bb region in teleosts.

In addition to detailing the evolutionary history of the myosin V gene family, I present evolutionary rate data comparing duplicated genes. These evolutionary rate comparisons highlight a high degree of sequence conservation at codons linked with functionality for the Myosin 5 proteins. Using phylogenetic and syntenic analyses along with evolutionary rate comparisons, my data imply that these duplications have persisted over evolutionary time with a high degree of conservation at specific sites. This finding could support an evolutionary pathway leading to neofunctionalization or raise the question as to why sites are so highly conserved over hundreds of millions of years if these duplicated genes are non-functional.

Using dN/dS evolutionary rate comparisons, selection tests, and the identification of a high percentage of codons subject to extreme purifying selection, I present data linking the newly identified *myo5bb* clade with a high degree of conservation at functionally important amino acids, suggesting *myo5bb* is either a duplicate that has retained function or has acquired a neofunctionality. The high degree of conservation of specific sites linked with functionality supports an evolutionary pathway leading to neofunctionalization for the *myo5ab* duplicated genes (found in teleosts only) and *myo5bb* duplicated genes (found in birds, turtle, shark, coelacanth, spotted gar and teleosts).

I have utilized a family of duplicated genes with one of the duplicates known to play a role in the pigmentation process but the role of the duplicates of these genes remains to be identified. Teleosts seem to have a higher proportion of pigment related genes in duplicate compared to non-teleosts (Braasch *et al.* 2009a). It is possible that the duplicates may still be functional, and the duplicates may be expressed at a different time in development or in a different type of cell. It is also possible that a neofunctional role may have evolved in one of the duplicates. Although, I suspect that *myo5aa* is carrying out the melanosome shuttling role similar to *myo5a* in non-teleosts, the role of *myo5ab* remains to be determined. Additionally, I suspect that *myo5ba* in fish are carrying out the same role as *myo5b* (more accurately, *myo5ba*) in non-teleosts but what is taking place among the newly identified *myo5bb* clade remains a mystery. Due to the high degree of conservation in the motor domain, I suspect that the proteins encoded by these genes still have a functional role and I suspect that the new role is related to the variability and positive selection I have identified in the cargo binding domain.

The data presented for percentage of invariant codons or codons under extreme purifying selection demonstrated this high level of purifying selection remains in fish and non-fish vertebrates in duplicated versions of the *myo5* genes.  As far as the first and second codon position, there seems to be high conservation at the codons linked with functionality, supporting the idea that these duplicated genes are likely functional, active and subject to selection.  For a large percentage of codons, the third codon positions are highly conserved over hundreds of millions of years of evolution. I speculate that this conservation may reflect post-transcriptional regulation of gene expression by microRNAs.  Conserved sequences as short as six to eight nucleotides in length may provide an opportunity for microRNA binding (Brennecke *et al*., 2005, Krek *et al*., 2005, Lewis *et al*., 2005) I identified invariant codons to exist throughout my alignments of duplicated genes and at times these invariant codons were clustered in groups of as many as ten invariant codons (30 identical nucleotides), raising the possibility that some duplicates may be regulated by microRNAs.  The data presented provide insights into molecular evolution and underscores the usefulness of teleosts in helping to understand the evolutionary consequences of gene duplication events.

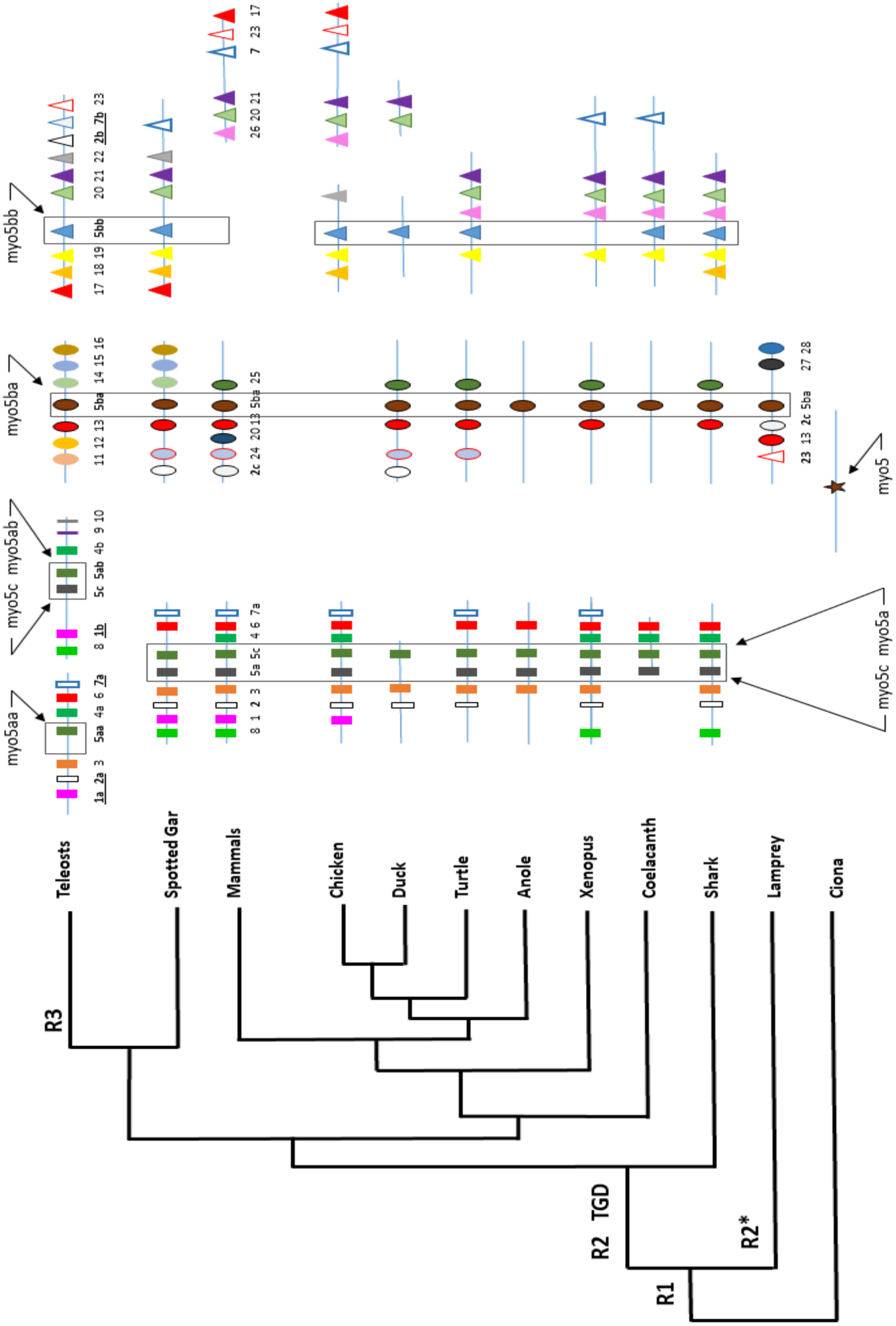


Figure 1. Cladogram and synteny for *myo5.*

**Figure 1 (preceding page). Cladogram and synteny for *myo5.*** Cladogram and syntenic diagram supporting key hypothesized evolutionary events in the history of *myo5* genes in teleosts and other chordates. Putative vertebrate genome duplication events (R1 and R2) led to the creation of three *myo5* copies, *myo5a* (green rectangle), *myo5ba (*brown oval) and *myo5bb* (blue triangle) in jawed vertebrates. The other copy that should have been created from two whole rounds of genome duplication likely became a pseudogene. The timing of the second genome duplication (R2*) of lamprey have been debated (see text for details). Each shape represents a gene, and the numbers under the teleost and mammal genes are listed below with the gene name. Shapes that are bordered and unshaded are identified as being orthologous and paralogous. Shaded shapes show orthologous relationships. Gene 1 is *cyp19a1*and it is colored as a pink rectangle. It is found near *myo5aa* and *myo5ab/myo5c* in teleosts and near *myo5a/myo5c* in spotted gar, mammals and chicken. Gene 2 is *mapk6* and it is shown as a black bordered, unshaded rectangle. It is found near *myo5aa* in teleosts and near *myo5a/myo5c* in spotted gar, mammals, chicken, duck, turtle, *Xenopus*, and shark. I found other *mapk* genes near *myo5ba* (unshaded, black bordered oval) and near *myo5bb* (unshaded, black triangle) gene families. I propose a tandem gene duplication event (TGD) that occurred before the divergence of jawed vertebrates which led to the formation of *myo5a* and *myo5c* as neighboring genes (green and grey rectangles in box). The TGD could have taken place before or after R2. The third whole genome duplication specific to teleosts (R3) led to the formation of *myo5aa* and *myo5ab* with a subsequent loss of a duplicated *myo5c* next to teleost *myo5aa*. The chromosomal locations for these genes on zebrafish are as follows: *myo5aa* chromosome 18, *myo5ab* chromosome 25, and *myo5c* chromosome 25 directly downstream of *myo5ab* in teleosts. The location for *myo5c* is directly downstream of *myo5a* in non-teleost vertebrates. Similar syntenic observations were made for other teleosts and non-teleosts, supporting the inference that *myo5a* and *myo5c* are tandem duplicates. Gene names corresponding with numbers listed under teleost and mammal genes are as follows: 1a-*cyp19a1*, 1b-*cyp19b*, 2a-*mapk6*, 2b-*map2ka*, 2c-*mapk4*, 3-*gnb5,* 4a-*arpp19a,* 4b-*arpp19b,* 5aa-*myo5aa*, 5ab-*myo5ab*, 5ba-*myo5ba*, 5bb-*myo5bb*, 5c-*myo5c*, 6-*fam214a,* 7a-*onecut1*, 7b-*onecut3*, 8- *ap4e1*, 9-*rsl24d1*, 10-*prtgb*, 11-*pvrl1a*, 12-*chek1*, 13-*cfap53*, 14-*il7r* , 15-*capslb*, 16-*lmbrd2*, 17- *btbd2* , 18- *hmg20b* , 19- *unk13a* , 20- *mbd3b* , 21- *tcf3*, 22- *zbtb7*, 23-*atp8b2*, 24-*skai*, 25- *acaa2* , 26-*mex3d* , 27-*ensab,* 28-*pigo*. See Table 1 for *myo5* gene identifiers and chromosomal locations.

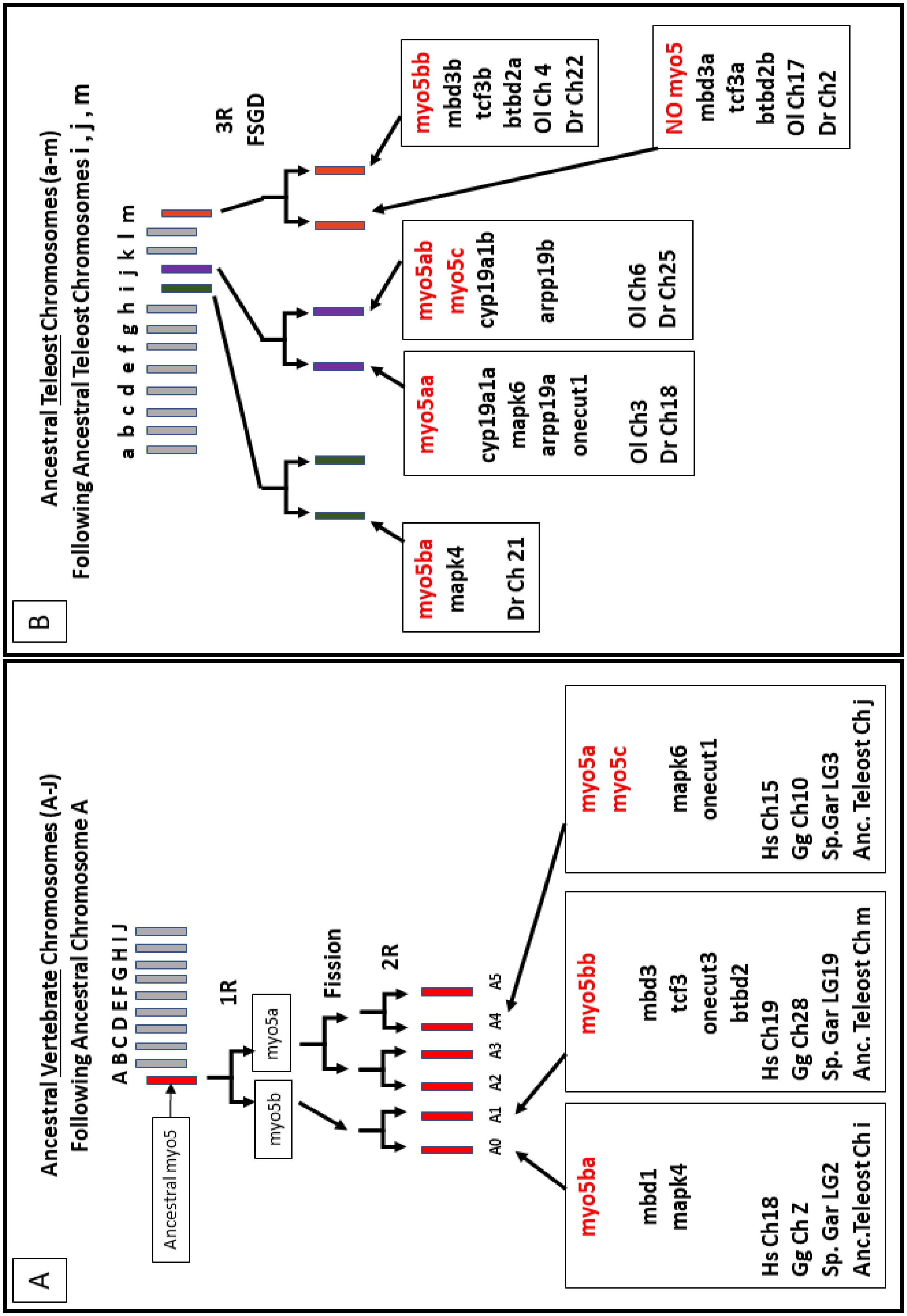


**Figure 2. Ancestral chromosomes for *myo5.***

**Figure 2 (Preceding page). Ancestral chromosomes for *myo5*.** Ten ancestral vertebrate proto-chromosomes have been previously described along with thirteen ancestral teleost chromosomes (Nakatani *et al*., 2007; Bian *et al*., 2015). All *myo5* genes were traced back to ancestral vertebrate chromosome A (Panel A). After two whole rounds of genome duplication and a fission event, six chromosomal fragments (A0-A5) existed. *myo5* genes and select co-duplicated genes are shown in the boxed region along with what ancestral chromosome fragment these genes are derived from. Organisms and the chromosomal regions the boxed genes are located on are found in the bottom of the boxed regions. Hs=*H.sapiens*, Gg=*G.gallus*, Sp. Gar=spotted gar, Anc. Teleost=Ancestral teleost, Ch=Chromosome, LG=linkage group. In panel B, ancestral teleost chromosomes are shown along with the 3 chromosomes that gave rise to the *myo5* genes in teleosts (Dr=*D.rerio*, Ol=*O.latipes*).

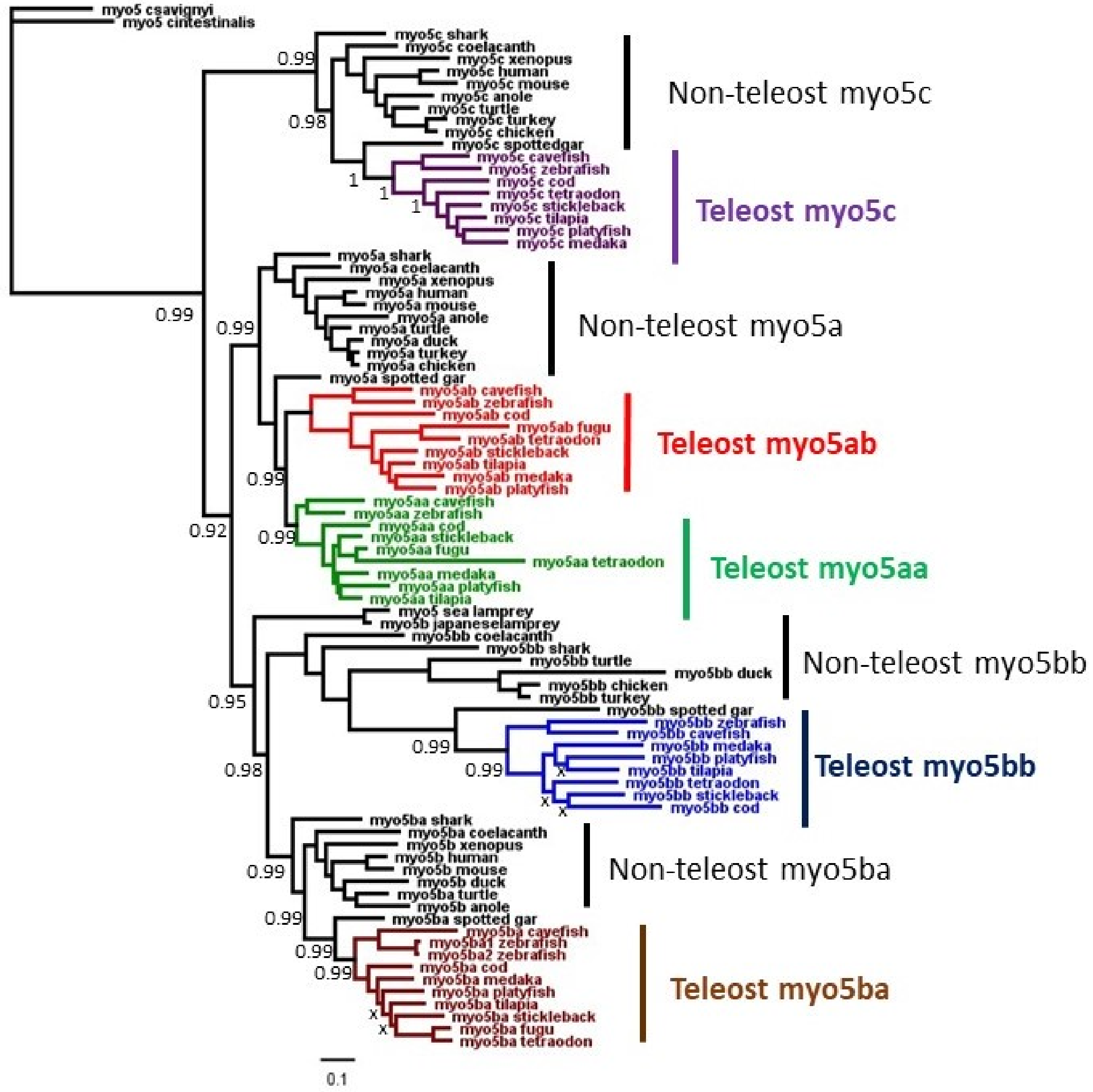


**Figure 3. Bayesian phylogenetic tree for full length *myo5* sequence.** Phylogenetic tree for full length coding sequence (6468 bp) of *myo5* using Mr.Bayes 3.1 and a GTR+I+G model of evolution. Teleost *myo5bb* clade shown in blue with an extended branch leading up to the clade. Posterior probability values are provided for some nodes. If not shown, the posterior probability value ranges from 0.94 to 1. "X" labels in myo5ba and myo5bb clades denote posterior probability values between 0.62 and 0.78. The scale bar represents 0.1 substitutions per site.

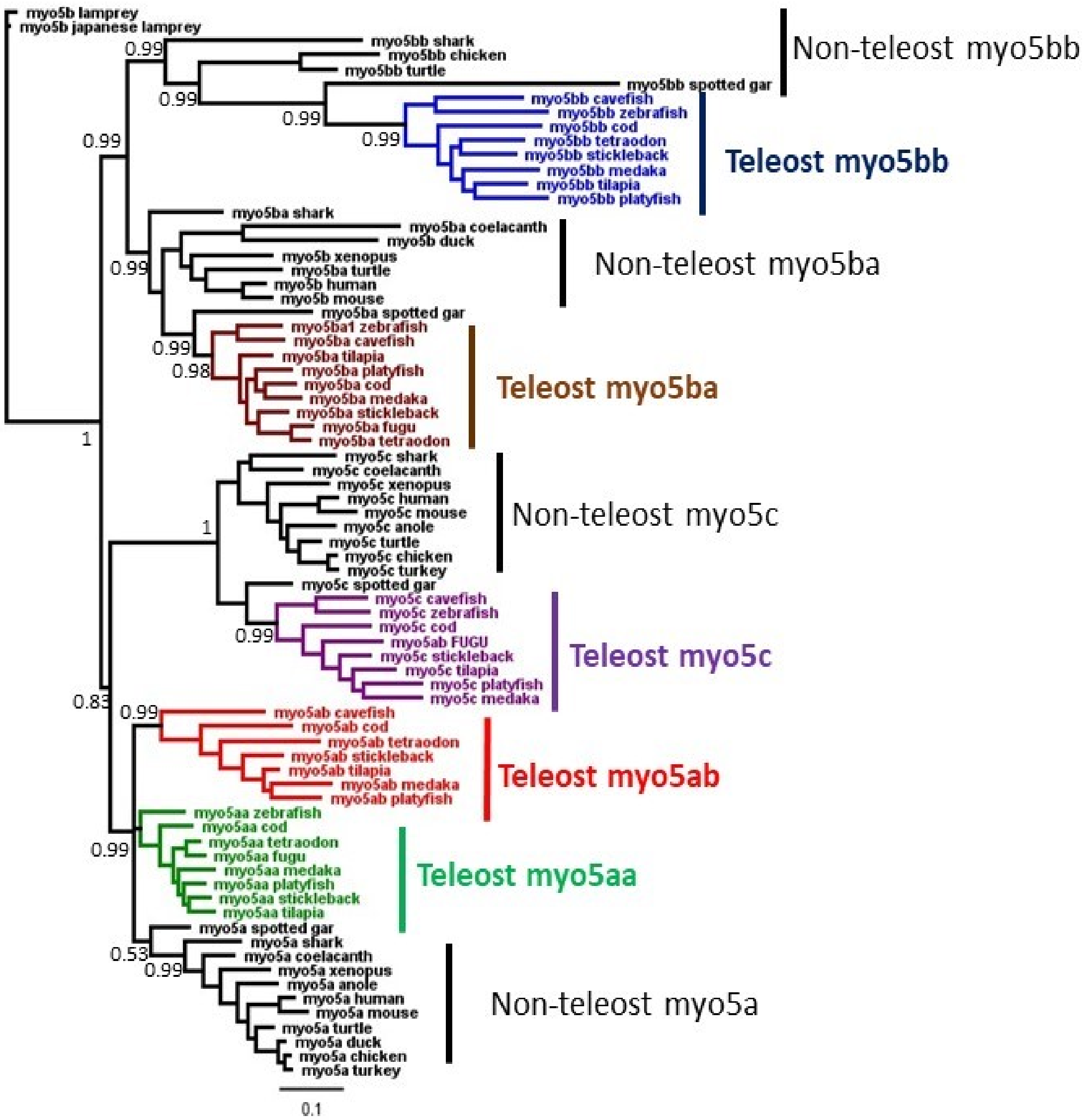


**Figure 4. Bayesian phylogenetic tree for *myo5* sequence encoding the cargo binding domain.** Phylogenetic tree for the cargo-binding domain (1035 bp fragment) of *myo5* using Mr.Bayes 3.1 and a GTR+I+G model of evolution. This 1035 bp fragment is found at the 3′ end of the *myo5* gene and includes the coding sequence for the dilute domain for *myo5a*. Teleost *myo5bb* clade is shown in blue with an extended branch leading up to the clade.

