Duplicated Myosin V Genes in Teleosts Show Evolutionary Rate Variations among the Motor and Cargo-Binding Domains

Richard J. Nuckels[1,2,*], Chris C. Nice[1], and Dana M. García[1]

[1]Department of Biology, Texas State University, San Marcos

[2]Department of Biology, The University of Texas at San Antonio

*Corresponding author: E-mail: richard.nuckels@utsa.edu.

Accepted: November 27, 2018

Abstract

We 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 we refer to now as *myo5bb*. Analysis using d*N*/d*S* 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, we 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. The high number of invariant codons coupled with relatively high d*N*/d*S* values in the region of the *myo5* genes encoding the ATP-binding domain suggests the encoded proteins retain function and may have acquired novel functions associated with changes to the cargo-binding domain.

Key words: gene duplication, myosin V, cargo-binding domain, motor domain, invariant codon, d*N*/d*S*, evolutionary rates.

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 the 2R hypothesis such that it is currently widely accepted. An additional genome duplication event is thought to have occurred in the teleost lineage around 300 Ma (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, Brunet, et al. 2009). Although the total number of genes in fish is not much different from tetrapods, Braasch, Brunet, et al. (2009) found that there are approximately 30% more pigmentation-related genes comparted with tetrapods. Duplicated genes related to pigmentation have provided new opportunities for phenotypic diversity among fishes (Braasch, Liedtke, et al. 2009) in addition to opening the evolutionary door for neofunctionalization for one of the duplicated genes to acquire a nonpigmentation-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 nonpigment-related cell types including neurons.

Together, the expression patterns of these two duplicated genes approximate the expression pattern of the nonduplicated *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 2008; 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 (CBD). 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 2008; 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; Swiatecka-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 (Swiatecka-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 nonsynonymous 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, 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, we characterize the *myo5* duplicates and their evolutionary history in vertebrates. We identify a branch in a phylogeny of the myosin gene family for a duplicated gene (*myo5bb*) in teleosts and spotted gar. We 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 CBD is encoded. In addition to presenting data that supports previously described genome duplication events, namely the vertebrate R1/R2 and fish-specific genome duplications, we identify a tandem gene duplication event for the *myo5a* and *myo5c* genes, and we propose a model for the evolution of the *myo5* gene family. In our proposed evolutionary model of the *myo5* gene family, we 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 our analysis of codons, we 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

We collected *myosin 5* sequences using Ensembl's genomic database (Ensembl Release 86), NCBI, and the Japanese Lamprey Genome Project. 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,

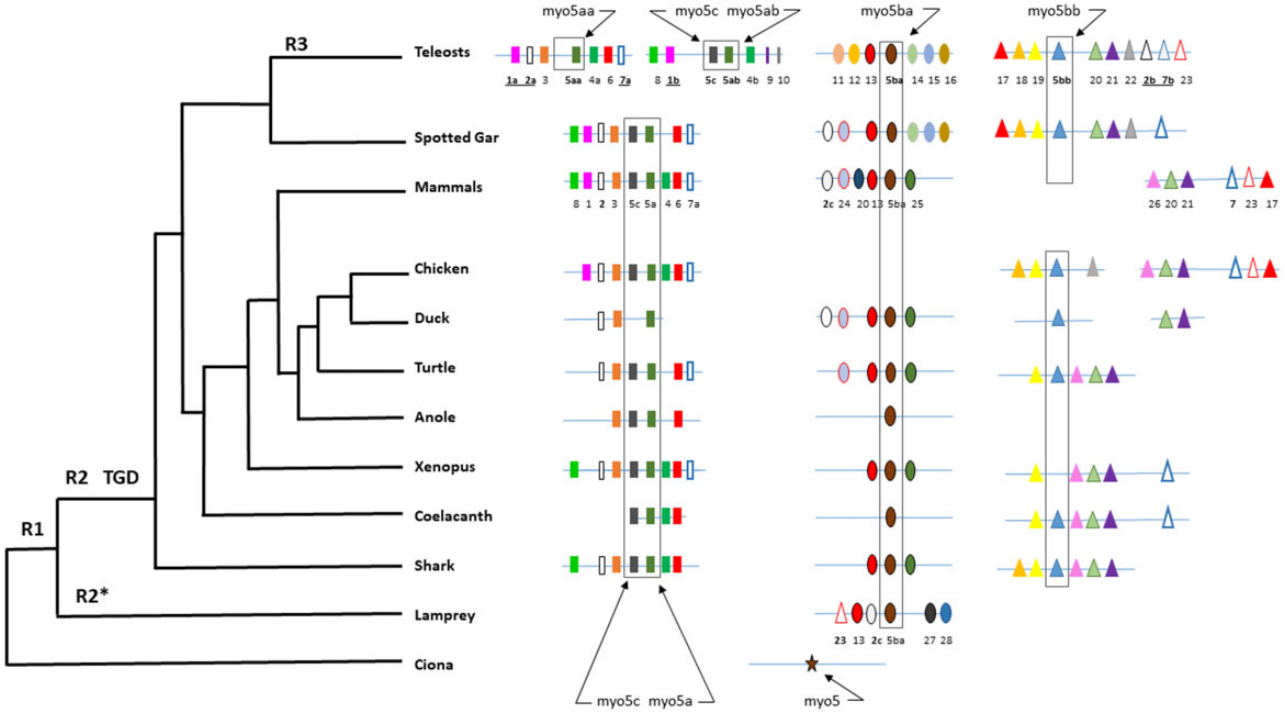


Fig. 1.—Cladogram and syntenic diagram supporting key hypothesized evolutionary events in the history of *myo5* genes in teleosts and other chordates. Putative vertebrate genome duplication events (R1 and R2) led to the creation of three *myo5* copies, *myo5a* (green rectangle), *myo5ba (*brown oval), 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 has 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 corresponding 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. We found other *mapk* genes near *myo5ba* (unshaded, black bordered oval) and near *myo5bb* (unshaded, black triangle) gene families. We 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 gray 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*. The location of *myo5c* directly downstream from *myo5ab* was observed in all teleosts examined. Likewise, the location for *myo5c* is directly downstream of *myo5a* in nonteleost vertebrates. Similar syntenic observations were made for other teleosts and nonteleosts, 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*, and 28-*pigo*. See supplementary table 1, Supplementary Material online for *myo5* gene identifiers and chromosomal locations.

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, we 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, we used the same set of genes as in the *myo5ab* gene family because *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 we 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 we used approximately 30 genes total and about 10 genes in each *myo5* gene neighborhood. Each gene neighborhood generally contained genes within 200,000 bases of each *myo5* gene.

Ancestral Chromosome Mapping

We used ancestral chromosomal reconstructions from Nakatani et al. (2007) and Bian et al. (2016) 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 nonteleost fish (including spotted gar) split. Utilizing these two sets of chromosomal mapping data, we were able to identify whether our 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 (fig. 2).

Alignment and Phylogenetics

Eighty-seven sequences were aligned using ClustalW and Geneious Pro 6.0 (Biomatters Ltd). Sequences were virtually translated, verified to contain open reading frames, and then back translated. The ends of the aligned sequences were trimmed and smaller alignments from three regions (motor domain, neck, CBD) 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 MrBayes 3.1 