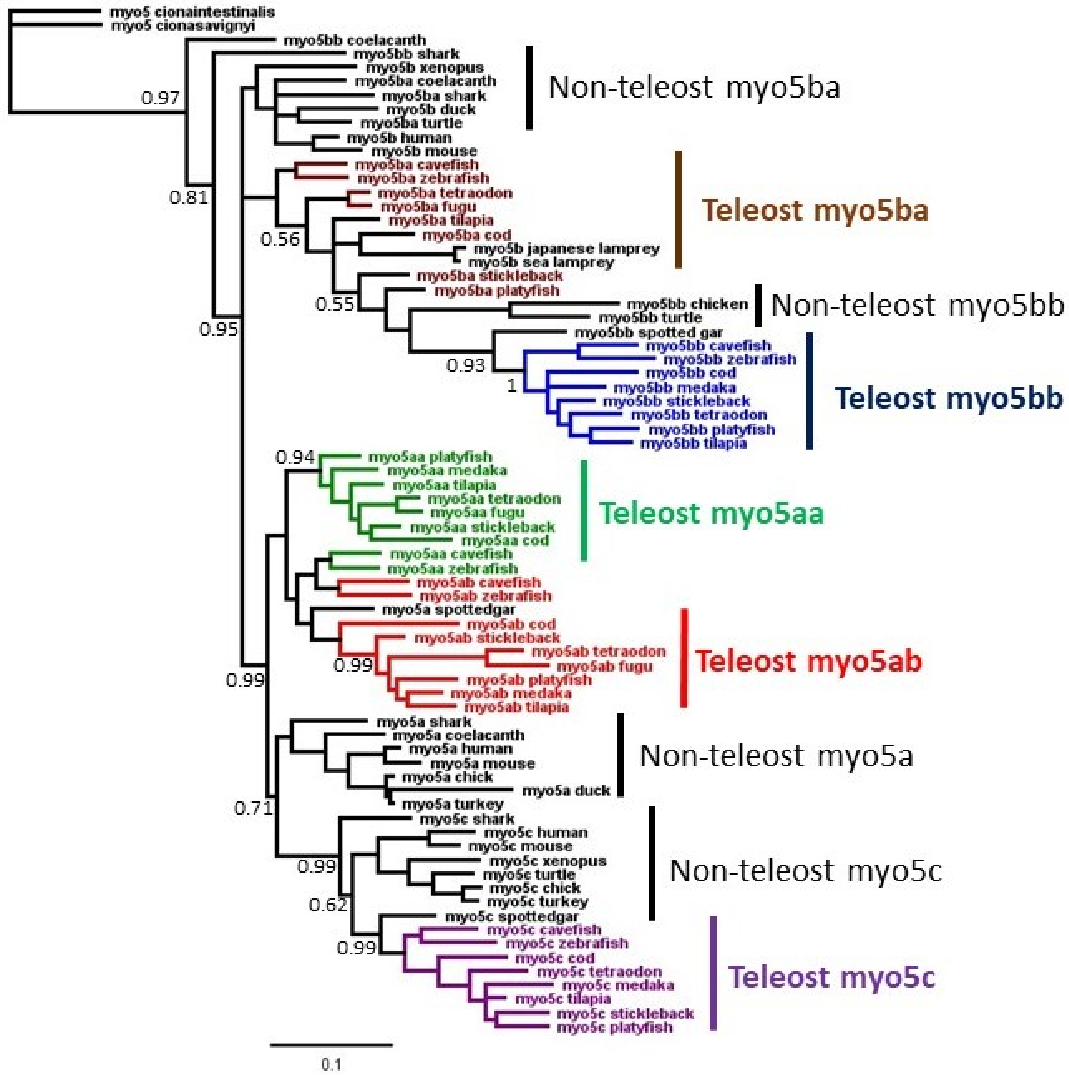


**Figure 5. Bayesian phylogenetic tree for *myo5* sequence encoding the motor domain.** Phylogenetic tree for the motor domain (a 651 bp fragment) of *myo5* using Mr.Bayes 3.1 and a GTR+I+G model of evolution. An alignment of 80 sequences from 18 different species was created, and the 5′end of the *myo5* gene including the ATP binding domain was used to generate this tree. The teleost *myo5bb* clade is shown in blue with an extended branch leading up to the clade. There are a few more branches that appear unresolved in this tree as a result of the high level of sequence conservation for the motor domain across taxa for the *myo5* gene clades.

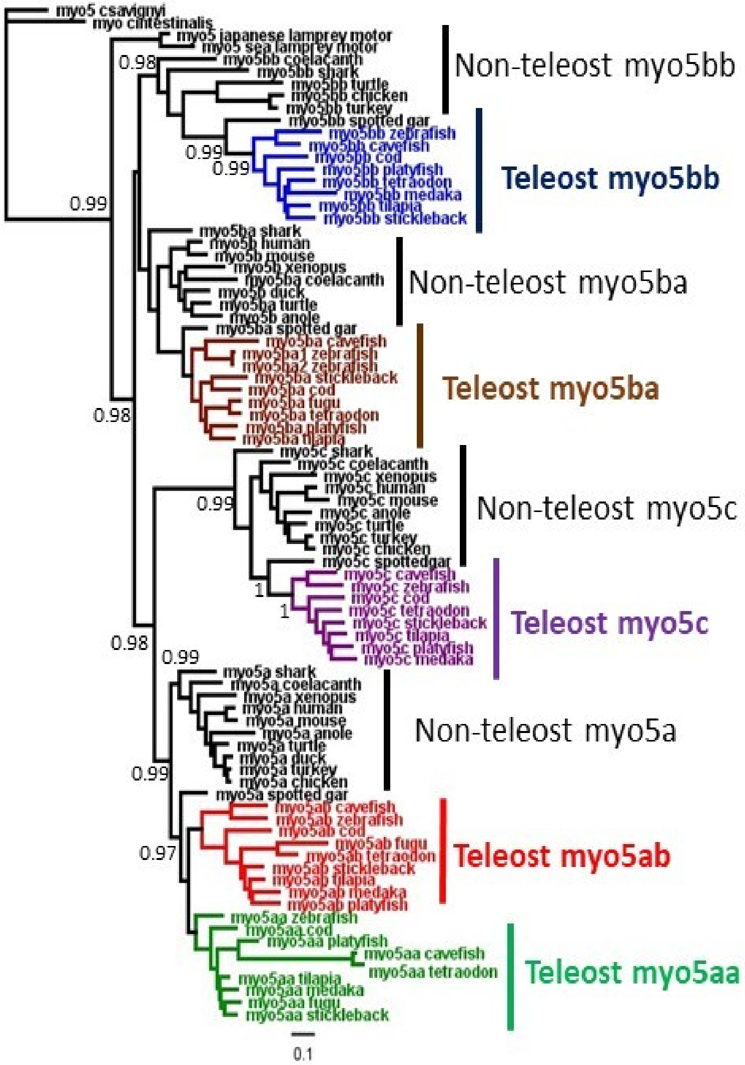


**Figure 6. Bayesian phylogenetic tree for *myo5* sequence encoding the neck region.** Phylogenetic tree for the neck and coiled coil domain (a 2505 bp fragment) of *myo5* using Mr.Bayes 3.1 and a GTR+I+G model of evolution. The region of the *myo5* gene used for this tree also includes a portion of the motor domain that includes the actin binding domain but excludes the ATP-binding domain. Teleost *myo5bb* clade shown in blue with an extended branch leading up to the clade. More of the branches are resolved compared to previously presented trees as a result of the diversity of the gene sequence in the neck region of the *myo5* gene family. Nodes without a posterior probability value are greater than 0.75 with most values being 1.

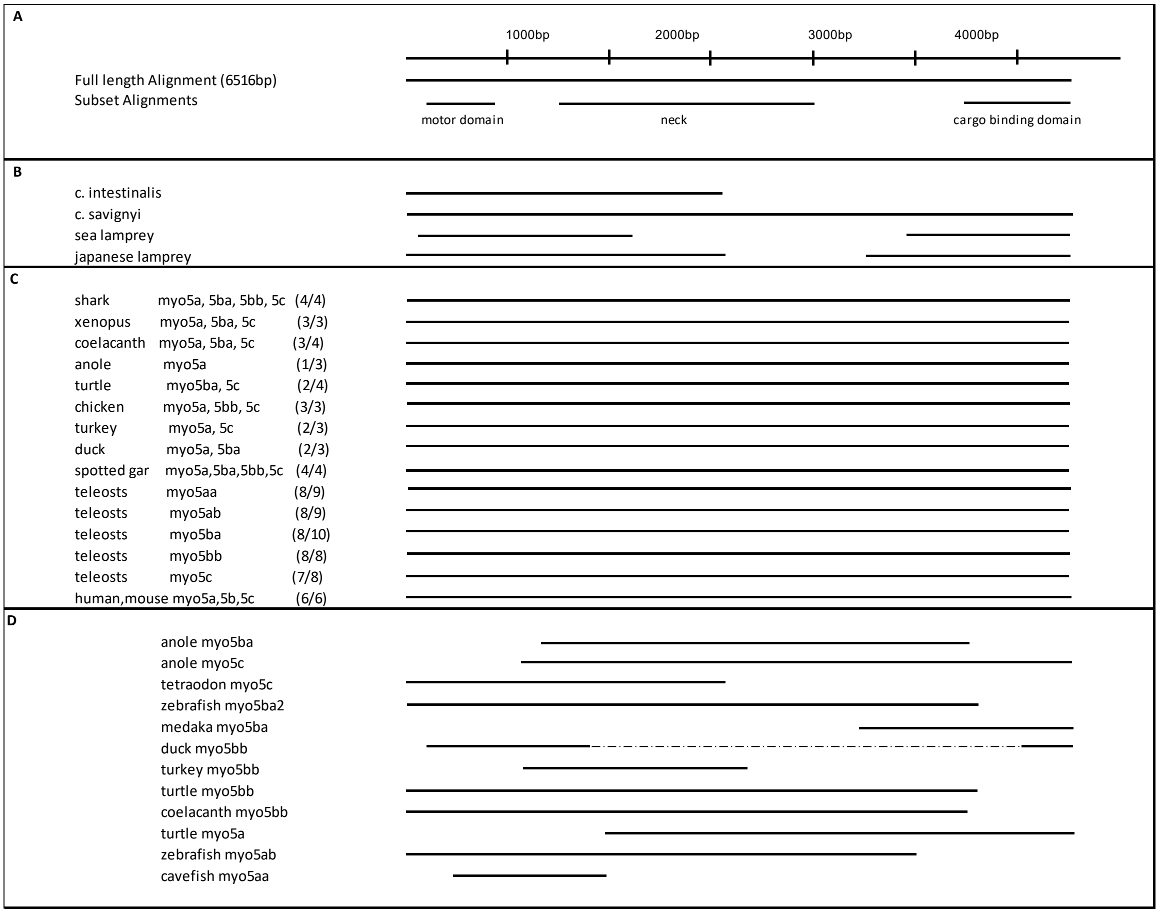


**Figure 7. Alignment of *myo5* sequences.** Panel A shows the full-length alignment size using 87 species along with the three smaller subsets (motor domain, neck, and cargo binding domain) that were used for further characterization of the *myo5* gene family. Panel B shows the smaller sequences found among lamprey and *C. intestinalis*. Panel C shows the sequences that are full length for the provided species or group of species with the first number in parentheses showing the number of full length sequences available for that species or group of species and the second number showing the total number of *myo5* sequences that have been found for that species or group of species. Panel D shows which sequences out of the total number of 87 sequences are truncated or missing some part of the full-length sequence.

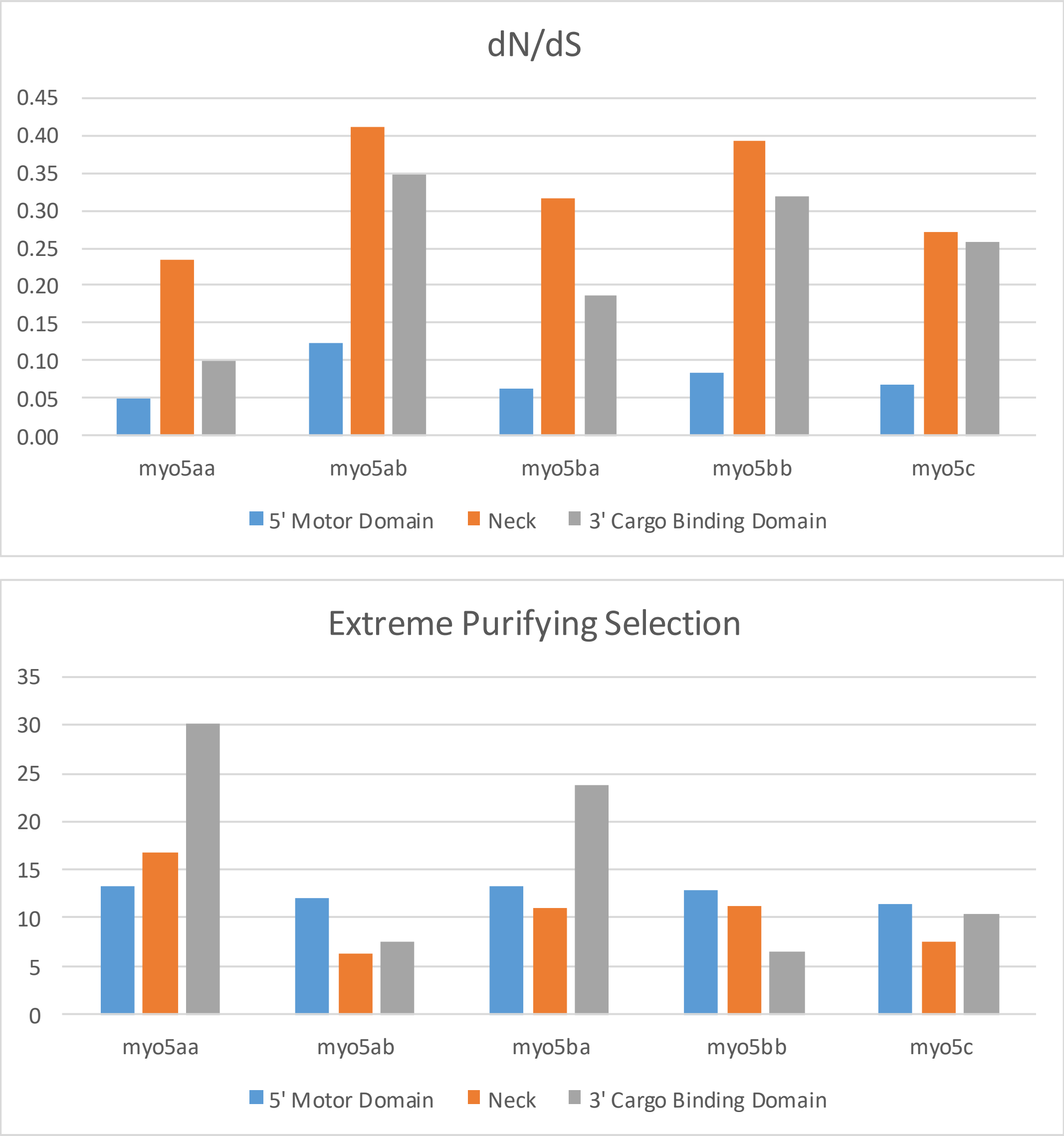


**Figure 8. dN/dS rates and percentage of invariant codons**.

**Figure 8. dN/dS rates and percentage of invariant codons**. dN/dS rates and percentage of codons that are invariant or under extreme purifying selection for all 5 *myo5* genes in teleosts. I see smaller dN/dS rates for *myo5aa* compared to *myo5ab* and for *myo5ba* compared to *myo5bb* for all cases using smaller regions of the *myo5* genes. The *myo5aa* gene has more invariant codons than its duplicate *myo5ab*. However, very similar percentages of invariant codons are observed for the motor and neck domain for the *myo5ba* and *myo5bb* duplicates. The *myo5ba* cargo binding domain has a much higher percentage of invariant codons compared to the paralogous *myo5bb* cargo binding domain, suggesting high conservation in the encoded protein as would be necessary to assure binding to the Rab11a cargo. The diversity seen in the *myo5bb* clade suggests that this duplicate has picked up a new function or ability to bind to other cargo.

**Table 1. Myosin 5 gene identifiers, isoforms, and locations.** List of organisms used in this study along with their Ensembl gene identifiers and their chromosomal or scaffold location.

| Organism | Ensembl Gene ID | myo5 isoform | chromosome #/ scaffold |
|---|---|---|---|
| Cave Fish | ENSAMXG00000003432 | 5aa | Scaffold KB872901.1 |
| Cave Fish | ENSAMXG00000017247 | 5ab | Scaffold KB871834.1 |
| Cave Fish | ENSAMXG00000003638 | 5b | Scaffold KB882129.1 |
| Cave Fish | ENSAMXG00000017029 | 5c | Scaffold KB871834.1 |
| Cod | ENSGMOG00000015730 | 5aa | GeneScaffold_3426 |
| Cod | ENSGMOG00000003963 | 5ab | GeneScaffold_1297 |
| Cod | ENSGMOG00000019264 | 5ba | GeneScaffold_691 |
| Cod | ENSGMOG00000008969 | 5bb | GeneScaffold_4456 |
| Cod | ENSGMOG00000003867 | 5c | GeneScaffold_1297 |
| Drerio | ENSDARG00000074622 | 5a | Chromosome 25: |
| Drerio | ENSDARG00000061635 | 5aa | Chromosome 18 |
| Drerio | ENSDARG00000025218 | 5ab | Chromosome 25 |
| Drerio | ENSDARG00000062003 | 5ba | Chromosome 21 |
| Drerio | ENSDARG00000061810 | 5bb | Chromosome 22 |
| Drerio | ENSDARG00000013782 | 5c | Chromosome 25 |
| Medaka | ENSORLG00000012865 | 5aa | Chromosome 3 |
| Medaka | ENSORLG00000005448 | 5ab | Chromosome 6 |
| Medaka | ENSORLG00000012866 | 5ba | ultracontig72 |
| Medaka | ENSORLG00000004814 | 5bb | Chromosome 4 |
| Medaka | ENSORLG00000005475 | 5c | Chromosome 6 |
| Platyfish | ENSXMAG00000018243 | 5aa | Scaffold JH556887.1 |
| Platyfish | ENSXMAG00000006730 | 5ab | Scaffold JH556946.1 |
| Platyfish | ENSXMAG00000013502 | 5ba | Scaffold JH557924.1 |
| Platyfish | ENSXMAG00000012695 | 5bb | JH556665.1 |
| Platyfish | ENSXMAG00000006735 | 5c | Scaffold JH556946.1 |
| Stickleback | ENSGACG00000016512 | 5aa | groupII |
| Stickleback | ENSGACG00000006025 | 5ab | groupXIX |
| Stickleback | ENSGACG00000013454 | 5ba | scaffold_196 |
| Stickleback | ENSGACG00000012760 | 5bb | groupVIII |
| Stickleback | ENSGACG00000006001 | 5c | groupXIX |
| Fugu | ENSTRUG00000008050 | 5aa | scaffold_174 |
| Fugu | ENSTRUG00000003810 | 5ba | scaffold_590 |
| Fugu | ENSTRUG00000006142 | 5ab | scaffold_198 |
| Tetraodon | ENSTNIG00000010232 | 5aa | Chromosome 5 |
| Tetraodon | ENSTNIG00000013708 | 5ab | Chromosome 13 |
| Tetraodon | ENSTNIG00000010628 | 5bb | Chromosome 1 |
| Tetraodon | ENSTNIG00000010379 | 5ba | Chromosome 4 |
| Tetraodon | ENSTNIG00000000200 | 5c | Chromosome 13 |
| Tilapia | ENSONIG00000002644 | 5aa | Scaffold GL831133.1 |

**Table 1. Myosin 5 gene identifiers, isoforms, and locations.**

| | | | |
|---|---|---|---|
| Tilapia | ENSONIG00000008315 | 5ab | Scaffold GL831160 |
| Tilapia | ENSONIG00000004825 | 5bb | GL831234.1 |
| Tilapia | ENSONIG00000013901 | 5b | Scaffold GL831403.1 |
| Tilapia | ENSONIG00000008313 | 5c | Scaffold GL831160.1 |
| Spotted Gar | ENSLOCG00000013374 | 5aa | Chromosome LG3 |
| Spotted Gar | ENSLOCG00000012796 | 5ba | Chromosome LG2 |
| Spotted Gar | ENSLOCG00000006065 | 5bb | Chromosome LG19 |
| Spotted Gar | ENSLOCG00000013362 | 5c | Chromosome LG4 |
| S. Lamprey | ENSPMAG00000003035 | 5-cbd | Scaffold GL476508 |
| S. Lamprey | ENSPMAG00000000443 | 5-motor | Scaffold GL479744 |
| J. Lamprey | J3181 | 5-cbd | Scaffold 00027 |
| J. Lamprey | JL3182 | 5-motor | Scaffold 00027 |
| Coelacanth | ENSLACG00000005276 | 5a | Scaffold JH128517.1 |
| Coelacanth | ENSLACG00000005498 | 5b | Scaffold JH127365 |
| Coelacanth | ENSLACG00000007284 | 5c | Scaffold JH128517.1 |
| Coelacanth | ENSLACG00000016232 | 5bb | Scaffold JH126673.1 |
| Human | ENSG00000197535 | 5a | Chromosome 15 |
| Human | ENSG00000167306 | 5b | Chromosome 18 |
| Human | ENSG00000128833 | 5c | Chromosome 15 |
| mouse | ENSMUSG00000034593 | 5a | Chromosome 9 |
| mouse | ENSMUSG00000025885 | 5b | Chromosome 18 |
| mouse | ENSMUSG00000033590 | 5c | Chromosome 9 |
| chicken | ENSGALG00000004624 | 5a | Chromosome 10 |
| chicken | ENSGALG00000012984 | 5bb | Chromosome 28 |
| chicken | ENSGALG00000004641 | 5c | Chromosome 10 |
| duck | ENSAPLG00000003054 | 5a | Scaffold KB746221.1 |
| duck | ENSAPLG00000005427 | 5b | Scaffold KB743255.1 |
| duck | ENSAPLG00000013765 | 5c | Scaffold KB745341.1 |
| turkey | ENSMGAG00000006197 | 5a | Chromosome 12 |
| turkey | ENSMGAG00000007708 | 5bb | Scaffold GL426528.1 |
| turkey | ENSMGAG00000006309 | 5c | Chromosome 12 |
| anole | ENSACAG00000010982 | 5a | Scaffold GL343479.1 |
| anole | ENSACAG00000017976 | 5b | Chromosome 2 |
| anole | ENSACAG00000029269 | 5c | Scaffold GL343479.1 |
| turtle | ENSPSIG00000009964 | 5a | Scaffold JH207514.1 |
| turtle | ENSPSIG00000007807 | 5c | Scaffold JH207514.1 |
| turtle | ENSPSIG00000005153 | 5ba | Scaffold JH212687.1 |
| turtle | ENSPSIG00000004377 | 5bb | Scaffold JH210437.1 |
| xenopus | ENSXETG00000020736 | 5a | Scaffold GL172839.1 |
| xenopus | ENSXETG00000020323 | 5b | Scaffold GL172853.1 |
| xenopus | ENSXETG00000020739 | 5c | Scaffold GL172853.1 |
| C.intestinalis | ENSCING00000002130 | 5 | Scaffold HT000045.1 |

**Table 1. Myosin 5 gene identifiers, isoforms, and locations.**

| | | | |
|---|---|---|---|
| C.savignyi | ENSCSAVG00000011586 | 5 | reftig_16 |
| Shark | XM_007906162.1 | 5a | NW_006890058.1 |
| Shark | XM_007906139.1 | 5c | NW_006890058.1 |
| Shark | XM_007906807.1 | 5ba | NW_006890225.1 |
| Shark | XM_007910183.1 | 5bb | NW_006890345.1 |

**Table 2. dN/dS average values for each clade**. dN/dS average values for each *myo5* clade from teleosts. dN/dS values are reflective of codon changes that lead to synonymous (S) or non-synonymous (N) codons. When comparing the *myo5ab* clade to the *myo5aa* clade there is a 33% increase in dN/dS values for the whole gene (1,915 codons) and 140%, 78%, and 250% increases for the 5′ end (217 codons), neck (742 codons), and 3′end (319 codons), respectively. Comparing the *myo5bb* clade to the *myo5ba* clade, I see a 58% increase in dN/dS values for the whole gene and 33%, 22%, and 68% for the 5′ end, neck, and 3′end, respectively. Additionally, *myo5c* evolves at a rate similar to *myo5aa*, which is found on a different chromosome. The *myo5c* gene evolves at a much slower rate than *myo5aa's* duplicate *myo5ab*. For the smaller subsets of codons encoding the ATP binding domain (21 codons), four of the five *myo5* genes in teleosts show higher levels of conservation than the larger 5′ region whereas the *myo5aa* clade has the same dN/dS value for the smaller subset of codons. For the smaller subset of codons related to actin binding (23 codons) there is strong conservation for both *myo5b* duplicates and for *myo5c*. For the smaller subset of codons encoding the cargo binding domain (10 codons), I see strong conservation for the *myo5ba* duplicate, suggesting the protein encoded likely binds to Rab11a, and the Myo5bb duplicate may bind to other cargo. There is not a value listed for *myo5c* and the smaller subset of ten codons in the cargo binding domain as it is unknown what amino acids are involved in this process for the orthologous *myo5c* in human.

| | whole gene | 5′ end & ATP binding | | neck & actin binding | | 3′ end & cargo binding domain | |
|---|---|---|---|---|---|---|---|
| | 1915 codons | 217 codons | 21 codons | 742 codons | 23 codons | 319 codons | 10 codons |
| *myo5aa* | 0.27 | 0.05 | 0.05 | 0.23 | 0.23 | 0.10 | 0.08 |
| *myo5ab* | 0.36 | 0.12 | 0.02 | 0.41 | 0.14 | 0.35 | 0.22 |
| *myo5ba* | 0.26 | 0.06 | 0.01 | 0.32 | 0.07 | 0.19 | 0.00 |
| *myo5bb* | 0.41 | 0.08 | 0.02 | 0.39 | 0.05 | 0.32 | 0.25 |
| *myo5c* | 0.26 | 0.07 | 0.00 | 0.27 | 0.04 | 0.26 | **** |
| Average | 0.31 | 0.08 | 0.02 | 0.32 | 0.11 | 0.24 | 0.14 |

**Table 3. Percentage increase for dN/dS rates for each clade.** Percentage increases for dN/dS rates between duplicated *myo5* genes and the three regions or domains within the *myo5* genes for teleosts. The dN/dS rates for *myo5aa* and *myo5ab* are compared in the top half of the table, and for *myo5ba* and *myo5bb* in the bottom half. Both *myo5ab* and *myo5bb* had higher dN/dS values than *myo5aa* and *myo5ba, respectively,* in all comparisons. The 3′ end has the highest percentage increase. This table compares data presented in Table 1.

| | % increase for dN/dS rates |
|---|---|
| Whole *myo5aa*↔*myo5ab* | 33 |
| 5′ end Motor Domain *myo5aa*↔*myo5ab* | 140 |
| Neck *myo5aa*↔*myo5ab* | 78 |
| 3′ end Cargo Binding Domain *myo5aa*↔*myo5ab* | 250 |
| Whole *myo5ba*↔*myo5bb* | 58 |
| 5′ end Motor Domain *myo5ba*↔*myo5bb* | 33 |
| Neck *myo5ba*↔*myo5bb* | 22 |
| 3′ end Cargo Binding Domain *myo5ba*↔*myo5bb* | 68 |

**Table 4. Percentage of invariant codons for each clade.** The percentage of invariant codons are shown for each teleost clade for each domain of each duplicate. The number of codons used in each alignment are shown along with the percentage of invariant sites (codons) for each alignment. For each clade there are 8-9 teleost sequences. For some of the regions there are large differences in the number of invariant sites found in the *myo5ab* clades compared to the *myo5aa* clades and when comparing the *myo5ba* clades to the *myo5bb* clades. The largest differences occur in the 3′ end of the *myo5* genes where the cargo binding domain is located. Extreme purifying selection is defined here as dN=dS=zero. No substitutions were identified in any of the 3 codon positions for these sites. For the cargo-binding domain (dilute domain) 30% of the codons for the teleost *myo5aa* clade showed extreme purifying selection, but only 7.5% of codons in the *myo5ab* clade showed extreme purifying selection. The data were generated using MEGA6 and HyPhy.

| | **Total codons** | **Invariants/Extreme Purifying Selection % codons where dN=dS=0** |
|---|---|---|
| *myo5aa* 5′ end | 217 | 13.4 |
| *myo5ab* 5′ end | 217 | 12.0 |
| *myo5ba* 5′ end | 217 | 13.4 |
| *myo5bb* 5′ end | 217 | 12.9 |
| *myo5c* 5′ end | 217 | 11.5 |
| | | |
| *myo5aa* neck | 728 | 16.8 |
| *myo5ab* neck | 742 | 6.2 |
| *myo5ba* neck | 748 | 11.0 |
| *myo5bb* neck | 734 | 11.2 |
| *myo5c* neck | 703 | 7.5 |
| | | |
| *myo5aa* 3′ end | 319 | 30.1 |
| *myo5ab* 3′ end | 322 | 7.5 |
| *myo5ba* 3′ end | 323 | 23.8 |
| *myo5bb* 3′ end | 336 | 6.6 |
| *myo5c* 3′ end | 327 | 10.4 |
| | | |
| *myo5aa* full length | 1908 | 16.7 |
| *myo5ab* full length | 1938 | 8.0 |
| *myo5ba* full length | 1904 | 13.6 |
| *myo5bb* full length | 1668 | 8.1 |
| *myo5c* full length | 1761 | 11.8 |

**Table 5. Summary of results for selection tests.** Summary of results from MEME (Mixed Effects Model of Evolution), REL (Random Effects Likelihood), and SLAC (Single Likelihood Ancestor Counting) hypothesis testing using HyPhy package from datamonkey.org. A large number of sites showing episodic diversifying selection in the neck region of the *myo5* gene are identified. The functional domains are in the motor domain and in the cargo binding domain (cbd). In the cbd I see a more episodic diversifying selection (MEME) in the *myo5bb* clade of teleosts versus the *myo5ba* clade of teleosts. I also see large variations between these two clades when comparing the REL results. The REL results show the number of sites (codons) experiencing positive (REL +) or negative/purifying (REL -) selection. Clades consisting of 8-9 sequences are only containing teleost sequences. The *myo5c* cbd clade consists of 8 teleost sequences and 6 non-teleost sequences. The clades with 22-25 sequences contain all the teleost sequences in that group (16-18 sequences) plus 6-8 non-teleost sequences.

| | | | MEME (# sites) | REL (# sites) | | SLAC (# of sites) | | SLAC |
|---|---|---|---|---|---|---|---|---|
| | # of sequences | Total codons | p<0.05 | REL + | REL - | SLAC + | SLAC - | dN/dS |
| motor-myo5a | 25 | 217 | 3 | 0 | 180 | 0 | 190 | 0.0830781 |
| motor-myo5aa | 9 | 217 | 2 | 0 | 217 | 0 | 88 | 0.058235 |
| motor-myo5ab | 9 | 217 | 0 | 2 | 183 | 0 | 99 | 0.081544 |
| motor-myo5b | 22 | 217 | 3 | 0 | 217 | 0 | 173 | 0.0627148 |
| motor-myo5ba | 8 | 217 | 1 | 0 | 217 | 0 | 82 | 0.03639 |
| motor-myo5bb | 8 | 217 | 1 | 0 | 217 | 0 | 81 | 0.048349 |
| motor-myo5c | 8 | 217 | 1 | 0 | 217 | 0 | 92 | 0.0424814 |
| neck-myo5a | 25 | 830 | 16 | 0 | 830 | 0 | 435 | 0.205097 |
| neck-myo5aa | 9 | 830 | 7 | 4 | 226 | 0 | 153 | 0.189734 |
| neck-myo5ab | 9 | 830 | 6 | 0 | 830 | 0 | 176 | 0.269787 |
| neck-myo5b | 25 | 830 | 26 | 1 | 408 | 1 | 404 | 0.223327 |
| neck-myo5ba | 9 | 831 | 20 | 0 | 242 | 0 | 168 | 0.192678 |
| neck-myo5bb | 8 | 830 | 17 | 0 | 377 | 0 | 159 | 0.241978 |
| neck-myo5c | 8 | 830 | 5 | 0 | 96 | 1 | 199 | 0.181709 |
| cbd-myo5a | 24 | 343 | 3 | 0 | 343 | 0 | 229 | 0.115414 |
| cbd-myo5aa | 8 | 343 | 1 | 2 | 132 | 0 | 94 | 0.0513929 |
| cbd-myo5ab | 8 | 343 | 1 | 1 | 103 | 0 | 73 | 0.210042 |
| cbd-myo5b | 23 | 343 | 5 | 2 | 180 | 0 | 198 | 0.178619 |
| cbd-myo5ba | 9 | 343 | 0 | 0 | 78 | 0 | 85 | 0.0993496 |
| cbd-myo5bb | 8 | 343 | 5 | 5 | 247 | 0 | 84 | 0.198824 |
| cbd-myo5c | 14 | 343 | 1 | 0 | 69 | 0 | 88 | 0.162034 |

**Table 6**. **Branch site REL results.** Branch Site-Random Effects Likelihood (BS-REL) test results. Using the BS-REL test through the Datamonkey server, the cargo binding domain and motor domain showed evidence of episodic diversifying selection. Twenty percent of the sites along the *myo5bb* cargo binding domain branch are subject to positive selection, 26% of the sites along the same branch are subject to neutral selection, and 54% of the sites along this branch are subject to purifying selection.

| | **Branch** | **p-value** | **Positive** | **Neutral** | **Purifying** |
|---|---|---|---|---|---|
| Motor Domain | myo5bb clade | 0.022 | 0.08 | 0.34 | 0.58 |
| Cargo Binding Domain | myo5b clade | 0.014 | 0.13 | 0.03 | 0.84 |
| | myo5ba clade | 0.014 | 0.04 | 0.03 | 0.93 |
| | myo5bb clade | 0.04 | 0.2 | 0.26 | 0.54 |

**Table 7. Rates of evolution of *myo5* sites encoding amino acids involved in ATP-binding.** Codon specific evolutionary rate comparisons between *myo5a* duplicates and between *myo5b* duplicates for the sites that are known to be involved in ATP binding and motor activity in *myo5a*. The column on the left shows the amino acid and site number based on the human *myo5a* ortholog. For the 21 amino acids that are listed in this table that have been shown to play a role in ATP binding or regulation of motor activity for Myo5, most of the corresponding codons have zero non-synonymous substitutions as represented by the dN column. We see lower dN/dS rates for all the *myo5* duplicates compared to the larger regions of the gene. For dN/dS values close to zero, this is a sign of purifying selection. We infer that the motor domains would still be functional and able to bind to ATP for all four *myo5a* and *myo5b* duplicates. Residues in switch I contact the ATP/Mg complex and may change conformation when no cofactor is bound

| | Hum. Myo 5a | dS | dN | dS | dN | dS | dN | dS | dN | dN/dS | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 5aa | | 5ab | | 5ba | | 5bb | | 5aa | 5ab | 5ba | 5bb |
| Motor activity | D134 | 0 | 0.8 | 3.28 | 0.42 | 0 | 0 | 1.67 | 0 | Un | 0.1 | | |
| | D136 | 1.6 | 0.9 | 3.28 | 0 | 1.69 | 0.4 | 1.84 | 0.4 | Un | 0 | | |
| MYO 5A ATP binding P-loop 163-170 | G163 | 1.0 | 0 | 3.00 | 0 | 3.00 | 0 | 2.00 | 0 | 0 | 0 | 0 | 0 |
| | E164 | 3.5 | 0 | 3.22 | 0.45 | 3.44 | 0 | 3.73 | 0 | 0 | 0.1 | 0 | 0 |
| | S165 | 4.0 | 0 | 3.00 | 0 | 3.00 | 0 | 3.00 | 0 | 0 | 0 | 0 | 0 |
| | G166 | 4.0 | 0 | 5.00 | 0 | 1.00 | 0 | 2.00 | 0 | 0 | 0 | 0 | 0 |
| | A167 | 2.0 | 0 | 4.00 | 0 | 4.00 | 0 | 3.00 | 0 | 0 | 0 | 0 | 0 |
| | G168 | 1.0 | 0 | 2.00 | 0 | 4.00 | 0 | 4.00 | 0 | 0 | 0 | 0 | 0 |
| | K169 | 1.7 | 0 | 4.83 | 0 | 3.53 | 0 | 0 | 0 | 0 | 0 | 0 | Un |
| | T170 | 2.0 | 0 | 2.00 | 0 | 2.00 | 0 | 3.00 | 0 | 0 | 0 | 0 | 0 |
| ATP binding Switch 1 region 209-219 | A209 | 3.0 | 0 | 1.00 | 0 | 4.00 | 0 | 3.00 | 0 | 0 | 0 | 0 | 0 |
| | K210 | 0 | 0 | 1.59 | 0 | 3.37 | 0 | 1.86 | 0 | Un | 0 | 0 | 0 |
| | T211 | 3.0 | 0 | 4.00 | 0 | 3.00 | 0 | 2.00 | 0 | 0 | 0 | 0 | 0 |
| | T212 | 3.0 | 0 | 3.37 | 0.47 | 3.00 | 0 | 2.02 | 0.5 | 0 | 0.1 | 0 | 0.3 |
| | R213 | 5.5 | 0.5 | 5.54 | 0 | 3.85 | 0 | 3.54 | 0 | 0.1 | 0 | 0 | 0 |
| | N214 | 4.8 | 0 | 4.67 | 0 | 4.41 | 0 | 1.52 | 0 | 0 | 0 | 0 | 0 |
| | D215 | 0 | 0 | 0 | 0 | 0 | 0 | 1.78 | 0 | Un | Un | Un | 0 |
| | N216 | 0 | 0 | 0 | 0 | 1.71 | 0 | 0 | 0 | Un | Un | 0 | Un |
| | S217 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | Un | Un | Un | Un |
| | S218 | 4.8 | 0 | 4.91 | 0 | 3.27 | 0 | 3.54 | 0 | 0 | 0 | 0 | 0 |
| | R219 | 2.0 | 0 | 1.97 | 0 | 1.00 | 0 | 3.96 | 0 | 0 | 0 | 0 | 0 |
| | Sum | 47 | 2.2 | 60.7 | 1.3 | 53.3 | 0.4 | 47.5 | 0.9 | | | | |
| | dN/dS | 0.047 | | 0.022 | | 0.008 | | 0.020 | | 0.05 | 0.02 | 0.01 | 0.02 |

**Table 8. Rates of evolution of *myo5* sites encoding amino acids involved in actin-binding.** Codon specific evolutionary rate comparisons between the *myo5a* duplicates (*myo5aa*-*myo5ab*) and between the *myo5b* duplicates (*myo5ba*-*myo5bb*) for the sites that are known to be involved in actin binding. When dN=dS=0, the dN/dS ratio has an undefined value. These invariant sites are under extreme purifying selection. U=Undefined

| Human 5a | myo5aa | | myo5ab | | myo5ba | | myo5bb | | dN/dS | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | dS | dN | dS | dN | dS | dN | dS | dN | 5aa | 5ab | 5ba | 5bb |
| H644 | 7.4 | 0.62 | 3.89 | 1.62 | 4.16 | 0.4 | 2.02 | 0 | 0.08 | 0.42 | 0.1 | 0 |
| L645 | 2.71 | 1.44 | 3.47 | 0.62 | 2.77 | 0 | 5.09 | 0.62 | 0.53 | 0.18 | 0 | 0.12 |
| L646 | 0.4 | 1.12 | 0.69 | 0 | 3.45 | 0 | 4.86 | 0 | 2.79 | 0 | 0 | 0 |
| M647 | 2.41 | 1.13 | 0 | 0 | 0 | 0 | 0 | 0 | 0.47 | U | U | U |
| E648 | 2.35 | 2.49 | 0 | 1.2 | 0 | 0.42 | 0 | 0.44 | 1.06 | U | U | U |
| T649 | 3.61 | 0.58 | 4.26 | 0.97 | 3 | 0 | 4.41 | 0.55 | 0.16 | 0.23 | 0 | 0.13 |
| L650 | 4.61 | 0 | 1.37 | 0 | 1.53 | 0 | 2.49 | 0 | 0 | 0 | 0 | 0 |
| N651 | 4.42 | 0 | 6.33 | 0 | 2.2 | 0 | 0 | 0.44 | 0 | 0 | 0 | U |
| A652 | 2.33 | 0.58 | 5 | 0.5 | 4 | 0 | 0 | 0.44 | 0.25 | 0.1 | 0 | U |
| T653 | 4.65 | 0 | 2 | 0 | 1 | 0 | 1.1 | 0 | 0 | 0 | 0 | 0 |
| T654 | 3.6 | 1.15 | 2.24 | 0.95 | 4 | 0 | 2.21 | 0 | 0.32 | 0.42 | 0 | 0 |
| P655 | 4.65 | 0.58 | 1 | 0 | 1 | 0 | 1.1 | 0 | 0.13 | 0 | 0 | 0 |
| H656 | 6.61 | 0 | 2.18 | 0 | 4.23 | 0 | 2.34 | 0 | 0 | 0 | 0 | 0 |
| Y657 | 5.45 | 1.43 | 6.06 | 0 | 6.6 | 0 | 2.22 | 0 | 0.26 | 0 | 0 | 0 |
| V658 | 5.82 | 0.58 | 1 | 0 | 4 | 0 | 2.21 | 0 | 0.1 | 0 | 0 | 0 |
| R659 | 1.16 | 1.2 | 0.99 | 0 | 0 | 0 | 1.1 | 0 | 1.04 | 0 | U | 0 |
| C660 | 2.12 | 1.05 | 0 | 0 | 2.23 | 0 | 4.67 | 0 | 0.5 | U | 0 | 0 |
| I661 | 4.67 | 0 | 1.35 | 0 | 1.33 | 0 | 6.07 | 0 | 0 | 0 | 0 | 0 |
| K662 | 2.09 | 1.01 | 5.28 | 0 | 6 | 0 | 2.15 | 0 | 0.48 | 0 | 0 | 0 |
| P663 | 1.16 | 1.18 | 3 | 0 | 1 | 0 | 0 | 0.55 | 1.02 | 0 | 0 | U |
| N664 | 4.2 | 0 | 4.04 | 0 | 4.23 | 0 | 6.95 | 0 | 0 | 0 | 0 | 0 |
| D665 | 2.21 | 0.47 | 2.13 | 0 | 4.35 | 0 | 2.35 | 0 | 0.21 | 0 | 0 | 0 |
| E666 | 2.05 | 1.98 | 3.71 | 2.46 | 5.33 | 3.56 | 3.05 | 0 | 0.96 | 0.66 | 0.67 | 0 |
| Sum | 80.68 | 18.59 | 59.99 | 8.33 | 66.41 | 4.39 | 56.39 | 3.03 | | | | |
| dN/dS | 0.23 | | 0.139 | | 0.066 | | 0.054 | | 0.23 | 0.14 | 0.07 | 0.05 |

**Table 9. Rates of evolution of *myo5* sites encoding amino acids involved in Mlph-binding.** Codon specific evolutionary rate comparisons between the *myo5a* duplicates (*myo5aa-myo5ab*) for the sites that are known to be involved in melanophilin (Mlph) binding. When dN=dS=0, the dN/dS ratio has an undefined value. These invariant sites are under extreme purifying selection.

| Human Myo5a | myo5aa | | myo5ab | | dN/dS | |
|---|---|---|---|---|---|---|
| | dS | dN | dS | dN | 5aa | 5ab |
| I1535 | 0 | 0 | 0 | 0 | undefined | Undefined |
| F1562 | 3.61 | 0 | 4.53 | 1.19 | 0 | 0.26 |
| L1588 | 0.69 | 0 | 2.12 | 0 | 0 | 0 |
| T1589 | 1 | 0.5 | 1.59 | 2.72 | 0.5 | 1.71 |
| N1590 | 0 | 0 | 4.22 | 0 | undefined | 0 |
| F1591 | 2.95 | 0 | 1.85 | 0 | 0 | 0 |
| D1592 | 0 | 0 | 5.82 | 0.42 | undefined | 0.07 |
| E1595 | 0 | 0 | 2.69 | 0.66 | undefined | 0.24 |
| Y1596 | 0 | 0 | 2.22 | 0.48 | undefined | 0.22 |
| L1600 | 0.68 | 0.66 | 3.08 | 1.03 | 0.97 | 0.33 |
| sum | 8.93 | 1.16 | 28.11 | 6.5 | | |
| dN/dS | 0.13 | | 0.231 | | 0.13 | 0.231 |

**Table 10. Rates of evolution of *myo5b* sites encoding amino acids involved in binding to Rab11a.** Codon specific evolutionary rate comparisons between the *myo5b* duplicates (*myo5ba* and *myo5bb*) for the sites that are known to be involved in Rab11a binding. When dN=dS=0, the dN/dS ratio has an undefined value. These invariant sites are under extreme purifying selection.

| human myo5b (5ba/5bb) | myo5ba | | myo5bb | | dN/dS | |
|---|---|---|---|---|---|---|
| | dS | dN | dS | dN | 5ba | 5bb |
| W1706 (W/C) | 0 | 0 | 0 | 0.48 | undefined | undefined |
| M1710 | 0 | 0 | 3.31 | 2.16 | undefined | 0.65 |
| Y1714 | 0 | 0 | 4.7 | 0 | undefined | 0 |
| E1721 | 1.91 | 0 | 2.19 | 0 | 0 | 0 |
| R1724 | 5.64 | 0 | 3.47 | 0 | 0 | 0 |
| Q1745 | 0 | 0 | 2.19 | 0 | undefined | 0 |
| Q1748 | 1.9 | 0 | 6.52 | 0 | 0 | 0 |
| L1749 (V/ M,I) | 2 | 0 | 2.57 | 1.35 | 0 | 0.52 |
| K1750(K/ S,K) | 0 | 0 | 1.9 | 1.6 | undefined | 0.84 |
| L1763 (L/T) | 2.19 | 0 | 1.71 | 1.65 | 0 | 0.96 |
| Sum | 13.63 | 0 | 28.57 | 7.23 | | |
| dN/dS | 0 | | 0.253 | | 0 | 0.253 |

# III. EVOLUTION OF SELECT *RAB11* DUPLICATES

## Abstract

I identify and describe the evolutionary history and relationships among new *rab11* genes. *rab11* is a *ras*-related oncogene that is ubiquitously expressed, and its protein product Rab11 has a primary role in membrane trafficking. There are three known Rab11 proteins, Rab11a, Rab11b, and Rab25 (also known as Rab11c) that have been characterized, but duplicates of these genes, which are present in numerous non-mammalian vertebrates, have not been studied, much less identified. Here, I identify five *rab11* genes (three of which are newly identified) present in fish: *rab11aa, rab11a1, rab11ba1, rab11ba2*, and *rab11bb.* I characterize the evolutionary history of these genes, finding evidence of selection, gene conversion between *rab11ba1* and *rab11ba2* and coevolution between *myo5bb* genes and *rab11a1* genes.

Duplicated genes were characterized using phylogenetics, syntenic analysis, dN/dS rate comparisons, and intron sequence comparisons. I found all five genes to be highly conserved. The 18 codons associated with specific binding motifs have an average dN/dS rate of 0.01 when comparing all 68 sequences from 24 different species and a taxonomic sampling ranging from yeast to human. Additionally, I identify a highly conserved intron between exons 2 and 3 that is entirely conserved between paralogs *rab11ba1* and r*ab11ba2* for the entire intron length of 77 base pairs in tetraodon and this same intron only has 1 nucleotide difference between paralogs for fugu (77 base pairs) and medaka (80 base pairs). Amazon molly also shows high conservation for this intron between paralogous sequences with an intron length of 266 base pairs. The high degree of sequence conservation in both coding and non-coding regions, the phylogenetic

relationships, and the location for these genes near each other on the same chromosome leads us to suspect gene conversion to have taken place in the evolutionary history of these fish. Alignments of all five *rab11* genes for eight teleosts show evidence of extreme purifying selection with a range of 18% to 34% of codons having zero synonymous or non-synonymous substitutions and remaining invariant among fish that diverged approximately 150 million years ago.

I initially questioned whether one or more of these genes in fish would show signs toward becoming a pseudogene or alternatively if these duplicates seemingly remained highly conserved and possibly still functional. Based on the data I present, it seems possible that these duplicated genes have retained their functionality.

## Introduction

Rab proteins are members of a superfamily of Ras-related proteins which are all small GTPase proteins. The Ras proteins were originally studied and identified as contributing to Rat sarcomas. The Rab proteins were identified as being related to Ras proteins and they were initially identified in the brain such that Ras related protein in the brain became Rab. In addition to Ras and Rab there are three other protein subfamilies in this GTPase superfamily, namely Rho, Ran and Arf. The general functional association for each of these families include cell proliferation for Ras, cell morphology for Rho, nuclear transport for Ran, and vesicle transport for Rab and Arf. Rab proteins make up the largest family of the five subtypes of Ras proteins (Stenmark and Olkkonen 2001).

Many Rab proteins have been identified and studied, including 66 Rab proteins in humans and ten Rab proteins in simple eukaryotes (Diekmann et al. 2011; Pereira-Leal 2008; Zhang et al. 2007). Many of the genes encoding these Rab proteins arose as a result of gene or genome duplication events, some of which are thought to have occurred over 500 million years ago. A common fate of a duplicated gene is to become a non-functional pseudogene. However, sometimes the duplicated gene evolves such that new functions or new patterns of expression are identified. In teleosts, a group of ray finned fish that account for approximately 96% of all ray finned fish, a fish-specific genome duplication event is thought to have occurred between 300-350 million years ago. Due to the presence of duplicated genes in fish that are not present in other vertebrates, evolutionary rate comparisons between fish and non-fish genes may provide insight into the functionality of these duplicated genes.

The 3' end of the *myo5* gene encodes a cargo binding domain that aids in binding to accessory proteins (including Rab proteins) which further aid in binding to specific cargo. In chapter 1, the evolutionary history of myosin 5 genes was described, and evolutionary rate variations were observed among different domains in duplicated myosin 5 genes. One of the duplicated myosin 5 genes in teleosts showed signs of episodic diversifying selection and generally had a faster rate of evolution compared to paralogous genes. Episodic diversifying selection allows codons to be placed in one, two, or three different rate classes ($\omega$) along a branch of a phylogeny (Smith et al. 2015). If one of those rate classes is greater than one by a statistically significant amount, then that branch is subject to episodes of diversifying selection. In episodic diversifying selection, some of the codons along a branch are placed in a rate category subject to positive selection ($\omega>1$), whereas the remainder of the codons are experiencing one or two rate classes of purifying selection ($\omega<1$).

One of the accessory proteins that has been shown to interact with Myo5b in human and mice is a Rab11a protein (Pylypenko et al. 2013). Considering the evolutionary rate variations identified in the teleost *myo5bb* clade and that Rab11 is known to interact with Myo5b, the question I seek to address is whether the duplicated *rab11* genes show any similar patterns of evolutionary rate variation among the duplicates. If the *myo5bb* gene evolved at a faster rate than its *myo5ba* duplicate, then one of the *rab11a* duplicates might also have evolved at a faster rate. Since there are still highly conserved codons linked with functionally significant amino acids in both *myo5bb* and *rab11a1* genes, I suspect these genes are retaining functionality and that these gene families are coevolving.

Here I identify five teleost *rab11* clades (*rab11a1*, *rab11aa*, *rab11ba*, *rab11ba1* and *rab11bb*); previously only two *rab11* genes (*RAB11a* and *RAB11b*) had been described and they had been described in non-teleost vertebrates. I found a large amount of purifying selection to be present for all the *rab11* clades but there were obvious differences with dN/dS values ranging from 0.01 to 0.09 for individual *rab11* clades. I traced the evolutionary history of the *rab11* genes back to ancestral teleost and ancestral vertebrate chromosomes. For two of the *rab11* clades (*rab11ba1* and *rab11ba2*), I suspect gene conversion contributed to a high amount of conservation in several intron sequences. I found that intron 2/3 for *rab11ba1* and *rab11ba2* for three of the teleost species was the same in sequence or varied by one nucleotide for the entire 77-80 base pairs of sequence. I utilized selection assays to determine whether episodic selection was taking place along specific branches of my phylogeny. In addition, I suspect coevolution to be taking place between the *rab11a1* clades and the *myo5bb* clades based on higher dN/dS ratios for these clades in addition to high conservation rates for codons associated with amino acids where the two proteins interact.

## Materials and Methods

### Sequence acquisition

I collected *rab11* sequences using the Ensembl genomic database (Ensembl Release 86). The following species and genomic assemblies were used for *rab11* sequence downloads: ten teleost species (Amazon molly, *Poecilia formosa*, Poecilia_formosa-5.1.2; cavefish, *Astyanax mexicanus,* AstMex102; cod, *Gadus morhua,* gadMor1; fugu, *Takifugu rubripes,* FUGU 4.0; medaka, *Oryzias latipes,* HdrR; platyfish, *Xiphophorus maculatus,* Xipmac4.4.2; stickleback, *Gasterosteus aculeatus,* BROAD S1; tetraodon, *Tetraodon nigroviridis,* TETRAODON 8.0; tilapia, *Oreochromis niloticus,* Orenil1.0; zebrafish, *Danio rerio,* GRCz10), one holostean fish (spotted gar, *Lepisosteus oculatus,* LepOcu1), one lobe finned fish (coelacanth, *Latimeria chalumnae,* LatCha1), one amphibian (western clawed frog, *Xenopus tropicalis,* JGI 4.2), seven sauropsids (chicken, *Gallus gallus,* Gallus_gallus-5.0; turkey, *Meleagris gallopavo* , Turkey_2.01 ; duck, *Anas platyrhynchos*, BGI_duck_1.0;  zebrafinch, *Taeniopygia guttata,* taeGut3.2.4, flycatcher, *Ficedula albicollis,* FicAlb_1.4, Chinese soft shell turtle, *Pelodiscus sinensis*, PelSin_1.0;green anole lizard, *Anolis carolinensis,* AnoCar2.0), three mammals (human, *Homo sapiens,* GRCh38.p7; mouse, *Mus musculus,* GRCm38.p5, opossum, *Monodelphis domestica,* monDom5), one insect (fruitfly, *Drosophila melanogaster*, BDGP6), one roundworm (*Caenorhabditis elegans*, WBcel235), one jawless vertebrate (sea lamprey, *Petromyzon marinus,* Pmarinus_7.0), and one urochordate (sea squirt, *Ciona intestinalis,* KH) and one fungus (yeast, *Saccharomyces cerevisiae*, R64-1-1). Sequence identifiers for each species are listed in Table 16.

### Syntenic analysis

Using Biomart in the Ensembl database, genes located within 1.5 megabases of each *rab11* gene were identified. Synteny maps were constructed based on conserved patterns of gene locations for each of the *rab11* gene families. Genomic data from ten teleosts along with representative bird, amphibian, reptile, and mammal genomes were mined for each *rab11* gene family.

### Alignment and phylogenetics

Sixty-eight sequences were aligned using ClustalW and Geneious Pro 6.0 (Biomatters Ltd). Sequences were virtually translated, verified to contain open reading frames, aligned based on the amino acid sequence, and then reverted to the nucleic acid sequence for further analysis. Model testing was performed for each of the four alignments (Table 12), and the model with the best AICc value was chosen for the generation of the phylogenetic trees using Geneious 6.0. Using Mr.Bayes 3.1 and a GTR+I+G model of evolution, trees were generated for the full length coding sequence (699 bp) of *rab11*. The parameters used in the Mr. Bayes-generated trees included three gamma categories with unconstrained branch lengths. Markov Chain Monte Carlo methods were used for 1,100,000 steps with thinning every 200 steps, four heated chains, and a preheated chain temperature of 0.2. A burn in length of 500 steps was used. Alternative models were tested using maximum likelihood (10,000 bootstrap replicates) and parsimony methods, and these provided similar topologies. Figure 9 shows the final tree generated for the alignment.

### dN/dS rates and identification of invariant codons

To determine the percentage of codons that are invariant and experiencing extreme purifying selection, I calculated dN and dS values for the original alignment using MEGA6. “dN” is defined as the ratio of non-synonymous substitutions (n) per non-synonymous site (N); “dS” is defined as the ratio of synonymous substitutions (s) per synonymous site (S). Synonymous and non-synonymous substitutions are based on the specific sequences that are in my alignment. Synonymous and non-synonymous sites are based on the possible changes in the three positions in a codon such that when a nucleotide in a codon is changed it either changes the amino acid (making that a non-synonymous site) or it doesn’t change the amino acid (making that a synonymous site). Maximum likelihood reconstructions were generated using a Muse-Gaut model (Muse and Gaut 1994) of codon substitution and a general time reversible model (Nei and Kumar 2000) for nucleotide substitution. I counted the number of codons in an alignment that had dN and dS values of zero and divided this by the total number of codons in the alignment to determine the percentage of codons that are invariant and experiencing extreme purifying selection.

### Selection Tests

I used the Datamonkey server and the HyPhy software package (Delport et al. 2010; Kosakovsky Pond et al. 2005) to test for purifying selection, positive selection, and episodic selection at the codon level and the branch level among the phylogenies I generated. I used BUSTED (Branch site Unrestricted Statistical Test for Episodic Diversification) to assess whether episodic diversification occurs on at least one branch and at least at one site in the phylogeny. The BUSTED test allows for varying rates of evolution (ω) applied to a constrained model of selection (null model) and an unconstrained model of selection (alternative model) using a Likelihood Ratio Test

(LRT). I tested my alignments using MEME (Mixed Effects Model of Evolution) and aBS-REL (adaptive Branch Site Random Effects Likelihood) tests. MEME identifies the number of sites (codons) showing episodic diversifying selection. Different evolutionary rates are allowed for each codon within an alignment. Trees that were generated as described previously using the Geneious Software package were saved as Nexus files and uploaded to the Datamonkey Server to run the selection tests. Methods for the tests I used in my analyses are further described in Murrell *et al.*, 2012 (MEME), Murrell *et al.*, 2015 (BUSTED), Smith *et al.*, 2015 (aBS-REL). The aBS-REL test determined which branches in the phylogeny showed evidence of episodic diversifying selection using a likelihood ratio test. Branches in a phylogeny are allowed one, two, or three rate classes. Sites along a branch are then subject to being placed in one of those rate classes. Branches that show positive selection for a percentage of the codons (sites) on that branch with statistical significance are provided upon completing the test for selection.

## Results

### Phylogenetics and synteny

New *rab11* genes were identified using phylogenetics and syntenic analyses. I identified two clades of *rab11* genes specific to teleosts, namely the *rab11ba2* clade and the *rab11bb* clade. I also identified a new clade of *rab11* genes (*rab11a1*) in both teleosts and several non-teleosts, including coelacanth, birds, reptiles, and spotted gar. Phylogenetic analysis failed to resolve *rab11ba1* and *rab11ba2* to separate clades (Figure 9). Most species of fish had both genes, suggesting it arose in a common ancestor to these species; however, for several species of fish, the two genes were monophyletic, suggesting subsequent gene conversion. These two genes are found on the same chromosome or linkage group with the *march2a* gene between the two *rab11* genes for six of the ten teleost species examined. Both copies of the two *rab11ba* genes are found in fugu, tetraodon, tilapia, cod, medaka, and amazon molly. One copy (either *rab11ba1* or *rab11ba2*) was present in stickleback, zebrafish, and cavefish. In contrast, no copies of *rab11ba1* or *rab11ba2* were detected in platyfish (Table 16). I suspect that high conservation among the 699 base pairs tested led to poor levels of support at some of the nodes (posterior probability < 0.8). The placement of *Ciona,* coelacanth, and spotted gar with low levels of nodal support could be due to the high conservation for this short set of sequences. Additionally, human and mouse *rab11b* do not seem to be placed properly and there is low support (posterior probability=0.53) for that node. The cavefish and zebrafish *rab11ba* genes are another example of sequences that seem misplaced and this could be due to the high conservation for these short alignments.

Synteny was observed among non-teleost vertebrates (boxed regions in Figure 10A) and among teleosts (boxed regions in Figure 10B). The chromosomes for these organisms have previously been mapped back to ancestral vertebrate (Nakatani *et al*., 2007) or ancestral teleost chromosomes (Bian *et al*., 2016). The *rab11ba1* and *rab11ba2* genes were identified on chromosome 4 in medaka with a *march 2a* gene in between. Syntenic regions were identified in other fish including Amazon molly, cod, tilapia, tetraodon, and fugu with a general gene order of *pip5k1ca*, *hnrnpm, rab11ba1*, *march2a*, and *rab11ba2*. Synteny was observed between the *rab11ba* genes and the *rab11bb* genes. Newly identified *rab11bb* genes in teleosts generally showed the order of neighboring genes as *pip5k1cb*, *rab11bb*, and *march2b*. Synteny was observed in non-teleost vertebrates with the *rab11b* gene neighboring the *march2* gene (red boxed regions in Figure 10).

**Ancestral chromosome mapping**

The *rab11* genes were mapped back to ancestral vertebrate chromosomes for human, chicken, spotted gar, medaka, and ancestral teleosts (Figure 10A). In Nakatani *et al.*, 2007, ancestral vertebrate and ancestral teleost chromosomes are mapped along with extant species of medaka, human, and chicken. Bian *et al*. provided chromosomal maps for zebrafish and spotted gar showing how the chromosomes for these species map back to the chromosomes of an ancestral teleost. For medaka and zebrafish, *rab11* genes were also mapped back to ancestral teleost chromosomes (Figure 10B). All three teleost *rab11b* genes mapped back to ancestral teleost chromosome m. This region mapped back to ancestral vertebrate chromosome segment A1 (red boxed regions in Figure 10). For the *rab11aa* genes in teleosts, this region mapped back to ancestral teleost chromosome j (Figure 10B), and this region further mapped back to ancestral vertebrate chromosome segment A4 (purple boxed regions in Figure 10). For the newly identified *rab11a1* gene family in teleosts, a *rab25b* gene was identified directly next to the *rab11a1* gene. On paralogous chromosomal regions, a *rab25a* gene was identified, and both regions mapped back to ancestral teleost chromosome b and ancestral vertebrate chromosome segment A5 (green boxed regions in Figure 10).

**dN/dS and invariant codons**

I determined the dN/dS rates for each of the five *rab11* clades in addition to the dN/dS rate for all 68 sequences (Table 13) and the dN/dS rate for 18 codons that code for amino acids that have been linked with a functional binding role in Rab11 proteins (Table 14). The *rab11* clade for teleosts with the lowest dN/dS ratio is the *rab11aa* clade. The *rab11aa* clade has a dN/dS value of 0.010 based on 215 codons and eight teleost sequences. dN/dS rates for the eighteen codons linked with functionally significant amino acids were 0 or 0/0 for each codon in each of the five clades. For these codons, the dN value and, more specifically, the number of non-synonymous substitutions was zero. Depending on the clade, there were three to six out of the eighteen codons with dS values and, more specifically the number of synonymous substitutions, that were also zero.

The *rab11ba1* clade has the highest percentage of invariant codons (Table 15) with 34.4% of the 218 codons having zero substitutions. There were 75 codons out of 218 codons in the *rab11ba1* clade that had zero synonymous or non-synonymous substitutions. The least conserved *rab11* clade in teleosts is the *rab11a1* clade based on both dN/dS values and the percentage of codons that are invariant. The *rab11a1* clade has a dN/dS value of 0.091, and 18.3% of the 202 codons were invariant among the nine species analyzed.

**Intron sequence conservation**

Several introns exhibited a highly conserved length across numerous teleost species for the *rab11ba1* and *rab11ba2* clades. Moreover, the sequence of Intron 2/3 was also highly conserved in sequence for paralogous *rab11ba1* and *rab11ba2* genes in fugu, tetraodon, and medaka (Figure 11). The number of nucleotide differences between paralogous *rab11ba1* and *rab11ba2* sequences in Intron 2/3 was zero out of 77 bases for tetraodon, one out of 77 bases for fugu, and one out of 80 bases for medaka (Table 11).

**Selection test results**

Using the BUSTED (Branch site Unrestricted Statistical Test for Episodic Diversification) selection assay, I see evidence of episodic diversifying selection in at least one codon on at least one branch within the *rab11* phylogeny. This test used a likelihood ratio test with my *rab11* phylogeny, and it showed evidence of episodic diversifying selection with strong statistical confidence ($p = 0.002$). I tested individual codon sites for evidence of selection using the MEME (Mixed Effects Model of Evolution) test, and I found evidence of episodic diversifying selection at three codon sites: codon 13, codon 210, and codon 220 ($p < 0.1$; Figure 12). In Murrell *et al.* (2012), a p-value < 0.1 is used to identify statistically significant sites for selection. Only codon 13 had a p-value less than 0.05 ($p=0.02091$); this codon encodes a functionally significant amino acid (K13) that aids in binding between a GTP bound Rab11a and Myo5b (Table 14). Although Table 14 shows there to be zero substitutions within any single clade for the K13 site, when using the entire phylogeny as I did with the MEME selection test, I found evidence for selection at the K13 site. The only organism in all the *rab11* duplicates that had sequence differences for the K13 site leading to a different amino acid was the flycatcher, with a K→A amino acid change in its *rab11a* gene. When using a single clade of 7-10 teleost sequences, they are so highly conserved with zero substitutions that only purifying selection is present. When I incorporate the entire phylogeny of 68 sequences, then there are variations taking place such that the test detects positive selection among some codons and purifying selection among others thus detecting episodic diversifying selection.

In testing specific branches, I found evidence of episodic diversifying selection to be present in the branch leading to the *rab11a* clade and in the branch leading to the *rab11a1* clade. For the *rab11a* clade, 91% of the sites have a dN/dS rate of 0.07, indicating purifying selection for these sites. However, 8.8% of the sites had an average dN/dS rate of 67.2, suggesting positive selection took place (Figure 13, panel B). I found only one other branch showing evidence of episodic diversifying selection, and that was for the *rab11a1* clade. Although, two rate classes were found on both the *rab11a1* branch and on the *rab11a* branch, the dN/dS rates were quite different. For the *rab11a1 branch*, there is a similar percentage of sites (94%) with dN/dS less than one as seen with the *rab11a* branch (91%). However, the dN/dS rate for the sites under positive selection was found to be 1090 for the *rab11a1* gene clade compared to 67.2 for the *rab11a* clade.

An aBS-REL (adaptive Branch Site Random Effects Likelihood) test was used to test for selection along specific branches. I found evidence of selection occurring on two of the branches in my *rab11* phylogeny (Figure 13A). Out of 133 branches in the phylogeny, 109 branches were found to have a single and often unique ω (dN/dS) rate values for each branch. For the other 24 branches out of the 133 branches tested, each branch had two rate classes, $\omega_1$ and $\omega_2$ (Figure 13A). For the *rab11a* branch in my phylogeny, I found two rate classes, $\omega_1 = 0.0720$ in 91% of the sites and $\omega_2 = 67.2$ for 8.8% of the sites (Figure 13, panels B and D). For the *rab11a1* branch in my phylogeny, I found two rate classes, $\omega_1 = 0.659$ in 94% of the sites and $\omega_2 = 1090$ for 6.1% of the sites (Figure 13, panels C and D).

## Discussion

Previously only three *rab11* genes had been described in vertebrates, *rab11a, rab11b,* and *rab11c*, although *rab11c* is usually referred to as *rab25*. This study is focused on the evolutionary history of the *rab11a* and *rab11b* clades as I seek to identify a correlation between the evolution of these genes and the previously studied *myo5* genes (See chapter 1). Here, I identify new *rab11* clades for non-teleosts (*rab11a1*) and for teleosts (*rab11a1, rab11ba2,* and *rab11bb*). My phylogenetic analysis revealed a *rab11aa* clade in teleosts, a *rab11a1* clade consisting of teleosts and non-teleosts, and a *rab11b* clade which is further divided into teleosts and non-teleosts (Figure 9). Some of the sequences for some of the organisms tested in my phylogeny did not sort into the expected clade. I suspect this is in part due to the high amount of conservation among the sequences along with using an alignment that was only 699 base pairs long. In the *rab11a* clade, some of the organisms that seemed unresolved or misplaced included spotted gar, coelacanth, and *Ciona.* These nodes had poor support with posterior probabilities between 0.75 and 0.8. For the *rab11b* clade, human and mouse were in an unexpected place and the posterior probability for the ancestral node was 0.53 which provides very weak support for this placement.

My syntenic analyses and my ancestral chromosomal mapping support my findings that *rab11a* and *rab11b* gene clades derive from ancestral vertebrate chromosome A (Figure 10), and this event is likely the result of the first genome duplication event (1R) in vertebrate evolutionary history. During the second whole genome duplication in vertebrates (2R), *rab11a* and *rab11a1* were formed and both gene clades have been maintained in numerous vertebrate families since then. These gene clades were mapped back to segments A4 and A5. Numerous genes around *rab11b, rab11a* and *rab11a1* seem to be co-duplicated. Some of the co-duplicated genes either in teleosts only or in teleosts plus other vertebrates include *rab25, mex, smad6, sema4,* *dennd4, pip5k1, map2k, march2, mbd, tcf,* and, *myo5b*

The newly identified *rab11a1* clade was found in teleost and non-teleost fish including spotted gar and coelacanth. I also found *rab11a1* genes present in two birds (chicken and flycatcher) and one reptile (turtle). The *rab11a1* clade was shown to have a higher dN/dS rate than other *rab11* clades and showed evidence of diversifying selection along the phylogenetic branch leading up to this clade. Episodic diversifying selection allowed for different evolutionary rates along each branch such that codons along a branch were placed in one, two, or three rate categories. Branches that were found to have a proportion of the codons in one of the rate categories greater than one (dN/dS>1) and the remainder of the codons to have a rate less than one with statistical confidence ($p<0.05$) are identified as experiencing episodic diversifying selection. As with other *rab11* clades, the codons linked with functionally significant amino acids have dN/dS values of zero for all 18 codons when examining one clade, such as *rab11aa*, and when including all five clades, the dN/dS rate is 0.01, indicating only slight variation from one clade to another for a small number of these 18 codons.

New *rab11* genes in teleost lineages were identified and shown to have a high percentage of sequence similarity. Two of the teleost specific genes, *rab11ba1* and *rab11ba2*, are found on the same chromosome or scaffold for six of the ten teleosts tested. The other four teleosts only have one of the rab11ba genes present and I suspect the other gene was lost early in evolutionary time. Highly conserved intron sequences were found to be present in the *rab11ba1* and *rab11ba2* clades. I found these genes to be paralogous and located on the same chromosome, and I found that these genes were separated by a *march2a* gene. The high degree of conservation in both coding and non-coding intron sequences suggests gene conversion has taken place. I suspect there is a regulatory component in these intron sequences that explains the high degree of conservation for sequences that are non-coding and diverged from each other most likely over 100 million years ago. The last common ancestor for the species that have both *rab11ba1* and *rab11ba2* genes is thought to have lived approximately 140-170 million years ago (timetree.org).

A high percentage of *rab11* codons (18-34%) were identified as having zero synonymous or non-synonymous substitutions for the ten teleosts used in this study. The clade with the highest percentage of codons having a dN value of zero is the *rab11aa* clade with a value of 97.67%. This clade of genes from teleosts, which consists of fish that diverged over 100 million years ago, has 210 out of 215 codons that have zero nonsynonymous changes.

I found evidence of episodic selection in specific codons and specific branches in my phylogeny using several selection assays. I identified three codons experiencing episodic diversifying selection and one of these codons (codon 13) coded for an amino acid (K13) previously identified as functionally important for the Rab11a protein (Pylypenko O *et al.*, 2013).

In chapter one, I identified a clade of *myosin 5* genes (*myo5bb*) that showed signs of high conservation along with positive selection, leading us to predict new functionality for this duplicated clade. Myosin 5b proteins have been shown to bind with Rab11a proteins. Since I have identified new duplicated clades for each of these previously described protein families, I examined the evolutionary rates for the duplicates (*myo5bb* and *rab11a1*) along with selective forces acting on these duplicate clades. I found a high degree of conservation in the codons that code for functionally significant amino acids. In chapter one and in this chapter, I found both duplicate clades have the highest dN/dS ratios compared to other paralogous clades.  These higher evolutionary rates for each clade along with the selection assay results and high degree of conservation for functionally linked codons lead us to infer that these clades are co-evolving across numerous taxa.

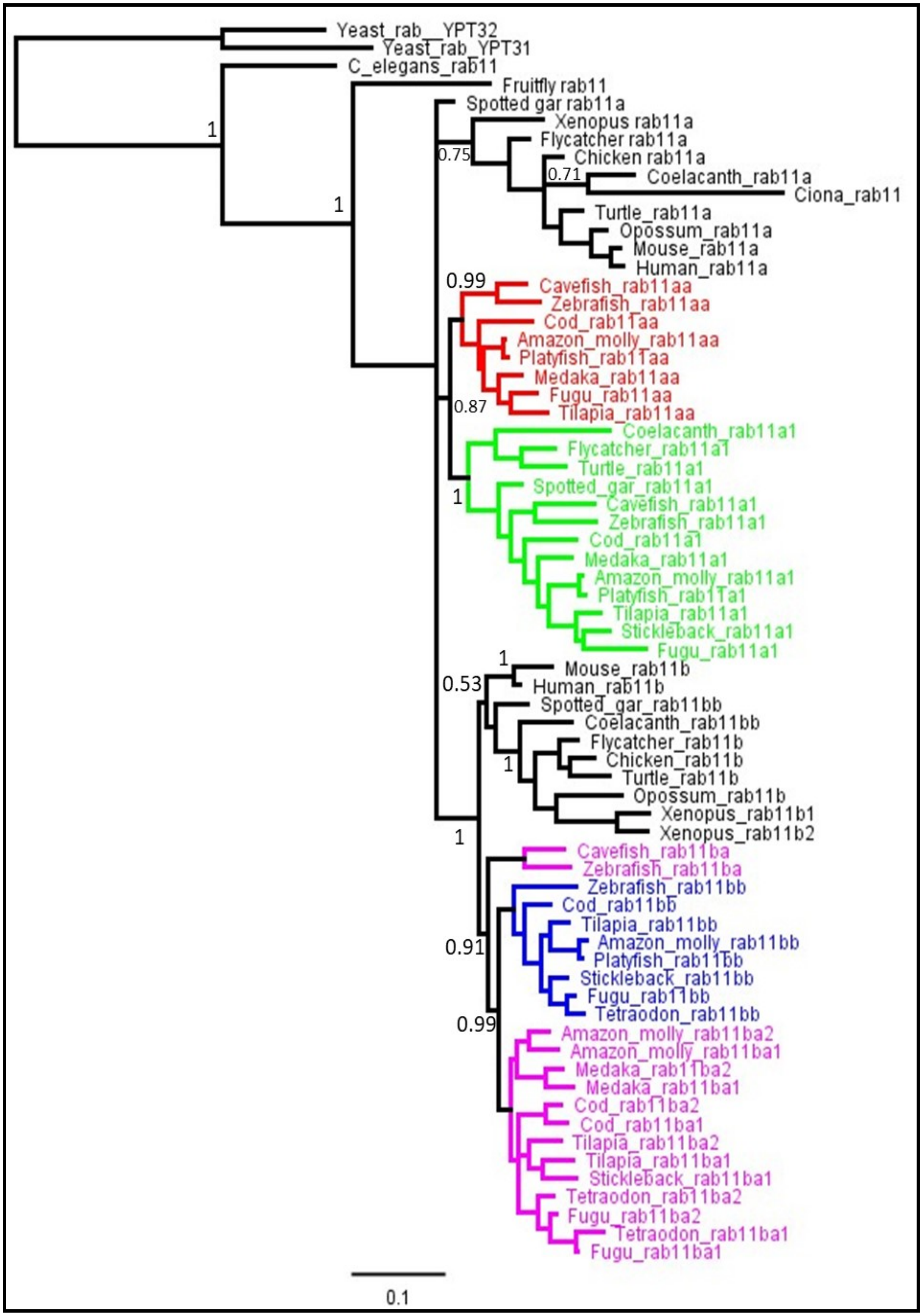


**Figure 9. Phylogenetic tree for 68 *rab11* sequences.**

**Figure 9. Phylogenetic tree for 68 *rab11* sequences (Preceding page).** Newly identified *rab11* genes include *rab11bb* (blue), *rab11ba2* (purple), and *rab11a1* (green). Gene conversion may be responsible for the relationships observed among *rab11ba1* and *rab11ba2* genes in teleosts (purple). The numbers at the nodes are posterior probabilities. Posterior probability values for nodes without a value shown are between 0.8 and 1. The scale bar represents substitutions per site.

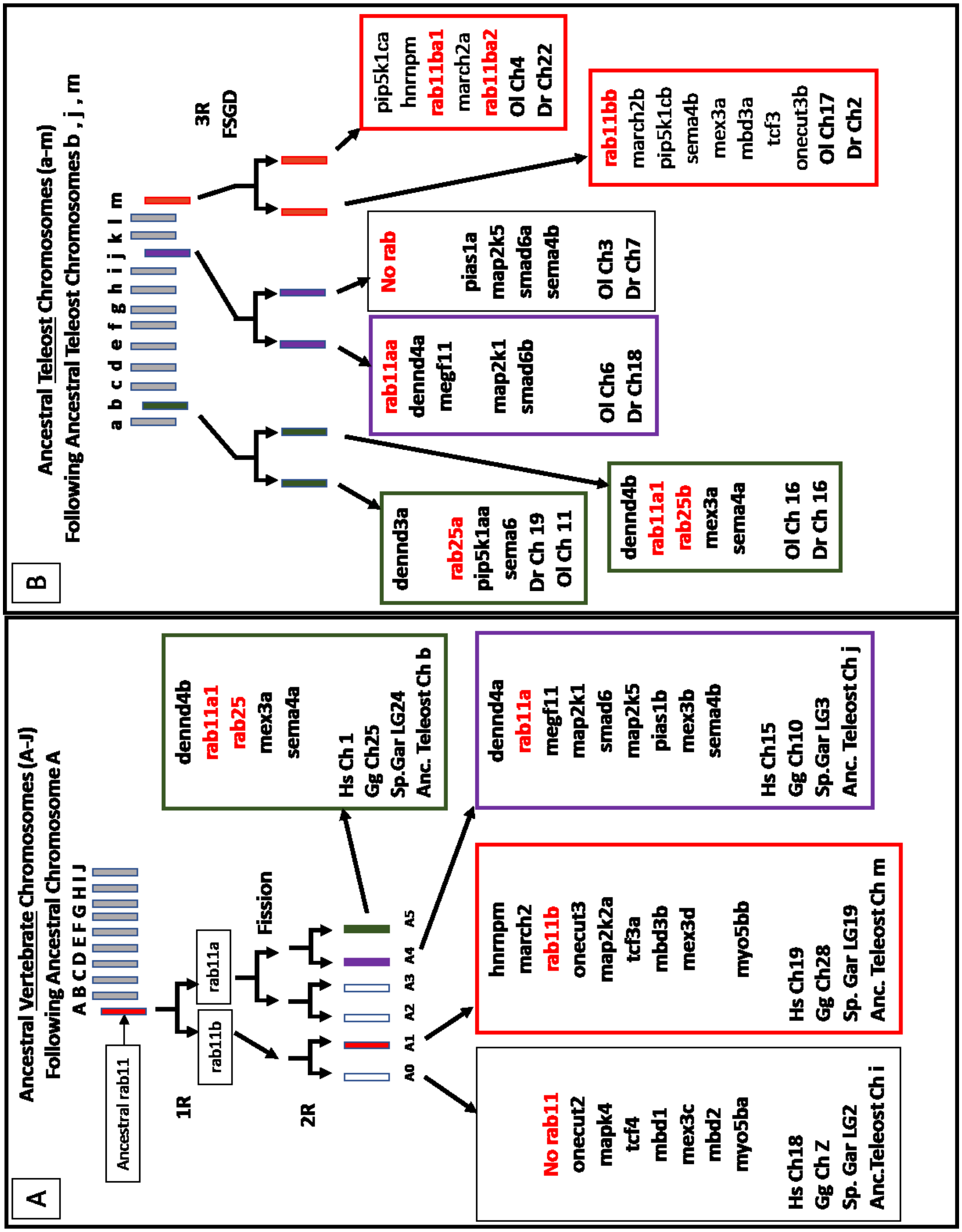


Figure 10. Ancestral chromosome mapping for *rab11*.

**Figure 10. Ancestral chromosome mapping for *rab11* (Preceding page).** The *rab11* genes in teleosts were mapped back to ancestral vertebrate chromosomes (A) and ancestral teleost chromosomes (B). All *rab11* genes map back to ancestral vertebrate chromosome A in panel A. After 1R, there were two *rab11* genes (*rab11a* and *rab11b*). After two rounds of whole genome duplication in ancestral vertebrates, three *rab11* genes are present (*rab11a, rab11a1*, and *rab11b*). After 2R, *rab11a* and the genes around it are found on ancestral chromosome segments A4 and A5. *rab25* is included since it is also known as *rab11c*. *rab11* genes are shown in red whereas syntenic genes are shown in black. Boxed regions include abbreviations for organisms and the chromosome or linkage group that have most or all of the genes in the boxed regions. Abbreviations for representative organisms are Hs (*Homo* sapiens), Gg (*Gallus* gallus), Sp. Gar (Spotted Gar), Anc.Teleost (ancestral teleost), Dr (*Danio* rerio), Ol (*Oryzias latipes*)

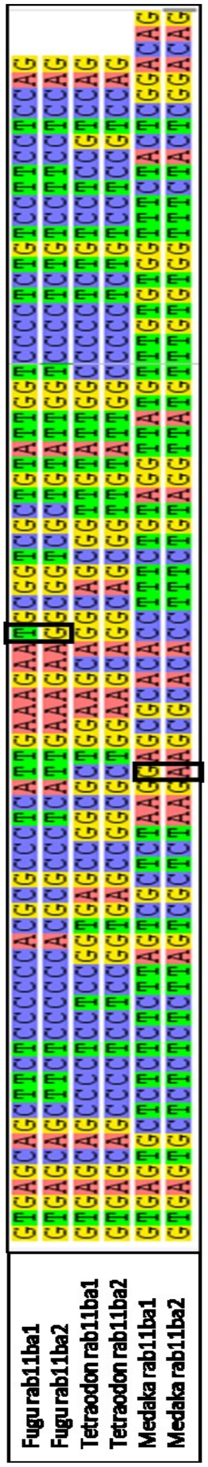
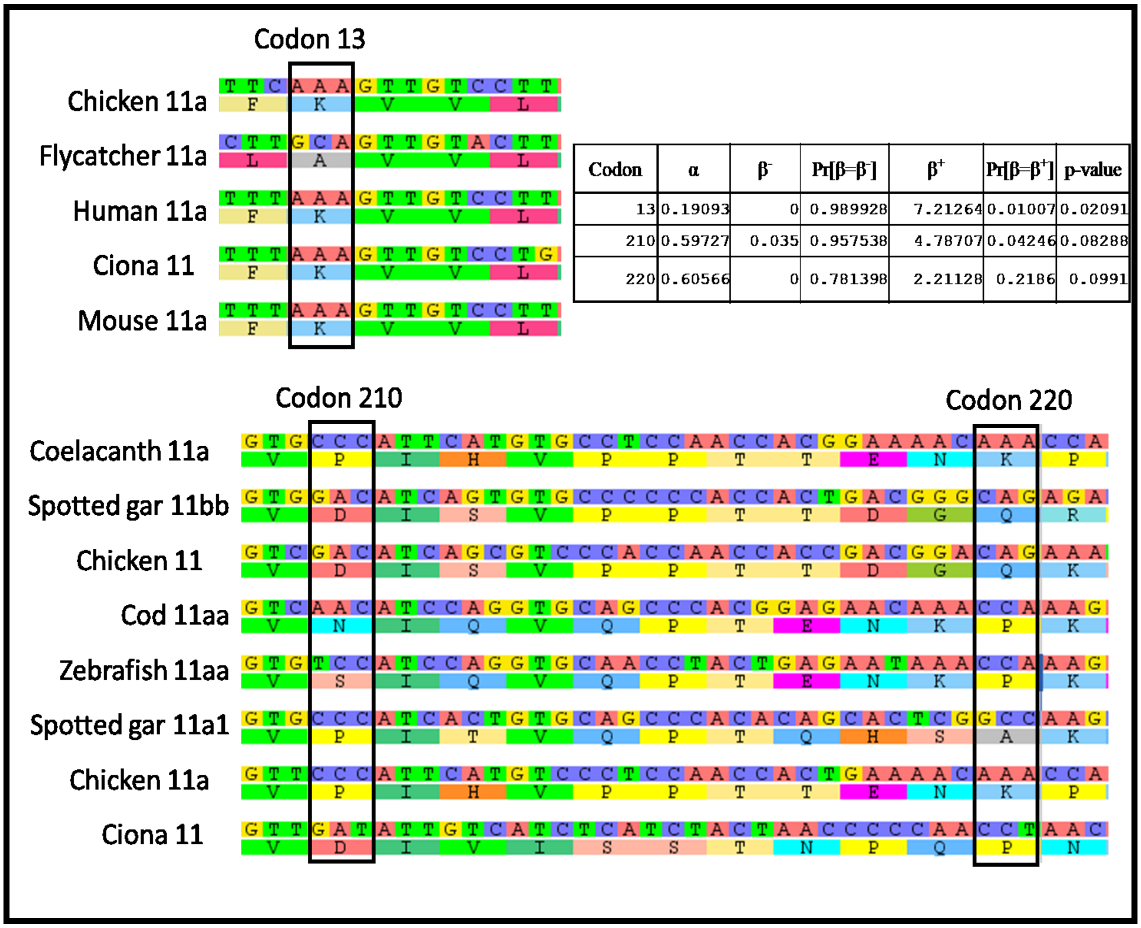


**Figure 11. Intron 2/3 sequences for *rab11ba1* and *rab11ba2*.**

**Figure 11. Intron 2/3 sequences for *rab11ba1* and *rab11ba2* (Previous page).** Paralogous intron sequences (intron 2/3) are highly conserved (98.7-100%) in three fish species for *rab11ba1* and *rab11ba2* genes. For fugu, tetraodon, and medaka there are one, zero, and one nucleotide differences, respectively, between the two paralogous sequences *rab11ba1* and *rab11ba2.* For the two paralogous fugu sequences, all the bases are identical except one at position 40. For tetraodon, all 77 bases of the paralogous intron sequences are identical. For medaka, there is one nucleotide difference between the paralogous sequences at position 31. Gene conversion may contribute to the high degree of conservation in non-coding regions.

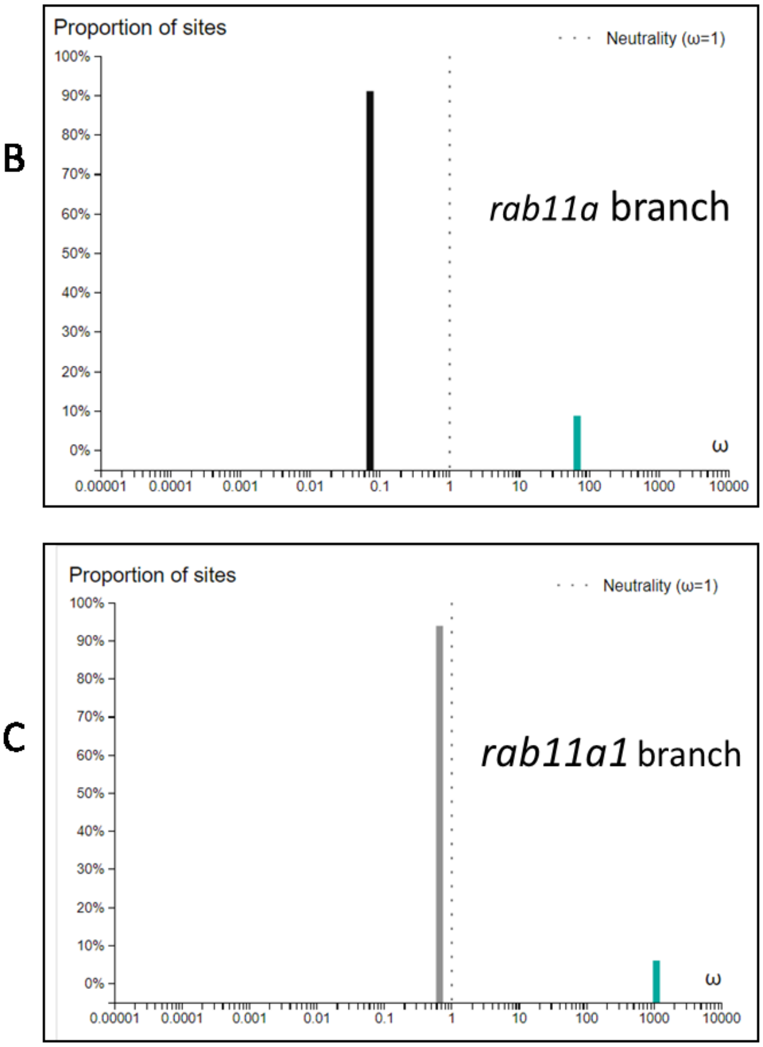

Codon 13
Chicken 11a
TTCAAAGTTGTCCTT
F K V V L
Flycatcher 11a
CTTGCAGTTGTACTT
L A V V L
Human 11a
TTTAAAGTTGTCCTT
F K V V L
Ciona 11
TTTAAAGTTGTCCTG
F K V V L
Mouse 11a
TTTAAAGTTGTCCTT
F K V V L
Codon 210
Codon 220
Coelacanth 11a
GTGCCCATTCATGTGCCTCCAACCACGGAAAACAAACCA
V P I H V P P T T E N K P
Spotted gar 11bb
GTGGACATCAGTGTGCCCCCCACCACTGACGGGCAGAGA
V D I S V P P T T D G Q R
Chicken 11
GTCGACATCAGCGTCCCACCAACCACCGACGGACAGAAA
V D I S V P P T T D G Q K
Cod 11aa
GTCAACATCCAGGTGCAGCCCACGGAGAACAAACCAAAG
V N I Q V Q P T E N K P K
Zebrafish 11aa
GTGTCCATCCAGGTGCAACCTACTGAGAATAAACCAAAG
V S I Q V Q P T E N K P K
Spotted gar 11a1
GTGCCCATCACTGTGCAGCCCACACAGCACTCGGCCAAG
V P I T V Q P T Q H S A K
Chicken 11a
GTTCCCATTCATGTCCCTCCAACCACTGAAAACAAACCA
V P I H V P P T T E N K P
Ciona 11
GTTGATATTGTCATCTCATCTACTAACCCCCAACCTAAC
V D I V I S S T N P Q P N

| Codon | α | β- | Pr[β=β-] | β+ | Pr[β=β+] | p-value |
|---|---|---|---|---|---|---|
| 13 | 0.19093 | 0 | 0.989928 | 7.21264 | 0.01007 | 0.02091 |
| 210 | 0.59727 | 0.035 | 0.957538 | 4.78707 | 0.04246 | 0.08288 |
| 220 | 0.60566 | 0 | 0.781398 | 2.21128 | 0.2186 | 0.0991 |

**Figure 12. Diversifying selection among *rab11* codons.** A MEME test was used to look for evidence of episodic diversifying selection at the codon site level. There were 68 *rab11* sequences analyzed and three sites, codon 13, codon 210, and codon 220, showed evidence of episodic diversifying selection. Non-synonymous substitution rates (β) and synonymous substitution rates ($\alpha$) are used in a MEME test. When $\beta < \alpha$, $\omega < 1$, and this is noted with $\beta^{-}$. When $\beta > \alpha$, $\omega > 1$ and this is noted with $\beta^{+}$. The probability that β is $\beta^{-}$ or $\beta^{+}$ is provided along with the p-value for each of the three codons identified as being subject to episodic diversifying selection.

A

| ω rate classes | # of branches | % of branches | # under selection |
|---|---|---|---|
| 1 | 109 | 82% | 0 |
| 2 | 24 | 18% | 2 |



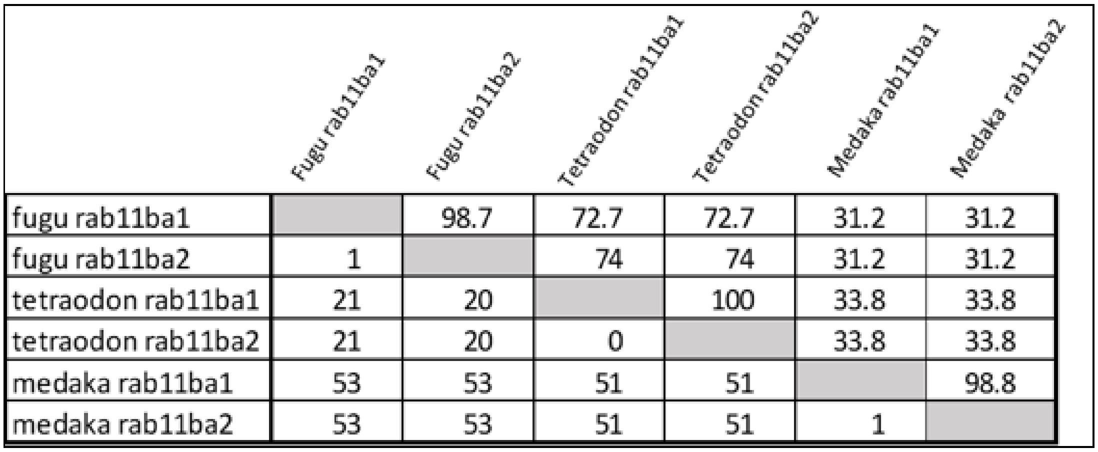
D

| Branch | LRT | Test p-value | Uncorrected p-value | ω distribution over sites |
|---|---|---|---|---|
| *rab11a1* | 18.7621 | 0.0018 | 0 | ω1 = 0.659 (94%)<br>ω2 = 1090 (6.1%) |
| *rab11a* | 16.6873 | 0.0051 | 0.0001 | ω1 = 0.0720 (91%)<br>ω2 = 67.2 (8.8%) |

**Figure 13. Diversifying selection among *rab11* branches.**

**Figure 13. Diversifying selection among *rab11* branches (Previous page).** An aBSREL test was used to identify evidence of episodic diversifying selection at the branch level. Out of 133 branches tested on the *rab11* tree, 2 branches showed evidence of episodic diversifying selection. Of the 133 branches tested for selection, there were 109 branches that were subject to a single rate class, ω, and there were 24 branches that were subject to two rate classes, $\omega_1$, $\omega_2$. Panels B and D shows that 94% of the sites along the *rab11a1* branch are subject to an $\omega_1$=0.659 value and 6.1% of the sites on this branch are subject to an $\omega_2$=1090 value with a corrected p-value of 0.0018. Panels C and D show the rates and percentage of codons for each rate for the *rab11a* branch.

**Table 11. Intron 2/3 percent identity for *rab11ba1* and *rab11ba2*.** Comparison of intron 2/3 in fugu, tetraodon, and medaka for *rab11ba1* and *rab11ba2*. The sizes of these introns are 77 bp for fugu and tetraodon and 80 bp for medaka. Each pair of paralogs are the same length and highly conserved as the table notes with only one nucleotide difference for medaka or fugu and zero differences for tetraodon. Below the diagonal: the number of nucleotide differences compared to two sequences. Above the diagonal: percent sequence similarity.



| | Fugu rab11ba1 | Fugu rab11ba2 | Tetraodon rab11ba1 | Tetraodon rab11ba2 | Medaka rab11ba1 | Medaka rab11ba2 |
|---|---|---|---|---|---|---|
| fugu rab11ba1 | | 98.7 | 72.7 | 72.7 | 31.2 | 31.2 |
| fugu rab11ba2 | 1 | | 74 | 74 | 31.2 | 31.2 |
| tetraodon rab11ba1 | 21 | 20 | | 100 | 33.8 | 33.8 |
| tetraodon rab11ba2 | 21 | 20 | 0 | | 33.8 | 33.8 |
| medaka rab11ba1 | 53 | 53 | 51 | 51 | | 98.8 |
| medaka rab11ba2 | 53 | 53 | 51 | 51 | 1 | |

**Table 12.** ***rab11*** **model testing using maximum likelihood methods.** A GTR+G+I model provided the best BIC and AIC$_c$ scores. Each model uses a different set of parameters with some models including a Gamma shape parameter, some models allowing for a proportion of sites to be invariant, and some models allowing for both conditions. In addition, each model provides a specific weighting for rates of substitutions for transitions and transversions.

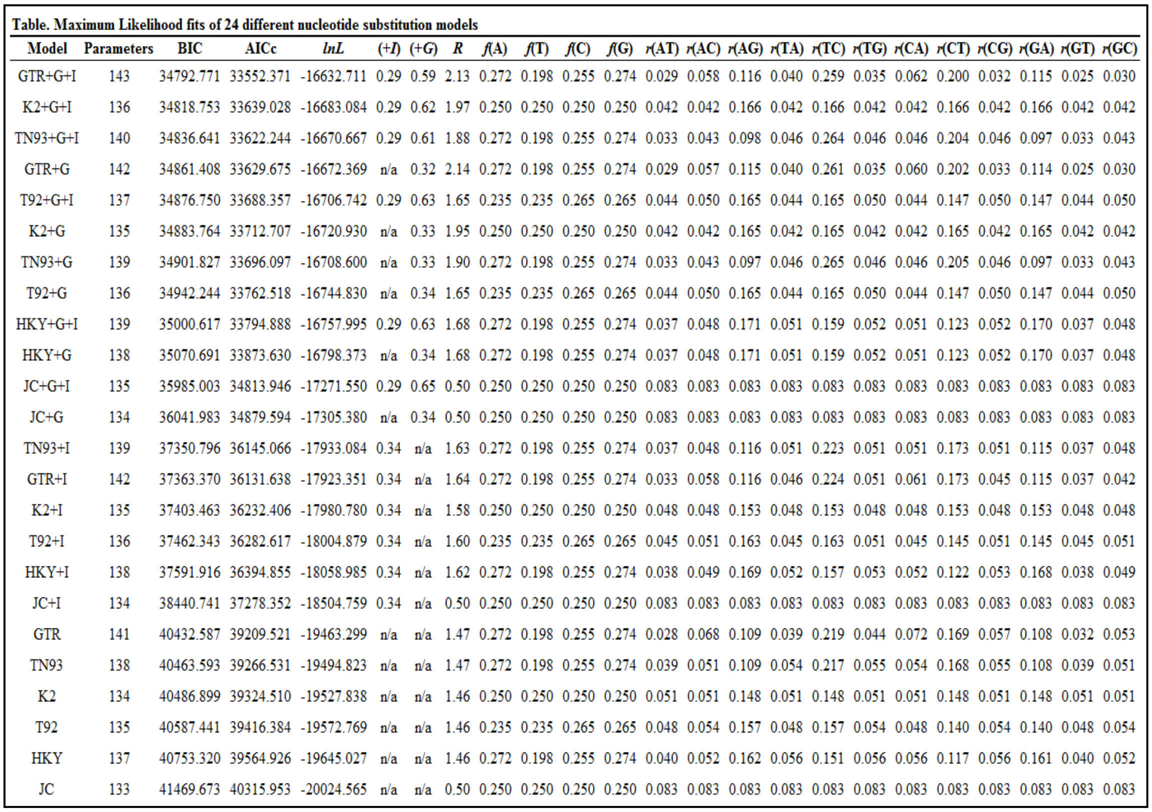


**Table. Maximum Likelihood fits of 24 different nucleotide substitution models**

| Model | Parameters | BIC | AICc | *lnL* | (+*I*) | (+*G*) | *R* | *f*(A) | *f*(T) | *f*(C) | *f*(G) | *r*(AT) | *r*(AC) | *r*(AG) | *r*(TA) | *r*(TC) | *r*(TG) | *r*(CA) | *r*(CT) | *r*(CG) | *r*(GA) | *r*(GT) | *r*(GC) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GTR+G+I | 143 | 34792.771 | 33552.371 | -16632.711 | 0.29 | 0.59 | 2.13 | 0.272 | 0.198 | 0.255 | 0.274 | 0.029 | 0.058 | 0.116 | 0.040 | 0.259 | 0.035 | 0.062 | 0.200 | 0.032 | 0.115 | 0.025 | 0.030 |
| K2+G+I | 136 | 34818.753 | 33639.028 | -16683.084 | 0.29 | 0.62 | 1.97 | 0.250 | 0.250 | 0.250 | 0.250 | 0.042 | 0.042 | 0.166 | 0.042 | 0.166 | 0.042 | 0.042 | 0.166 | 0.042 | 0.166 | 0.042 | 0.042 |
| TN93+G+I | 140 | 34836.641 | 33622.244 | -16670.667 | 0.29 | 0.61 | 1.88 | 0.272 | 0.198 | 0.255 | 0.274 | 0.033 | 0.043 | 0.098 | 0.046 | 0.264 | 0.046 | 0.046 | 0.204 | 0.046 | 0.097 | 0.033 | 0.043 |
| GTR+G | 142 | 34861.408 | 33629.675 | -16672.369 | n/a | 0.32 | 2.14 | 0.272 | 0.198 | 0.255 | 0.274 | 0.029 | 0.057 | 0.115 | 0.040 | 0.261 | 0.035 | 0.060 | 0.202 | 0.033 | 0.114 | 0.025 | 0.030 |
| T92+G+I | 137 | 34876.750 | 33688.357 | -16706.742 | 0.29 | 0.63 | 1.65 | 0.235 | 0.235 | 0.265 | 0.265 | 0.044 | 0.050 | 0.165 | 0.044 | 0.165 | 0.050 | 0.044 | 0.147 | 0.050 | 0.147 | 0.044 | 0.050 |
| K2+G | 135 | 34883.764 | 33712.707 | -16720.930 | n/a | 0.33 | 1.95 | 0.250 | 0.250 | 0.250 | 0.250 | 0.042 | 0.042 | 0.165 | 0.042 | 0.165 | 0.042 | 0.042 | 0.165 | 0.042 | 0.165 | 0.042 | 0.042 |
| TN93+G | 139 | 34901.827 | 33696.097 | -16708.600 | n/a | 0.33 | 1.90 | 0.272 | 0.198 | 0.255 | 0.274 | 0.033 | 0.043 | 0.097 | 0.046 | 0.265 | 0.046 | 0.046 | 0.205 | 0.046 | 0.097 | 0.033 | 0.043 |
| T92+G | 136 | 34942.244 | 33762.518 | -16744.830 | n/a | 0.34 | 1.65 | 0.235 | 0.235 | 0.265 | 0.265 | 0.044 | 0.050 | 0.165 | 0.044 | 0.165 | 0.050 | 0.044 | 0.147 | 0.050 | 0.147 | 0.044 | 0.050 |
| HKY+G+I | 139 | 35000.617 | 33794.888 | -16757.995 | 0.29 | 0.63 | 1.68 | 0.272 | 0.198 | 0.255 | 0.274 | 0.037 | 0.048 | 0.171 | 0.051 | 0.159 | 0.052 | 0.051 | 0.123 | 0.052 | 0.170 | 0.037 | 0.048 |
| HKY+G | 138 | 35070.691 | 33873.630 | -16798.373 | n/a | 0.34 | 1.68 | 0.272 | 0.198 | 0.255 | 0.274 | 0.037 | 0.048 | 0.171 | 0.051 | 0.159 | 0.052 | 0.051 | 0.123 | 0.052 | 0.170 | 0.037 | 0.048 |
| JC+G+I | 135 | 35985.003 | 34813.946 | -17271.550 | 0.29 | 0.65 | 0.50 | 0.250 | 0.250 | 0.250 | 0.250 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 |
| JC+G | 134 | 36041.983 | 34879.594 | -17305.380 | n/a | 0.34 | 0.50 | 0.250 | 0.250 | 0.250 | 0.250 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 |
| TN93+I | 139 | 37350.796 | 36145.066 | -17933.084 | 0.34 | n/a | 1.63 | 0.272 | 0.198 | 0.255 | 0.274 | 0.037 | 0.048 | 0.116 | 0.051 | 0.223 | 0.051 | 0.051 | 0.173 | 0.051 | 0.115 | 0.037 | 0.048 |
| GTR+I | 142 | 37363.370 | 36131.638 | -17923.351 | 0.34 | n/a | 1.64 | 0.272 | 0.198 | 0.255 | 0.274 | 0.033 | 0.058 | 0.116 | 0.046 | 0.224 | 0.051 | 0.061 | 0.173 | 0.045 | 0.115 | 0.037 | 0.042 |
| K2+I | 135 | 37403.463 | 36232.406 | -17980.780 | 0.34 | n/a | 1.58 | 0.250 | 0.250 | 0.250 | 0.250 | 0.048 | 0.048 | 0.153 | 0.048 | 0.153 | 0.048 | 0.048 | 0.153 | 0.048 | 0.153 | 0.048 | 0.048 |
| T92+I | 136 | 37462.343 | 36282.617 | -18004.879 | 0.34 | n/a | 1.60 | 0.235 | 0.235 | 0.265 | 0.265 | 0.045 | 0.051 | 0.163 | 0.045 | 0.163 | 0.051 | 0.045 | 0.145 | 0.051 | 0.145 | 0.045 | 0.051 |
| HKY+I | 138 | 37591.916 | 36394.855 | -18058.985 | 0.34 | n/a | 1.62 | 0.272 | 0.198 | 0.255 | 0.274 | 0.038 | 0.049 | 0.169 | 0.052 | 0.157 | 0.053 | 0.052 | 0.122 | 0.053 | 0.168 | 0.038 | 0.049 |
| JC+I | 134 | 38440.741 | 37278.352 | -18504.759 | 0.34 | n/a | 0.50 | 0.250 | 0.250 | 0.250 | 0.250 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 |
| GTR | 141 | 40432.587 | 39209.521 | -19463.299 | n/a | n/a | 1.47 | 0.272 | 0.198 | 0.255 | 0.274 | 0.028 | 0.068 | 0.109 | 0.039 | 0.219 | 0.044 | 0.072 | 0.169 | 0.057 | 0.108 | 0.032 | 0.053 |
| TN93 | 138 | 40463.593 | 39266.531 | -19494.823 | n/a | n/a | 1.47 | 0.272 | 0.198 | 0.255 | 0.274 | 0.039 | 0.051 | 0.109 | 0.054 | 0.217 | 0.055 | 0.054 | 0.168 | 0.055 | 0.108 | 0.039 | 0.051 |
| K2 | 134 | 40486.899 | 39324.510 | -19527.838 | n/a | n/a | 1.46 | 0.250 | 0.250 | 0.250 | 0.250 | 0.051 | 0.051 | 0.148 | 0.051 | 0.148 | 0.051 | 0.051 | 0.148 | 0.051 | 0.148 | 0.051 | 0.051 |
| T92 | 135 | 40587.441 | 39416.384 | -19572.769 | n/a | n/a | 1.46 | 0.235 | 0.235 | 0.265 | 0.265 | 0.048 | 0.054 | 0.157 | 0.048 | 0.157 | 0.054 | 0.048 | 0.140 | 0.054 | 0.140 | 0.048 | 0.054 |
| HKY | 137 | 40753.320 | 39564.926 | -19645.027 | n/a | n/a | 1.46 | 0.272 | 0.198 | 0.255 | 0.274 | 0.040 | 0.052 | 0.162 | 0.056 | 0.151 | 0.056 | 0.056 | 0.117 | 0.056 | 0.161 | 0.040 | 0.052 |
| JC | 133 | 41469.673 | 40315.953 | -20024.565 | n/a | n/a | 0.50 | 0.250 | 0.250 | 0.250 | 0.250 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 |

**Table 13. Evolutionary rates for *rab11*.** Rates (dN/dS) were determined for each teleost *rab11* clade and for all 68 sequences analyzed. A dN/dS rate of 0.01 was found for codons associated with functionally significant amino acids for all 68 sequences. Other dN/dS rates for entire *rab11* genes for each clade ranged from 0.01-0.09.

| | # sequences | # codons | dN/dS | % invariant codons |
|---|---|---|---|---|
| rab11ba1 | 7 | 218 | 0.05 | 34.4% |
| rab11ba2 | 8 | 216 | 0.03 | 23.1% |
| rab11bb | 8 | 218 | 0.02 | 30.3% |
| rab11aa | 8 | 215 | 0.01 | 23.7% |
| rab11a1 | 9 | 202 | 0.09 | 18.3% |
| rab11 (All codons) | 68 | 215 | 0.09 | 0.9% |
| rab11 (functionally significant codons) | 68 | 18 | 0.01 | 5.6% |

**Table 14. Evolutionary rates for 18 codons for *rab11.*** dN/dS rates were 0 or 0/0 for each of the codons linked with functionally significant amino acids for each of the five *rab11* clades in teleosts. Eighteen codons linked with functionality have a dN value of 0 for each *rab11* clade and several codons are invariant across all teleosts examined such that dS is also 0.

| | | *rab11* gene dN/dS values for functionally linked codons | | | | |
|---|---|---|---|---|---|---|
| amino acid | codon | *rab11aa* | *rab11a1* | *rab11ba1* | *rab11ba2* | *rab11bb* |
| K13 | AAA | 0/0 | 0 | 0/0 | 0/0 | 0 |
| L16 | CTG | 0 | 0 | 0 | 0 | 0 |
| R33 | AGA | 0 | 0 | 0 | 0 | 0 |
| I44 | ATC | 0 | 0 | 0 | 0 | 0 |
| G45 | GGG | 0 | 0 | 0 | 0 | 0 |
| V46 | GTG | 0 | 0 | 0/0 | 0/0 | 0/0 |
| F48 | TTC | 0 | 0 | 0 | 0 | 0 |
| T50 | ACC | 0 | 0 | 0 | 0 | 0 |
| Q63 | CAA | 0/0 | 0 | 0 | 0 | 0/0 |
| W65 | TGG | 0/0 | 0/0 | 0/0 | 0 | 0/0 |
| T67 | ACG | 0 | 0 | 0 | 0 | 0 |
| A68 | GCT | 0 | 0 | 0 | 0 | 0 |
| E71 | GAA | 0/0 | 0 | 0 | 0 | 0 |
| Y73 | TAC | 0 | 0/0 | 0/0 | 0/0 | 0/0 |
| A75 | GCC | 0/0 | 0 | 0 | 0 | 0/0 |
| I76 | ATC | 0 | 0/0 | 0/0 | 0 | 0/0 |
| T77 | ACC | 0 | 0 | 0 | 0 | 0 |
| Y80 | TAT | 0 | 0 | 0 | 0 | 0 |

**Table 15. Evolutionary rates for teleost *rab11* duplicates.** Each *rab11* clade consists of seven to nine teleost sequences consisting of 202 to 218 codons. The number and percentage of codons with dN and dS values equal to zero (dN/dS=0/0) are noted as invariant. These codons have zero substitutions for the seven to nine teleosts that diverged from a common ancestor approximately 148 million years ago. The percentage of invariant codons ranges from 18.32% to 34.4%. The total number of codons are further organized to see how many codons within a clade have at least one synonymous substitution (dS>0), the number and percentage of codons that have zero nonsynonymous substitutions but at least one synonymous substitution (dN/dS=0), the number and percentage of codons that have a dN/dS greater than 0, (dN/dS>0) or the percentage of codons that don't have any non-synonymous substitutions (dN=0).

| | Total codons | # codons invar. | % codons invar. | # codons dS>0 | # codons dN/dS =0 | % codons dN/dS =0 | # codons dN/dS >0 | % codons dN/dS >0 | % codons dN=0 |
|---|---|---|---|---|---|---|---|---|---|
| rab11aa | 215 | 51 | 23.72 | 164 | 159 | 73.95 | 5 | 2.33 | 97.67 |
| rab11a1 | 202 | 37 | 18.32 | 165 | 132 | 65.35 | 33 | 16.34 | 83.66 |
| rab11ba1 | 218 | 75 | 34.40 | 143 | 124 | 56.88 | 19 | 8.72 | 91.28 |
| rab11ba2 | 216 | 50 | 23.15 | 166 | 147 | 68.06 | 19 | 8.80 | 91.20 |
| rab11bb | 218 | 66 | 30.28 | 152 | 143 | 65.60 | 9 | 4.13 | 95.87 |

**Table 16. Organisms, gene identifiers, genes and loci used in this study.**

| Organism | Ensembl Gene ID | rab11 gene | chromosome # |
|---|---|---|---|
| Amazon molly | ENSPFOG00000009284.1 | rab11aa | Scaffold KI520333.1 |
| Amazon molly | ENSPFOG00000011316.2 | rab11a1 | Scaffold KI519873.1 |
| Amazon molly | ENSPFOG00000006159 | rab11ba1 | KI520025.1: 205,707-211,019:1 |
| Amazon molly | ENSPFOG00000006224 | rab11ba2 | KI520025.1: 221,890-226,813:1 |
| Amazon molly | ENSPFOG00000024114 | rab11bb | KI519782.1: 893,679-904,591:1 |
| Cave Fish | ENSAMXG00000003155 | rab11aa | KB882123.1 |
| Cave Fish | ENSAMXG00000024632 | rab11a1 | KB872819.1 |
| Cave Fish | ENSAMXG00000016024 | rab11ba2 | KB882160.1 |
| Cod | ENSGMOG00000004096 | rab11aa | GeneScaffold_2898 |
| Cod | ENSGMOG00000006450 | rab11a1 | GeneScaffold_4655 |
| Cod | ENSGMOG00000007390 | rab11ba1 | GeneScaffold_1985 |
| Cod | ENSGMOG00000009668 | rab11ba2 | GeneScaffold_1986 |
| Cod | ENSGMOG00000009211 | rab11bb | GeneScaffold_3237 |
| Drerio | ENSDARG00000041450 | rab11aa | 18 |
| Drerio | ENSDARG00000014340 | rab11a1 | 16 |
| Drerio | ENSDARG00000041878 | rab11ba2 | 22 |
| Drerio | ENSDARG00000090086 | rab11bb | 2 |
| Medaka | ENSORLG00000018331 | rab11aa | ultracontig37 |
| Medaka | ENSORLG00000016126 | rab11a1 | 16 |
| Medaka | ENSORLG00000015307 | rab11ba1 | 4:28326921 |
| Medaka | ENSORLG00000015284 | rab11ba2 | 4:28313863 |
| Medaka | ENSORLG00000013938 | rab11bb | 17:21757475 |
| Platyfish | ENSXMAG00000013563 | rab11aa | JH556927.1 |
| Platyfish | ENSXMAG00000013416 | rab11a1 | JH556728.1 |
| Platyfish | ENSXMAG00000018847 | rab11bb | JH556675.1 |
| Stickleback | ENSGACG00000013259 | rab11a1 | groupXX |
| Stickleback | ENSGACG00000013297 | rab11ba1 | groupVIII |
| Stickleback | ENSGACG00000015479 | rab11bb | groupIII |
| Fugu | ENSTRUG00000010482 | rab11aa | scaffold_1 |
| Fugu | ENSTRUG00000007556 | rab11a1 | scaffold_196 |
| Fugu | ENSTRUG00000018273 | rab11ba1 | Scaffold 25:1632368 |
| Fugu | ENSTRUG00000018283 | rab11ba2 | Scaffold 25:1640890 |
| Fugu | ENSTRUG00000000223 | rab11bb | scaffold_214 |
| Tetraodon | ENSTNIG00000013184 | rab11ba1 | Ch 1: 12928539 |
| Tetraodon | ENSTNIG00000013182 | rab11ba2 | Ch 1: 12921001 |
| Tetraodon | ENSTNIG00000011081 | rab11bb | 15 |

**Table 16(continued). Organisms, gene identifiers, genes and loci used in this study.**

| | | | |
|---|---|---|---|
| Tilapia | ENSONIG00000015072 | rab11aa | GL831436.1 |
| Tilapia | ENSONIG00000002281 | rab11a1 | GL831233.1 |
| Tilapia | ENSONIG00000002535 | rab11ba1 | GL831140.1 : 5879795 |
| Tilapia | ENSONIG00000002544 | rab11ba2 | GL831140.1 : 5897791 |
| Tilapia | ENSONIG00000010435 | rab11bb | GL831134.1 |
| Spotted Gar | ENSLOCG00000013835.1 | rab11aa | LG 3 |
| Spotted Gar | ENSLOCG00000008865 | rab11a1 | LG 24 |
| Spotted Gar | ENSLOCG00000004490 | rab11bb | LG 19 |
| Lamprey | ENSPMAG00000009335 | rab11bb or a1 | GL479037 |
| Coelacanth | ENSLACG00000012184.2 | rab11aa | JH126619.1 |
| Coelacanth | ENSLACG00000011380 | rab11a1 | JH127154.1 |
| Coelacanth | ENSLACG00000016857 | rab11bb | JH126613.1 |
| Human | ENSG00000103769 | rab11a | Ch 15 |
| Human | ENSG00000185236 | rab11b | Ch 19 |
| Mouse | ENSMUSG00000004771 | rab11a | Ch 9 |
| Mouse | ENSMUSG00000077450 | rab11b | Ch 17 |
| Chicken | ENSGALG00000007615 | rab11a | Ch 10 |
| Chicken | ENSGALG00000037802 | rab11a1 | Ch 25 |
| Chicken | ENSGALG00000000613 | rab11b | Ch 28 |
| Flycatcher | ENSFALG00000010052 | rab11a | JH603210.1 |
| Flycatcher | ENSFALG00000002594 | rab11a1 | JH603485.1 |
| Flycatcher | ENSFALG00000011828 | rab11b | JH603352.1 |
| Turtle | ENSPSIG00000011930 | rab11a | JH210454.1 |
| Turtle | ENSPSIG00000013143 | rab11a1 | JH206249.1 |
| Turtle | ENSPSIG00000010242 | rab11b | JH209348.1 |
| Xenopus | ENSXETG00000006639 | rab11a | GL172999.1 |
| Xenopus | ENSXETG00000025484 | rab11b1 | GL173022.1 |
| Xenopus | ENSXETG00000021890 | rab11b2 | GL173022.1 |
| *C.intestinalis* | ENSCING00000009526 | rab11 | Ch 3 |
| Yeast | YER031C | rab11-1 | Ch V |
| Yeast | YGL210W | rab11-2 | Ch VII |
| Fruitfly | FBgn0015790 | rab11 | Ch 3R |
| *C. elegans* | WBGene00004274 | rab-11.1 | Ch 1: 108670 |
| Opossum | ENSMODG00000009810 | rab11a | Ch 1 |
| Opossum | ENSMODG00000003841 | rab11b | Ch 3 |

# IV. EVOLUTION OF DUPLICATED *rab27* GENES

## Abstract

Two known *rab27* genes have been identified and studied in vertebrates. Here I identify a third *rab27* gene found in fish and infer that this gene is a result of the fish-specific genome duplication. I used phylogenetics, syntenic analysis, and evolutionary selection tests to characterize duplicated *rab27* genes. I identified a third and new *rab27bb* clade present in teleosts and characterized the different *rab27* clades. I determined the evolutionary rates for each duplicated gene clade in the form of dN/dS values ranging from 0.11 to 0.24 for teleost fish and non-teleost vertebrates. In addition, I found that the percentage of invariant codons varied among the different duplicated gene clades for teleosts and non-teleosts ranging from 8.6% to 31.5%. The newly identified *rab27bb* clade had the highest dN/dS rate at 0.24. However, the *rab27a* non-teleost clade had the lowest dN/dS rate (0.11) while also having the lowest percentage of codons as invariant (8.6%). Using selection tests to determine if evolutionary rates varied at specific codon sites or along specific branches of my phylogeny, I identified evidence of episodic diversifying selection on the *rab27bb* branch in teleost fish with 13% of the sites on that branch identified as being subject to positive selection and 87% of the sites subject to purifying selection.

## Introduction

Pigmentation of animals is important in natural and sexual selection. Although there are many factors that contribute to pigmentation, one factor is the disposition of the pigment granules in the skin and fur, feathers or scales of the individual. The placement of pigment granules is mediated by molecular motors and accessory proteins that mediate the attachment of those motors to their cargo (reviewed in Wasmeier *et al.*, 2008). Rab27a, melanophilin, and Myo5a have been shown to interact and bind with each other to transport melanosomes across actin cytoskeletal tracks (Hammer & Wu 2007). In this assembly, Myo5a functions as the motor, while Rab27a and melanophilin mediate its attachment to melanin-containing pigment granules, the former through a direct interaction with the Myo5a cargo-binding domain.

In chapter one of this dissertation, I characterized the evolutionary history and types of selection taking place among *myo5* duplicates. I characterized differences between *myo5a* clades which included duplicated clades in teleosts, *myo5aa* and *myo5ab,* and differences in *myo5b* clades, which included the duplicated clades *myo5ba* and *myo5bb* (which represented a vertebrate genome duplication event, R2). I showed that there was significant variability in the cargo binding domain between duplicated genes, while there was high conservation in the motor domain of the duplicated genes. This result supports the hypothesis that these duplicates encode functional proteins. If the evolutionary rate of *myo5ab* clade cargo binding domain is faster than that of the *myo5aa* clade cargo binding domain, then it would be reasonable to expect duplicated gene clades of myosin accessory proteins to similarly co-evolve at faster rates to maintain functional protein-protein interactions (Pazos *et al.*,1997; Goh and Cohen, 2002). Therefore, one of the hypotheses tested in this chapter is that a duplicate Rab27 accessory protein, encoded by *rab27bb*, is evolving (and therefore diverging) at a faster rate than the more conserved "founder" gene *rab27b*. Within the duplicated gene clade, a high degree of conservation could indicate functional activity that has not yet been identified or characterized.

Rab proteins are numerous in type and they have evolved to play a wide range of roles in all living organisms from bacteria to humans. They are part of a superfamily of Ras related G-proteins with functions related to cell signaling, organelle transport, and endocytic processes. There are currently over 30 different numbered types of Rab related proteins with numerous subtypes bringing the total number of Rab related proteins above 60 for humans and many other species (Diekmann *et al.*, 2011). Rab27 is a protein that has had two subtypes previously identified and characterized, Rab27a and Rab27b.

In this chapter, I analyze the gene for Rab27. I seek to determine whether there is evidence of coevolution among duplicated genes for accessory proteins and whether *rab27* shows similar patterns of evolution as have been identified for *myo5* and *rab11* (see Chapters 1 and 2). I identify a third subtype, *rab27bb*, present in teleost fish, that arose due to the teleost specific genome duplication, and show that the *rab27bb* clade is subject to episodic diversifying selection.

## Materials and Methods

### Sequence acquisition

I collected *rab 27* sequences from the Ensembl genomic database (Ensembl Release 86). The following species and genomic assemblies were used for *rab27* sequence downloads: ten teleost species (Amazon molly, *Poecilia formosa*, Poecilia_formosa-5.1.2; cavefish, *Astyanax mexicanus,* AstMex102; cod, *Gadus morhua,* gadMor1; fugu, *Takifugu rubripes,* FUGU 4.0; medaka, *Oryzias latipes,* HdrR; platyfish, *Xiphophorus maculatus,* Xipmac4.4.2; stickleback, *Gasterosteus aculeatus,* BROAD S1; tetraodon, *Tetraodon nigroviridis,* TETRAODON 8.0; tilapia, *Oreochromis niloticus,* Orenil1.0; zebrafish, *Danio rerio,* GRCz10), one holostean fish (spotted gar, *Lepisosteus oculatus,* LepOcu1), one lobe finned fish (coelacanth, *Latimeria chalumnae,* LatCha1), one amphibian (western clawed frog, *Xenopus tropicalis,* JGI 4.2), seven sauropsids (chicken, *Gallus gallus,* Gallus_gallus-5.0; turkey, *Meleagris gallopavo* , Turkey_2.01 ; duck, *Anas platyrhynchos*, BGI_duck_1.0; zebrafinch, *Taeniopygia guttata,* taeGut3.2.4, flycatcher, *Ficedula albicollis,* FicAlb_1.4, Chinese soft shell turtle, *Pelodiscus sinensis*, PelSin_1.0;green anole lizard, *Anolis carolinensis,* AnoCar2.0), four mammals (human, *Homo sapiens,* GRCh38.p7; mouse, *Mus musculus,* GRCm38.p5; opossum, *Monodelphis domestica,* monDom5; platypus, *Ornithorhynchus anatinusI,* OANA5), one insect (fruitfly, *Drosophila melanogaster*, BDGP6), one roundworm (*Caenorhabditis elegans*, WBcel235), one jawless vertebrate (sea lamprey, *Petromyzon marinus,* Pmarinus_7.0), and two urochordates (sea squirt, *Ciona intestinalis,* KH; sea squirt, *Ciona savignyi,* CSAV 2.0) and one fungus (yeast, *Saccharomyces cerevisiae*, R64-1-1). Sequence identifiers for each species are listed in Table 20.

**Syntenic analysis**

Using Biomart in the Ensembl database, genes located within 1.5 megabases of each *rab27* gene were identified. Synteny maps were constructed based on conserved patterns of gene locations for each of the *ra27* gene families, and consolidated results are presented in boxed regions in Figure 15. Construction of syntenic regions are based on genomic data mined from ten teleosts along with representative bird, amphibian, reptile, and mammal genomes for each *rab27* gene family.

**Alignment and phylogenetics**

Sixty-one sequences were aligned using ClustalW and Geneious Pro 6.0 (Biomatters Ltd). Sequences were virtually translated and verified to contain open reading frames. Model testing was performed for each of the four alignments, and the model with the best AICc value was chosen for the generation of the phylogenetic trees using Geneious 6.0. Using Mr.Bayes 3.1 and a GTR+I+G model of evolution, trees were generated for the full length coding sequence (714 bp) of *rab27*. The parameters used in the Mr. Bayes-generated trees included three gamma categories with unconstrained branch lengths. Markov Chain Monte Carlo methods were used for 1,100,000 steps with thinning every 200 steps, four heated chains, and a preheated chain temperature of 0.2. A burn-in length of 500 steps was used. Alternative models were tested using maximum likelihood (using 10,000 bootstrap replicates) and parsimony methods, and these provided similar topologies. Figures 14 shows the final tree generated for the alignment.

**dN/dS rates and identification of invariant codons**

To determine the percentage of codons that are invariant and experiencing extreme purifying selection, I calculated dN and dS values for the original alignment using MEGA6. "dN" is defined as the ratio of non-synonymous substitutions per non-synonymous site; "dS" is defined as the ratio of synonymous substitutions per synonymous site. Maximum likelihood reconstructions were generated using a Muse-Gaut model (Muse and Gaut 1994) of codon substitution and a general time reversible model (Nei and Kumar 2000) for nucleotide substitution. I counted the number of codons in an alignment that had dN and dS values of zero and divided this by the total number of codons in the alignment to determine the percentage of codons that are invariant and experiencing extreme purifying selection.

**Selection Tests**

I used the Datamonkey server and the HyPhy software package (Delport *et al.*, 2010; Kosakovsky Pond *et al.*, 2005) to test for purifying selection, positive selection, and episodic selection at the codon level and the branch level among the phylogenies that I generated. I used BUSTED (Branch site Unrestricted Statistical Test for Episodic Diversification) to assess whether episodic diversification occurs on at least one branch and at least at one site in the phylogeny.  The BUSTED test allows for varying rates of evolution (ω) applied to a constrained model of selection (null model) and an unconstrained model of selection (alternative model) using a Likelihood Ratio Test (LRT).  I tested my alignments using MEME (Mixed Effects Model of Evolution) and aBS-REL (adaptive Branch Site Random Effects Likelihood) tests.  MEME identifies the number of sites (codons) showing episodic diversifying selection.  Different evolutionary rates are allowed for each codon within an alignment.  Trees that were generated as described previously using the Geneious Software package were saved as Nexus files and uploaded to the Datamonkey Server to run the selection tests.  Methods for the tests I used in my analyses are further described in Murrell *et al.* (2012; MEME), Murrell et al. (2015; BUSTED), Smith et al. (2015; aBS-REL).  The aBS-REL test determined which branches in the phylogeny showed evidence of diversifying selection using a likelihood ratio test and $p \leq 0.05$.

## Results

### Phylogenetics and synteny

Using 61 sequences and 711 bases of DNA for my alignment, I generated a phylogenetic tree with a mix of low to high support values ranging from 0.56 to 1 (Figure 14). The *rab27a* genes formed a monophyletic clade with one branch containing only teleost genes and a sister branch containing a mix of fishes and tetrapods.  A lamprey *rab27* gene is grouped within this clade with a posterior probability value of 0.77 on the lamprey/cod node and posterior probability values of 0.56 and 0.57 on the two nodes that precede this node. Teleost *rab27b* and *rab27bb* genes sort into two distinct clades and in a position on the tree consistent with the genes having been duplicated in the fish specific genome duplication.

Syntenic analysis and ancestral chromosomal mapping placed the *rab27b* genes and other syntenic genes (green boxed regions) on segment A0 derived from ancestral vertebrate chromosome A (Figure 15, Panel A). The *rab27a* genes and other syntenic genes came from segment A4 (purple boxed region, Figure 15A). Both genes, *rab27a* and *rab27b*, were duplicated due to the 1R vertebrate genome duplication event approximately 550 million years ago. There were no identifiable duplicates of *rab27a* or *rab27b* due to the second whole genome duplication event in vertebrate evolutionary history (2R). However, I did find examples of duplicated and syntenic genes on ancestral vertebrate chromosomal segment A1. Genes that are syntenic among human, chicken, spotted gar and other non-teleosts are shown in the green boxed region in panel A. Many genes were lost in evolutionary time when comparing the black boxed region which is derived from segment A1 with the green boxed region derived from segment A0.

In teleosts, *rab27b* and *rab27bb* duplicates derived from ancestral teleost chromosome i and syntenic regions are shown in green boxed regions in Figure 15, panel B. Teleost *rab27a* genes derived from ancestral teleost chromosome j and syntenic genes are shown in purple boxed regions. Although duplicates for *rab27a* genes in teleosts were not found, other duplicated syntenic genes were identified (*nptna, prtgb*), and these are shown in the black boxed region in Figure 15, panel B.

dN/dS rate comparisons and invariant codons

I determined the dN/dS rates for each *rab27* clade for teleosts and non-teleost using 8-10 sequences per clade (Table 17). The dN/dS rates for teleosts were 0.18 (*rab27a*), 0.11(*rab27b*), and 0.24 (*rab27bb*). The dN/dS rates for non-teleosts were 0.11 (*rab27a*) and 0.13 (*rab27b*). The number of codons that were invariant for each clade varied from 19 (8.6% of 221 codons in the *rab27a* non-teleost clade) to 68 (31.48% of 216 codons in the *rab27b* teleost clade). In addition to finding the percentage of codons that were invariant, I found the percentage of codons where dN/dS = 0 (Table 17). This percentage includes codons that had zero non-synonymous substitutions (dN = n/N, where n = nonsynonymous substitutions and N = nonsynonymous site) but had a dS value greater than zero. The invariant codons are the codons in which both dN and dS are equal to zero. I found the percentage of codons with a dN/dS = 0 value to be between 39.57% (*rab27bb* teleosts) and 62.90% (*rab27a* non-teleosts). I found three codons out of 206 codons to be invariant across all 61 *rab27* sequences, which included fish and non-fish vertebrates, fly, *Ciona*, and fungus. The dN/dS rate for all 61 of these sequences was 0.2 and 10.19% of the 206 codons had a dN/dS = 0.

The evolutionary history for the *rab27* family along with the ω (dN/dS) values is shown in Figure 16. The extant *rab27* genes are placed such that *rab27a* and *rab27b* diverged from each other after the 1R vertebrate genome duplication event. Duplicates for these two genes that were a result of the R2 event were not identified in any of the species tested; thus, an “X” is placed where these missing duplicates should be (Figure 16). There is less rate variation among the non-teleost *rab* genes (0.11 to 0.13) compared to the rate variation that I identified among teleost *rab* genes (0.11 to 0.24). The *rab27bb* teleost clade had the highest ω value of 0.24 leading us to test this branch and other branches for selection.

**Selection**

I tested the *rab27* alignment (61 sequences and 237 sites) for evidence of diversifying selection at the codon level and at the branch level. Using BUSTED, a statistical test to determine whether episodic diversifying selection is taking place on at least one branch with at least one codon, I found evidence of selection in my phylogeny ($p = 0.002$). The unconstrained model provided three rate classes (Table 18) with 80.84% of the sites having $\omega_1 = 0$ (purifying selection), 14.67% of the sites having $\omega_2 = 0.18$ (purifying selection), and 4.49% of the sites having $\omega_3 = 2.23$ (positive selection). The constrained model provided three rate classes with 76.08% of the sites having $\omega_1 = 0$ (purifying selection), 13.53% of the sites having $\omega_2 = 0.03$ (purifying selection), and 10.39% of the sites having $\omega_3 = 1$ (neutral).

I tested 117 branches in my phylogeny for evidence of diversifying selection using the aBSREL test, which uses a likelihood ratio test (LRT). I found evidence of diversifying selection on five of the 117 branches tested (Table 19). The five branches that showed signs of selection with statistical support ($p < 0.05$) were stickleback *rab27bb*, medaka *rab27bb*, coelacanth *rab27b*, spotted gar *rab27b*, and the branch leading up to the teleost clade for *rab27bb* (Figure 17). All five of these branches had two rate classes along with other branches in my phylogeny (54 branches total with two rate classes). Of the 117 branches tested, 62 branches were identified as having one rate class and one branch was identified as having three rate classes.

All five of the branches identified as experiencing episodic diversifying selection were among the *rab27b* branches. Two of these were non-teleost species: spotted gar

(Figure 17A) and coelacanth (Figure 17B). The proportion of codons (or sites) within a specific rate class are shown on the y-axes of panels A, B, and C in Figure 17 and the ω for each rate class are shown on the x-axes of these panels. The teleost *rab27bb* branch (blue clade in tree) was found to be experiencing episodic diversifying selection with 87% of the sites having ω of 0.124 and 13% of the sites on this branch having an ω of 13.90 (Figures 17C and 17D). The table in Figure 17 shows the level of statistical support ($p < 0.05$) for each branch identified as experiencing episodic diversifying selection along with the proportion of sites on the branch that have a specific ω rate.

## Discussion

The evolutionary history and selection patterns for *rab27a* and *rab27b* gene families were characterized using phylogenetics, syntenic analysis, and selection assays. Newly identified in this study is a clade of *rab27bb* genes in teleosts that I found to exhibit episodic diversifying selection and higher dN/dS rates compared to the other *rab27* clades. Nearly 40% of the codons in this clade of eight teleosts have a dN/dS rate of zero with an additional 23% of the codons having a dN/dS rate of 0/0, representing invariant codons. Together, 63% of the 187 codons have zero non-synonymous substitutions for the *rab27bb* clade (Table 17). I suspect that the *rab27bb* clade in teleosts is functional due to the extent of purifying selection taking place among most of the codons for this gene family in addition to a large percentage of codons that are invariant when comparing teleost species that diverged approximately 148 million years ago. The combined rates for dN/dS = 0 and dN/dS = 0/0 are higher for the other two teleost clades (*rab27a* and *rab27b*). For *rab27a* in teleosts, the combined rate is over 73%, and for the *rab27b* clade the combined rate is over 82%, representing the percentage of codons having zero non-synonymous substitutions.

Based on my phylogenetics and syntenic analysis, the *rab27b* and *rab27bb* clades arose due to the fish specific genome duplication event thought to have taken place over 300 million years ago (Figure 14). The dN/dS rates for these two clades show that one clade (*rab27b*) is more conserved with dN/dS = 0.11, compared to the non-duplicated clade (*rab27a*) which had a dN/dS = 0.18. However, the other duplicated clade (*rab27bb*) is less well conserved with a dN/dS = 0.24 showing that there are more non-synonymous substitutions within that clade compared to the non-duplicated *rab27a* clade.

The non-teleost clades which represent a more diverse taxonomic sampling than the teleost clades and includes organisms that diverged from a common ancestor over 400 million years ago (http://www.timetree.org) show generally low dN/dS rates for both duplicates. The non-teleost clades have a dN/dS rate of 0.11 for the *rab27a* clade and a dN/dS rate of 0.13 for the *rab27b* clade. For the non-teleost clades, the duplication of these *rab27a* and *rab27b* genes took place approximately 550 million years ago consequent to the first (1R) vertebrate genome duplication event (Figure 15). I identify two relatively similar rates of evolution for gene clades that diverged approximately 550 million years ago.

I identified synteny to be more maintained among orthologous chromosomal regions compared to paralogous chromosomal regions. For the green boxed region in Figure 15A, the genes listed are found among human, chicken, spotted gar, and most other non-teleost vertebrates. There is a high degree of synteny in the green boxed region among these orthologous species which diverged from a common ancestor approximately 450 million years ago. However, only a couple of these genes are found on their paralogous chromosomes which diverged consequent to the 1R genome duplication event. It appears that there was a large amount of chromosomal rearrangement taking place after the 1R event and before the divergence of the species examined in this study as identified in my syntenic analysis and that there continued to be a high degree of conservation at the individual gene level as identified by my dN/dS rates for the *rab27a* and *rab27b* clades in non-teleosts.

The teleost duplicated genes *rab27b* and *rab27bb*, which represent a duplication event taking place approximately 300-350 million years ago, have a larger amount of evolutionary rate variation (0.11 and 0.24) compared to the non-teleost rate variation from a duplication event taking place approximately 550 million years ago (0.11 for *rab27a* and 0.13 for *rab27b*). These evolutionary rate variations may indicate that one of the duplicated clades (*rab27bb*) has been subjected to selective forces in a way that has led to new functional roles for this duplicated gene (Figure 16). In (Opazo et al. 2013)

The *rab27bb* clade in teleost was shown to have episodic diversifying selection present (p = 0.0005) with 13% of the sites on that branch subject to positive selection and 87% of the sites on that branch subject to purifying selection with a dN/dS rate of 0.124. I suspect the teleost clade of *rab27bb* genes has acquired a new function due to the strong statistical support for episodic diversifying selection taking place along this branch of genes.

Since Rab27b has been shown to bind with a melanophilin and myosin 5a complex (Strom et al. 2002), I suspect that the teleost *rab27bb* gene and *myo5ab* genes may be coevolving. In chapter one, I discussed the evolution of *myo5a* which included a fish specific duplication leading to teleost genes *myo5aa* and *myo5ab.* The data presented in chapter one shows that teleost *myo5aa* gene is more well conserved (higher percentage of invariant codons) and evolving at a slower rate (lower dN/dS rates) compared to the *myo5ab* teleost gene. I suspect that the duplicates that are more highly conserved (*rab27b* and *myo5aa*) are potentially interacting with each other, coevolving, and subject to a larger amount of purifying selection. Further, I suspect that the duplicates that are less well conserved (*rab27bb* and *myo5ab*) are possibly co-evolving new functional roles.

Using phylogenetics and syntenic analysis, I identified a new *rab27bb* clade present only in teleost fish. Future studies may identify whether the *rab27bb* genes are functional in some teleosts. New functional roles or a partitioning of functionality (spatially or temporally) may be elucidated with further expression studies in addition to the creation of knockout or knockin mutants. Comparisons among species that have different combinations of these duplicated genes may also help unravel the molecular relationships that have evolved among the Myosin and Rab proteins. Further experiments may further support the idea of coevolution taking place among duplicated genes that evolve at different rates.

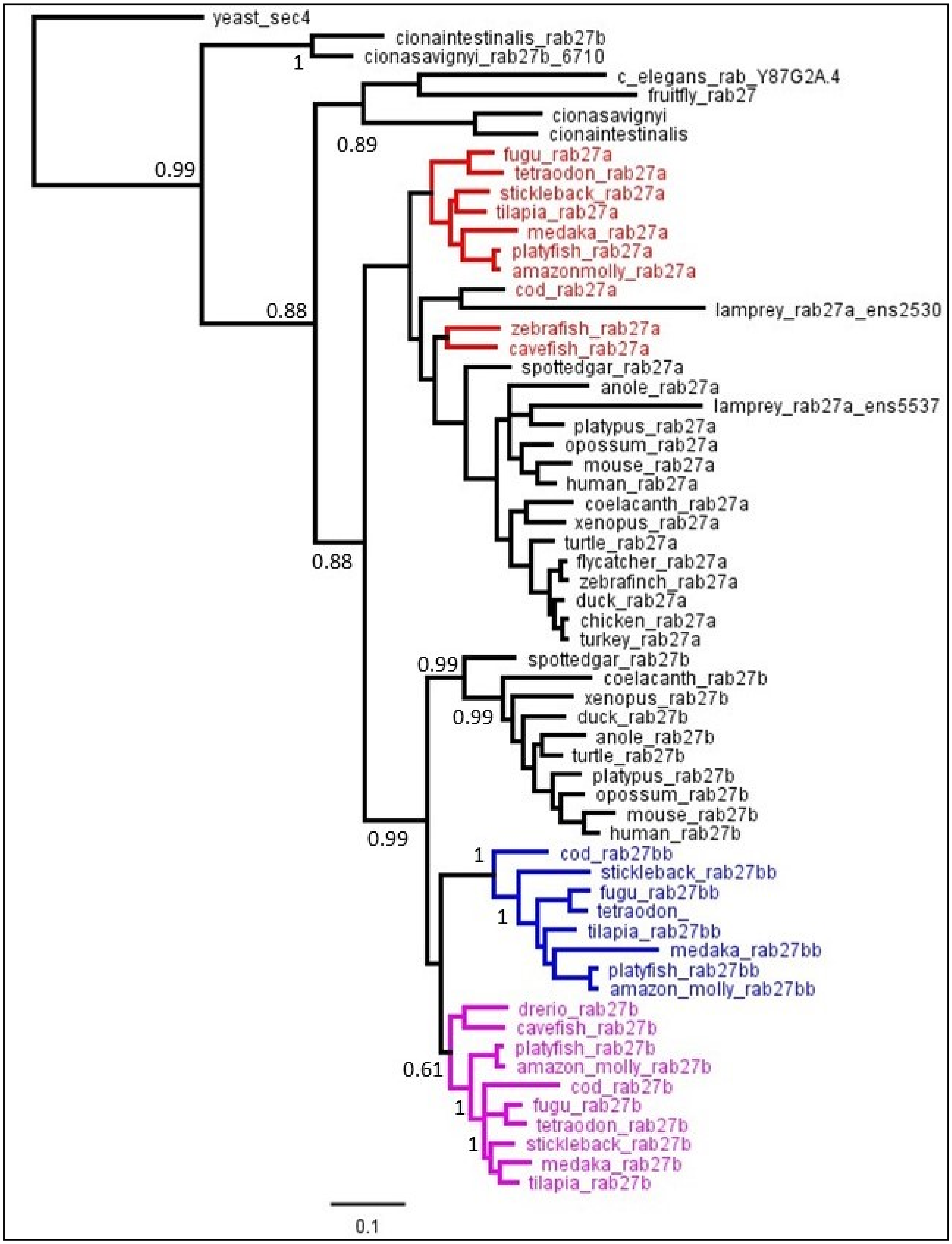


**Figure 14. Bayesian tree for *rab27a* and *rab27b* gene families.**

**Figure 14. Bayesian tree for *rab27a* and *rab27b* gene families (Previous page).** Teleost genes are colored as follows: *rab27b* is in purple (bottom most clade), *rab27bb* is in blue (2nd clade from bottom), and *rab27a* is in red (clade near the top). Posterior probability values are provided for select nodes. Posterior probability values range from 0.56 to 1 for nodes without a value shown.

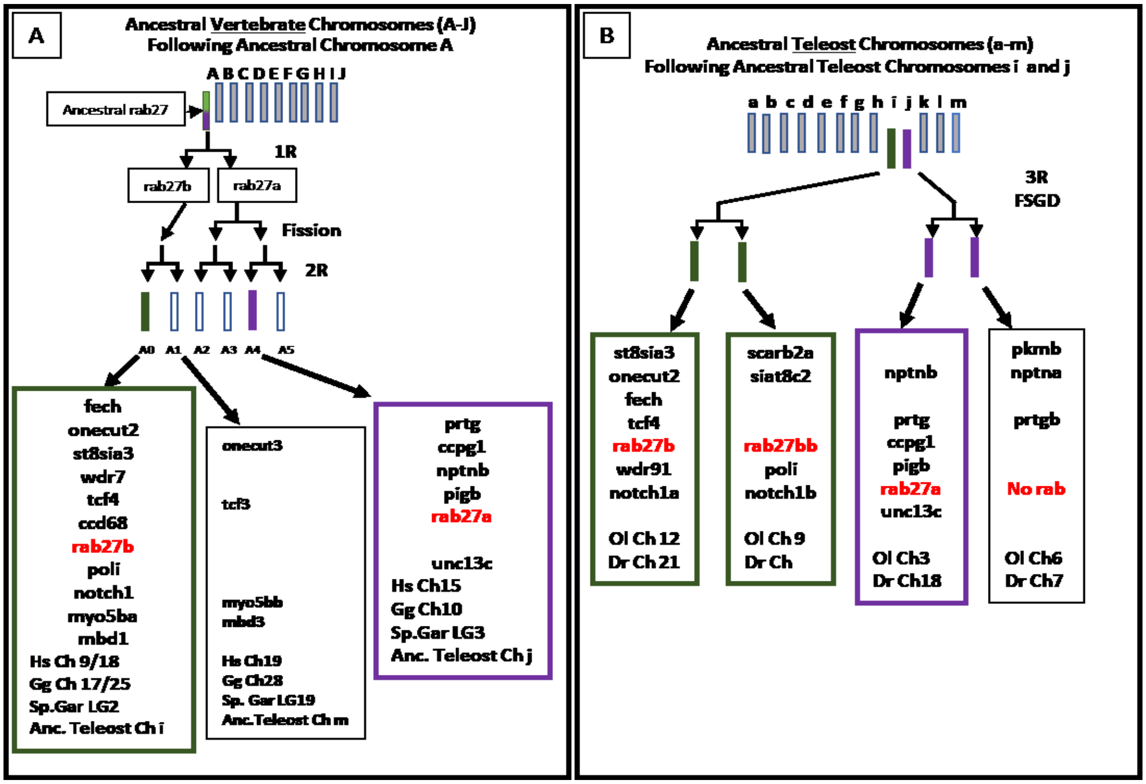


**Figure 15. Ancestral chromosomes and synteny for *rab27*.** An ancestral *rab27* gene on ancestral vertebrate chromosome A was duplicated after 1R leading to the *rab27a* and *rab27b* genes (panel A). After the 2R event, duplicates were lost but one copy of each *rab* gene remained on segment A0 and A4. Syntenic and other select genes are shown in the boxed regions. In addition, the chromosomal locations or linkage groups are shown for human (Hs), chicken (Gg), spotted gar (Sp.Gar), ancestral teleost (Anc.teleost), medaka (Ol), and zebrafish (Dr). Genes listed in the green boxed region in panel A are syntenic among humans, chicken, spotted gar, and other non-teleosts. However, paralogous synteny is not as well maintained when comparing the green boxed region coming from segment A0 to the black boxed region coming from segment A1. After a fish specific genome duplication (FSGD), also annotated as 3R, two *rab27b* duplicates remain. Two *rab27b* genes have been identified in eight of the ten teleosts used in this study.

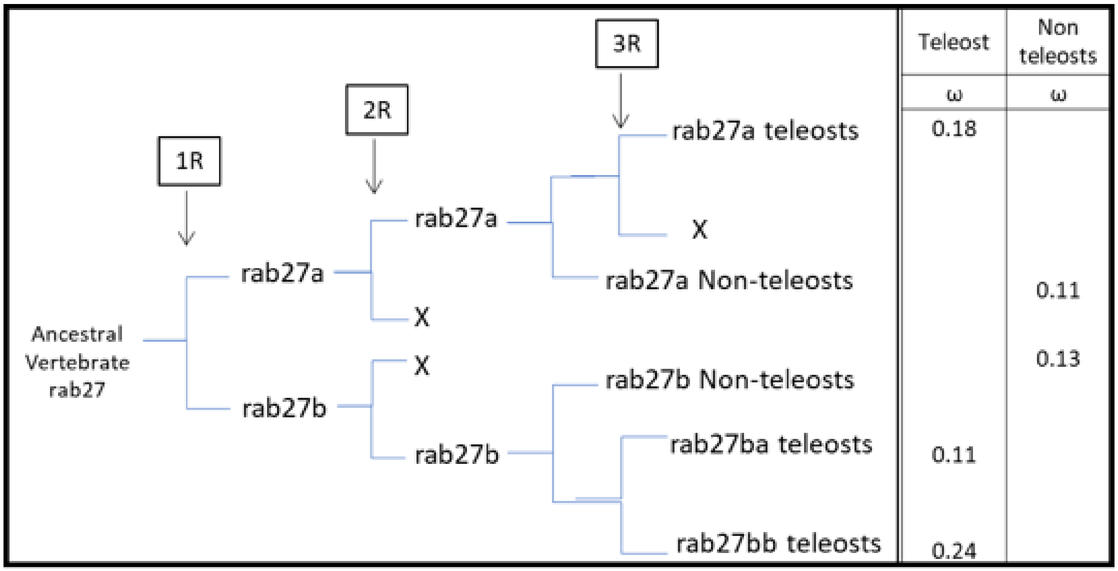
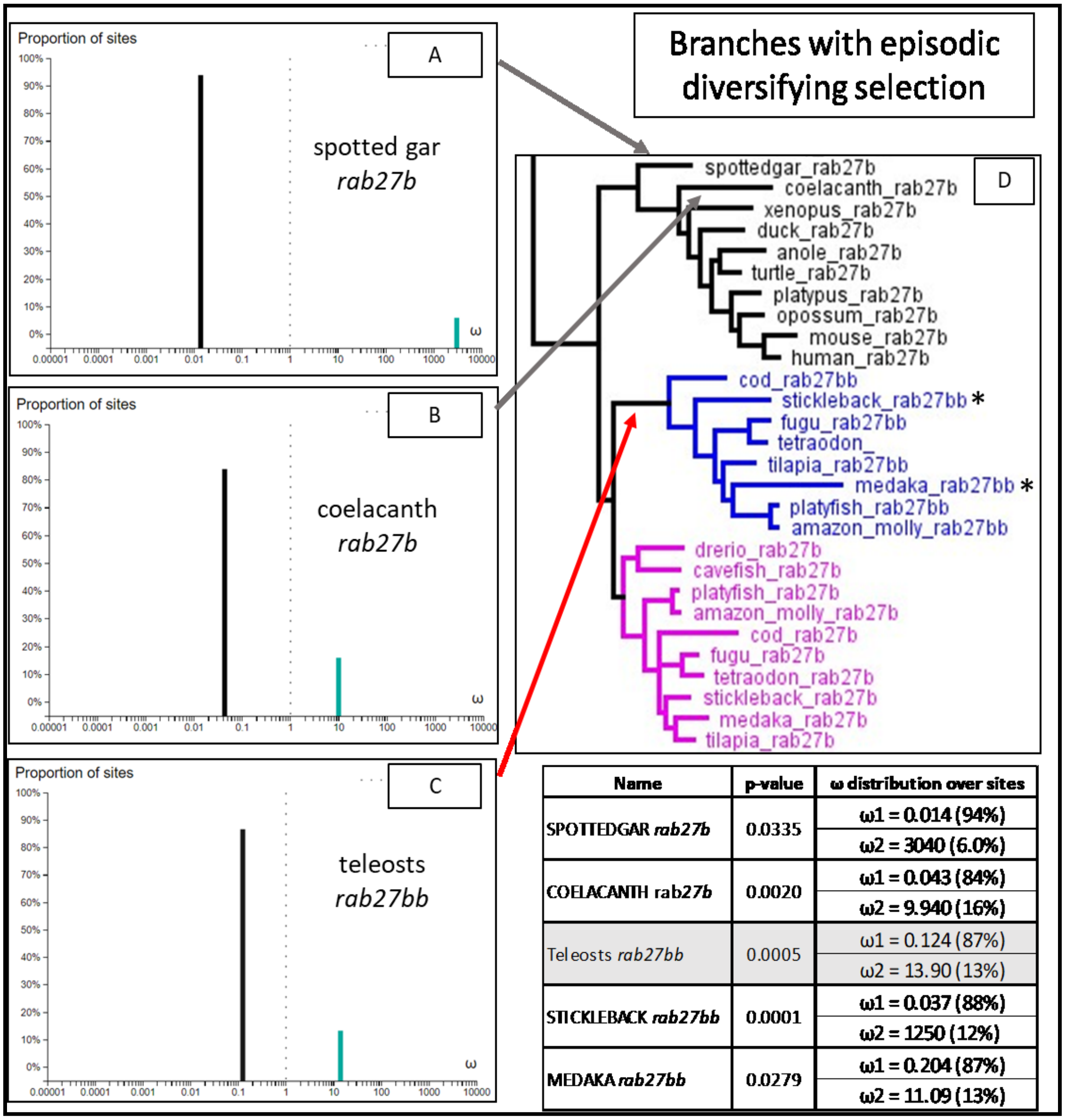


**Figure 16. Evolutionary history of *rab27.*** Evolutionary history of *rab27* family along with a comparison of ω (dN/dS rates) for duplicated teleost or non-teleost clades. An "X" denotes where a duplicate of *rab27* is expected based on genome duplication events but was not found in any species examined in this study. The non-teleost ω values are closer to each other (0.11 vs. 0.13) compared to the teleost ω values (0.11 vs. 0.24).



| Name | p-value | ω distribution over sites |
|---|---|---|
| SPOTTEDGAR *rab27b* | 0.0335 | ω1 = 0.014 (94%) |
| | | ω2 = 3040 (6.0%) |
| COELACANTH rab*27b* | 0.0020 | ω1 = 0.043 (84%) |
| | | ω2 = 9.940 (16%) |
| Teleosts *rab27bb* | 0.0005 | ω1 = 0.124 (87%) |
| | | ω2 = 13.90 (13%) |
| STICKLEBACK *rab27bb* | 0.0001 | ω1 = 0.037 (88%) |
| | | ω2 = 1250 (12%) |
| MEDAKA *rab27bb* | 0.0279 | ω1 = 0.204 (87%) |
| | | ω2 = 11.09 (13%) |

**Figure 17. Diversifying selection among five branches of *rab27* alignment.**

**Figure 17. Diversifying selection among five branches of *rab27* alignment (Previous page).** Five branches in my phylogeny showed evidence of diversifying selection ($p < 0.05$) using a likelihood ratio test. Using an adaptive branch site random effects likelihood (aBS-REL) test, two rate classes ($\omega_1$ and $\omega_2$) were found for each of the five branches. Panels A and B show the proportion of sites along a branch that have a rate class $\omega_1$ representing purifying selection and another rate class ω2, representing positive selection. Panel C shows the branch leading up to teleosts (red arrow panel D) is subject to two rate classes $\omega_1 = 0.124$ (purifying selection) for 87% of the sites on this branch and $\omega_2 = 13.90$ (positive selection) for 13% of the sites on this branch. The table shows the ω values for the two rate classes along each branch and the percentage of sites on that branch that are represented by the listed rate class ω. Diversifying selection in the form of two rate classes was also identified for stickleback and medaka *rab27bb* (asterisks in panel D and rate values in table).

**Table 17. Percentage of *rab27* codons invariant or with dN/dS=0**. The percentage of invariant codons for each *rab27* clade ranged from 8.6% to 31.5%. Each select group of organisms and sequences represent divergence from a common ancestor from 148 to 435 million years ago, for non-teleosts and teleosts, respectively. Rates of evolution (dN/dS values) for each select clade ranged from 0.11 to 0.24.

| | # Sequ. | # codons | #codons dN/dS=0 | # codons invariant | dN/dS | % codons dN/dS=0 | % codons invariant |
|---|---|---|---|---|---|---|---|
| rab27a teleosts | 8 | 212 | 109 | 47 | 0.18 | 51.42 | 22.17 |
| rab27b teleosts | 8 | 216 | 110 | 68 | 0.11 | 50.93 | 31.48 |
| rab27bb teleosts | 8 | 187 | 74 | 43 | 0.24 | 39.57 | 22.99 |
| rab27a non-teleosts | 10 | 221 | 139 | 19 | 0.11 | 62.90 | 8.60 |
| rab27b non-teleosts | 10 | 218 | 114 | 28 | 0.13 | 52.29 | 12.84 |
| rab27 (All codons) | 61 | 206 | 21 | 3 | 0.2 | 10.19 | 1.46 |

**Table 18. Diversifying selection identified in *rab27* alignment using BUSTED.** The unconstrained model provided three rate classes and the percentage of sites within each rate class, $\omega_1$=0 ,80.84% of sites; $\omega_2$=0.18, 14.67% of sites; and $\omega_3$=2.23, 4.49% of sites. A likelihood ratio test was used showing the unconstrained model (mix of purifying and positive selection) had stronger statistical support than the constrained model which represents the null model or neutral evolution for some of the sites and purifying selection for others (p=0.002).

| Model | log L | # parameters | AICc | ω1 | ω2 | ω3 |
|---|---|---|---|---|---|---|
| Unconstrained model | -19501.8 | 136 | 39278.1 | 0.00 (80.84%) | 0.18 (14.67%) | 2.23 (4.49%) |
| Constrained model | -19508.1 | 135 | 39288.9 | 0.00 (76.08%) | 0.03 (13.53%) | 1.00 (10.39%) |

**Table 19. Diversifying selection among five branches for *rab27* alignment.** Five branches out of 117 branches in my phylogeny showed evidence of diversifying selection with two rate classes. Overall, 62 branches in my phylogeny are shown to have one rate class, 54 branches are shown to have two rate classes, and one branch was shown to have three rate classes (aBSREL, LRT, $p < 0.05$)

| ω rate classes | # of branches | % of branches | % of tree length | # branches under selection |
|---|---|---|---|---|
| 1 | 62 | 53% | 0.13% | 0 |
| 2 | 54 | 46% | 100% | 5 |
| 3 | 1 | 0.85% | 0.04% | 0 |

**Table 20. *rab27* genes and gene identifiers used in this study.** Organismal names, *rab27* gene names, gene identifiers and location, number of nucleotides for the RNA transcripts, the number of amino acids for the proteins, and the number of exons for the gene are provided.

| Organism | Gene | Gene Id | Genomic location | Tran. size | Prot. size (aa) | # exons |
|---|---|---|---|---|---|---|
| Zebrafish | rab27a | ENSDARG00000103935 | Chromosome 18: 1,353,767 | 1474 | 222 | 6 |
| Zebrafish | rab27b | ENSDARG00000087762 | Chromosome 21: 1,504,774 | 3425 | 224 | 5 |
| Fugu | rab27a | ENSTRUG00000017930 | scaffold_14: 1,513,283 | 663 | 220 | 5 |
| Fugu | rab27b | ENSTRUG00000003528 | scaffold_216: 215,359 | 771 | 256 | 5 |
| Fugu | rab27bb | ENSTRUG00000016761 | scaffold_4: 3,182,421 | 771 | 256 | 5 |
| Tetraodon | rab27a | ENSTNIG00000009317 | Chromosome 5: 5,295,870 | 657 | 218 | 5 |
| Tetraodon | rab27b | ENSTNIG00000010328 | Chromosome Un_random: 23,567,018 | 651 | 216 | 5 |
| Tetraodon | rab27bb | ENSTNIG00000004112 | Chromosome Un_random: 66,627,237 | 1140 | 214 | 6 |
| medaka | rab27a | ENSORLG00000008731 | Chromosome 3: 20,383,618 | 663 | 220 | 5 |
| medaka | rab27b | ENSORLG00000014981 | Chromosome 12: 28,441,898 | 981 | 218 | 7 |
| medaka | rab27bb | ENSORLG00000000417 | Chromosome 9: 717,130 | 642 | 213 | 7 |
| cod | rab27a | ENSGMOG00000002736 | GeneScaffold_3322: 117,673 | 651 | 216 | 5 |

**Table 20 (continued).** ***rab27*** **genes and gene identifiers used in this study.**

| | | | | | | |
|---|---|---|---|---|---|---|
| cod | rab27b | ENSGMOG00000011291 | GeneScaffold_3652: 17,873 | 378 | 125 | 3 |
| cod | rab27bb | ENSGMOG00000017684 | GeneScaffold_1580: 7,341 | 573 | 191 | 6 |
| platyfish | rab27a | ENSXMAG00000014238 | Scaffold JH556861.1: 452,592 | 666 | 221 | 6 |
| platyfish | rab27b | ENSXMAG00000011803 | Scaffold JH557267.1: 52,863 | 4949 | 218 | 6 |
| platyfish | rab27bb | ENSXMAG00000003211 | Scaffold JH556717.1: 1,428,615 | 678 | 225 | 5 |
| stickleback | rab27a | ENSGACG00000015855 | groupII: 11,586,268 | 672 | 223 | 6 |
| stickleback | rab27b | ENSGACG00000017867 | groupXIV: 10,149,089 | 657 | 218 | 6 |
| stickleback | rab27bb | ENSGACG00000013851 | groupXIII: 17,400,861 | 821 | 220 | 6 |
| tilapia | rab27a | ENSONIG00000005740 | Scaffold GL831150.1: 2,579,061 | 4150 | 220 | 6 |
| tilapia | rab27b | ENSONIG00000012725 | Scaffold GL831141.1: 3,018,247 | 1499 | 218 | 6 |
| tilapia | rab27bb | ENSONIG00000016904 | Scaffold GL831307.1: 22,470 | 675 | 224 | 5 |
| amazon molly | rab27a | ENSPFOG00000001824 | Scaffold KI519973.1: 128,952 | 3116 | 220 | 5 |
| amazon molly | rab27b | ENSPFOG00000010261 | Scaffold KI520009.1: 458,255 | 5017 | 218 | 7 |
| amazon molly | rab27bb | ENSPFOG00000006245 | Scaffold KI519702.1: 685,678 | 1127 | 219 | 5 |
| cavefish | rab27a | ENSAMXG00000021170 | Scaffold KB882149.1: 2,580,374 | 1339 | 222 | 5 |
| cavefish | rab27b | ENSAMXG00000009878 | Scaffold KB871656.1: 558,672 | 3647 | 217 | 6 |
| spotted gar | rab27a | ENSLOCG00000013450 | Chromosome LG3: 40,998,843 | 672 | 223 | 6 |
| spotted gar | rab27b | ENSLOCG00000003498 | Chromosome LG2: 9,097,415 | 847 | 223 | 6 |
| coelacanth | rab27a | ENSLACG00000014087 | Scaffold JH126859.1: 958,835 | 1885 | 221 | 7 |
| coelacanth | rab27b | ENSLACG00000013259 | ScaffoldJH126777.1: 830,534 | 66 | 220 | 7 |

**Table 20 (continued).** ***rab27*** **genes and gene identifiers used in this study.**

| | | | | | | |
|---|---|---|---|---|---|---|
| xenopus | rab27a | ENSXETG00000007375 | Scaffold GL173632.1 112,712 | 1904 | 221 | 6 |
| xenopus | rab27b | ENSXETG00000003366 | Scaffold GL172733.1: 1,177,701 | 1922 | 218 | 8 |
| chicken | rab27a | ENSGALG00000004435 | Chromosome 10: 7,968,909 | 1913 | 221 | 6 |
| duck | rab27a | ENSAPLG00000002447 | Scaffold KB742572.1: 108,437 | 759 | 221 | 5 |
| duck | rab27b | ENSAPLG00000010537 | Scaffold KB742564.1: 361,477 | 998 | 215 | 6 |
| flycatcher | rab27a | ENSFALG00000012003 | Scaffold JH603201.1: 5,597,706 | 3068 | 221 | 6 |
| turkey | rab27a | ENSMGAG00000006007 | Chromosome 12: 7,721,098 | 1656 | 221 | 6 |
| zebrafinch | rab27a | ENSTGUG00000006508 | Chromosome 10: 7,711,439 | 666 | 221 | 5 |
| anole | rab27a | ENSACAG00000006873 | Contig AAWZ02036658: 721 | 513 | 170 | 4 |
| anole | rab27b | ENSACAG00000008824 | Scaffold GL343213.1: 321,304 | 657 | 218 | 5 |
| turtle | rab27a | ENSPSIG00000011050 | Scaffold JH205794.1: 40,671 | 3377 | 221 | 5 |
| turtle | rab27b | ENSPSIG00000014730 | Scaffold JH206968.1: 6,028,229 | 1530 | 218 | 6 |
| opossum | rab27a | ENSMODG00000007980 | Chromosome 1: 166,255,227 | 2356 | 221 | 5 |
| opossum | rab27b | ENSMODG00000020460 | Chromosome 3: 252,327,309 | 3635 | 218 | 6 |
| platypus | rab27a | ENSOANG00000001183 | UltraContig Ultra366: 1,170,011 | 1537 | 181 | 4 |
| platypus | rab27b | ENSOANG00000002261 | Chromosome 3: 25,762,628 | 3047 | 218 | 6 |
| mouse | rab27a | ENSMUSG00000032202 | Chromosome 9: 73,044,854 | 3168 | 221 | 6 |
| mouse | rab27b | ENSMUSG00000024511 | Chromosome 18: 69,979,131 | 6891 | 218 | 7 |
| human | rab27a | ENSG00000069974 | Chromosome 15: 55,202,966 | 3549 | 221 | 6 |
| human | rab27b | ENSG00000041353 | Chromosome 18: 54,717,860 | 7281 | 218 | 6 |
| yeast | SEC4 | YFL005W | Chromosome VI: 130,334 | 648 | 215 | 1 |

**Table 20 (continued).** ***rab27*** **genes and gene identifiers used in this study.**

| | | | | | | |
|---|---|---|---|---|---|---|
| lamprey | rab27 | ENSPMAG00000005537 | Scaffold GL476596: 510,564 | 753 | 219 | 5 |
| lamprey | rab27 | ENSPMAG00000002530 | Scaffold GL485242: 6,306 | 525 | 174 | 3 |
| fruitfly | rab27 | FBgn0025382 | Chromosome X: 1,473,884 | 1564 | 230 | 2 |
| *Ciona savignyi* | rab27a | ENSCSAVG00000003516 | reftig_41: 481,888 | 932 | 226 | 7 |
| *Ciona savignyi* | rab27 | ENSCSAVG00000006710 | reftig_60: 660,220 | 1186 | 220 | 5 |
| *Ciona intestinalis* | rab27a | ENSCING00000007626 | Chromosome 4: 2,240,648 | 833 | 233 | 8 |
| *Ciona intestinalis* | rab27b | ENSCING00000019994 | Chromosome 5: 4,261,517 | 1140 | 232 | 6 |
| *C.elegans* | aex-6 | WBGene00000089 | Chromosome I: 13,543,079 | 961 | 215 | 6 |

# V. CONCLUSIONS

In the work presented herein, I determined whether a common theme existed among evolutionary rates for duplicated genes versus non-duplicated genes using *myo5, rab11,* and *rab27*. I found evolutionary rate variation among duplicated gene families for duplicated genes versus the rates of evolution of non-duplicated orthologs. In three of the six cases studied, both duplicated clades were shown to have higher dN/dS rates compared to an orthologous non-duplicated clade. In the other three cases, there was evidence of one duplicate experiencing a faster rate of evolution (*myo5ab, myo5bb,* or *rab27bb*) and the other duplicate experiencing a slower rate of evolution (*myo5aa, myo5ba,* or *rab27b*) compared to the orthologous non-duplicated genes (*myo5a, myo5b*, or *rab27a*) in non-teleosts (Table 21).

When considering the rates of two duplicated genes versus the rate of non-duplicated orthologs, it might be natural to expect one of the duplicated genes to have a higher evolutionary rate than the other duplicated gene. However, based on this summary of six cases, it seems like one of the duplicate genes has a well conserved evolutionary rate. Meaning, one of the duplicated genes for each clade has a rate near or below 0.1. I suspect this to be a functional duplicated gene with such a low evolutionary rate and high level of sequence conservation. My prediction for the other duplicated gene with a higher evolutionary rate is that it has had its evolutionary constraints lifted. I suspect the selective pressures have been loosened for the duplicated genes with higher evolutionary rates, allowing for the possibility of a new function for that duplicated gene.

**Table 21. Rate comparisons for duplicated genes.** Rate comparisons for duplicated genes from different gene families (*myo5, rab11,* and *rab27*) from 2R and 3R duplication events. In comparing a clade that has been duplicated to a clade that wasn't duplicated or lost its duplicate, I find duplicate 2 always evolving at a faster rate (up arrow last column) than the non-duplicated clade. The other duplicate (duplicate 1) evolves at a slower rate in three out of the six scenarios (down arrow) or at a faster rate in the other three out of six scenarios compared to the non-duplicated clade.

| Case | # sequences | organism | divergence time organisms (MYA) | Dupl. event | duplicated genes rates, ω<br>dupl.1 \| dupl.2 | | Non-dupl. genes rates ω | |
|---|---|---|---|---|---|---|---|---|
| (1) | 6 | non-teleosts | 413-435 | 2R | myo5ba<br>0.05 | myo5bb<br>0.24 | myo5a<br>0.09 | ↓↑ |
| (2) | 8 | non-teleosts | 229-435 | 2R | rab11a<br>0.03 | rab11a1<br>0.19 | rab11b<br>0.01 | ↑↑ |
| (3) | 8 | teleosts/ non-teleosts | 229-435 | 3R | myo5aa<br>0.05 | myo5ab<br>0.12 | myo5a<br>0.09 | ↓↑ |
| (4) | 6 | teleosts/ non-teleosts | 229-435 | 3R | myo5ba<br>0.06 | myo5bb<br>0.08 | myo5b<br>0.05 | ↑↑ |
| (5) | 8 | teleosts | 148 | 3R | rab27b<br>0.11 | rab27bb<br>0.24 | rab27a<br>0.18 | ↓↑ |
| (6) | 8 | teleosts | 229 | Post 3R | rab11ba1<br>0.03 | rab11ba2<br>0.05 | rab11bb<br>0.02 | ↑↑ |

# REFERENCES

Anisimova M, Yang Z. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Molecular Biology and Evolution. 24:1219–1228.

Brennecke J, et al. 2005. Principles of MicroRNA–Target Recognition. PLOS Biology. 3:e85.

Bian C et al. 2016. The Asian arowana (Scleropages formosus) genome provides new insights into the evolution of an early lineage of teleosts. Scientific Reports. 6:1–17.

Bielawski JP, Yang Z. 2004. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. Journal of Molecular Evolution. 59:121–132.

Bielawski JP, Yang Z. 2003. Maximum likelihood methods for detecting adaptive evolution after gene duplication. Journal of Structural and Functional Genomics. 3:201–212.

Braasch I, Schartl M, Volff J-N. 2007. Evolution of pigment synthesis pathways by gene and genome duplication in fish. BMC Evolutionary Biology. 7:74.

Braasch I, Brunet F, Volff J-N, Schartl M. 2009a. Pigmentation Pathway Evolution after Whole-Genome Duplication in Fish. Genome Biology and Evolution. 1:479–493.

Braasch I, Liedtke D, Volff J, Schartl M. 2009b. Pigmentary function and evolution of tyrp1 gene duplicates in fish. Pigment Cell & Melanoma Research. 22:839–850.

Catchen JM, Conery JS, Postlethwait JH. 2009. Automated identification of conserved synteny after whole-genome duplication. Genome Research. 19:1497–1505.

Coureux P-D et al. 2003. A structural state of the myosin V motor without bound nucleotide. Nature. 425:419–423.

Delport W, Poon AF, Frost SDW and Kosakovsky Pond SL 2010.  Datamonkey 2010: A suite of phylogenetic analysis tools for evolutionary biology.  Bioinformatics. 26(19): 2455-2457.

Diekmann Y et al. 2011. Thousands of Rab GTPases for the cell biologist. PLoS Computational Biology. 7(10):  e1002217.

Force A et al. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 151:1531–1545.

Goh CS, Cohen FE. 2002. Co-evolutionary analysis reveals insights into protein–protein interactions. Journal of molecular biology. 324(1):177–192

Hammer JA, Wagner W. 2013. Functions of class v myosins in neurons. Journal of Biological Chemistry. 288:28428–28434.

Hammer JA, Wu X. 2007. Slip sliding away with myosin V. Proceedings of the National Academy of Sciences. 104:5255–5256.

Hodel C et al. 2014. Myosin VIIA is a marker for the cone accessory outer segment in zebrafish. Anatomical Record. 297:1777–1784.

Hughes, A. 1999. Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. Journal of Molecular Evolution 48(5): 565-576.

Jaillon O et al. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature. 431:946–957.

Karcher RL et al. 2001. Cell cycle regulation of myosin-V by calcium/calmodulin-dependent protein kinase II. Science. 293:1317–1320.

Kosakovsky Pond SL and Frost SDW 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments  Bioinformatics 21(10): 2531-2533

Kosakovsky Pond SL and Frost SDW 2005.  Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. Molecular Biology and Evolution 22(5): 1208-1222

Kosakovsky Pond SL et al. 2011. A random effects branch-site model for detecting episodic diversifying selection. Molecular Biology and Evolution. 28(11): 3033-3043

Krek A et al. 2005. Combinatorial microRNA target predictions. Nature Genetics. 37:495.

Kuraku S. 2013. Impact of asymmetric gene repertoire between cyclostomes and gnathostomes. Seminars in Cell & Developmental Biology. 24:119–127.

Kuraku S. 2010. Palaeophylogenomics of the Vertebrate Ancestor—Impact of Hidden Paralogy on Hagfish and Lamprey Gene Phylogeny. Integrative and Comparative Biology. 50:124–129.

Lewis BP, Burge CB, Bartel DP. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 120:15–20.

Li JF, Nebenführ A. 2008. The tail that wags the Dog: The globular Tail Domain defines the function of Myosin V/XI. Traffic. 9:290–298.

Li WH. 1980. Rate of gene silencing at duplicate loci. A theoretical study and interpretation of data from tetraploid fishes. Genetics. 95:237–258.

Lin-Jones J, Sohlberg L, Dose A, Breckler J, Hillman DW, Burnside B. 2009. Identification and localization of myosin superfamily members in fish retina and retinal pigmented epithelium. J Comp Neurol 513:209–223.

Mellgren EM, Johnson SL. 2005. kitb, a second zebrafish ortholog of mouse Kit . Development Genes and Evolution. 215:470–477.

Mills MG, Nuckels RJ, Parichy DM. 2007. Deconstructing evolution of adult phenotypes: genetic analyses of kit reveal homology and evolutionary novelty during adult pigment pattern development of Danio fishes. Development. 134:1081–1090.

Murrell B et al. 2012. Detecting Individual Sites Subject to Episodic Diversifying Selection. PLoS Genetics 8(7): e1002764

Murrell B et al. 2015. Gene-Wide identification of episodic selection. Molecular Biology and Evolution. 32:1365–1371.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Molecular Biology and Evolution. 11:715–724.

Nakatani Y, Takeda H, Kohara Y, Morishita S. 2007. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. Genome Research. 17:1254–1265.

Nascimento AAC, Amaral RG, Bizario JCS, Larson RE, Espreafico EM. 1997. Subcellular Localization of Myosin-V in the B16 Melanoma Cells, a Wild-type Cell Line for the  dilute Gene. Molecular Biology of the Cell. 8:1971–1988.

Nei M, Kumar S. 2000. Molecular Evolution and Phylogenetics. Oxford University Press.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics. 148:929–936.

Ohno S. 1970. *Evolution by Gene Duplication*. Springer Berlin Heidelberg: Berlin, Heidelberg .

Opazo JC, Butts GT, Nery MF, Storz JF, Hoffmann FG. 2013. Whole-genome duplication and the functional diversification of teleost fish hemoglobins. Molecular Biology and Evolution. 30:140–153.

Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein-protein interaction. Journal of Molecular Biology. 1997;271(4):511–523

Pereira-Leal JB. 2008. The Ypt/Rab family and the evolution of trafficking in fungi. Traffic. 9:27–38.

Putnam NH et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. Nature. 453:1064–1071.

Pylypenko O et al. 2013. Structural basis of myosin V Rab GTPase-dependent cargo recognition. Proceedings of the National Academy of Sciences of the United States of America. 110:20443–8.

Qiu H, Hildebrand F, Kuraku S, Meyer A. 2011. Unresolved orthology and peculiar coding sequence properties of lamprey genes: the KCNA gene family as test case. BMC Genomics. 12:325.

Ramakrishnan C, Dani VS, Ramasarma T. 2002. A conformational analysis of Walker motif A [GXXXXGKT (S)] in nucleotide-binding and other proteins. Protein Engineering Design and Selection. 15:783–798.

Rodriguez OC, Cheney RE. 2002. Human myosin-Vc is a novel class V myosin expressed in epithelial cells. Journal of Cell Science. 115:991–1004.

Sittaramane V, Chandrasekhar A. 2008. Expression of unconventional myosin genes during neuronal development in zebrafish. Gene Expression Patterns. 8:161–170.

Smith MD et al. 2015. Less is more: An adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Molecular Biology and Evolution. 32:1342–1353.

Sonal et al. 2014. Myosin Vb Mediated Plasma Membrane Homeostasis Regulates Peridermal Cell Size and Maintains Tissue Homeostasis in the Zebrafish Epidermis. PLoS Genetics. 10:1-19

Stenmark H, Olkkonen VM. 2001. The Rab GTPase family. Genome Biology. 2:Reviews3007.

Strom M, Hume AN, Tarafder AK, Barkagianni E, Seabra MC. 2002. A family of Rab27-binding proteins. Melanophilin links Rab27a and myosin Va function in melanosome transport. The Journal of Biological Chemistry. 277:25423–25430.

Swiatecka-Urban A et al. 2007. Myosin Vb is required for trafficking of the cystic fibrosis transmembrane conductance regulator in Rab11a-specific apical recycling endosomes in polarized human airway epithelial cells. Journal of Biological Chemistry. 282:23725–23736.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular evolutionary genetics analysis version 6.0. Molecular Biology and Evolution. 30:2725–2729.

Taylor JS, Peer Y Van De, Braasch I, Meyer A. 2001. Comparative genomics provides evidence for an ancient genome duplication event in fish. Philosophical Transactions of the Royal Society of London. 356:1661–1679.

Trybus KM. 2008. Myosin V from head to tail. Cellular and Molecular Life Sciences. 65:1378–1389.

Wu XS et al. 2002. Identification of an organelle receptor for myosin Va. Nat Cell Biol. 4:271–278.

Zhang J et al. 2007. Thirty-one flavors of Drosophila Rab proteins. Genetics. 176:1307–1322.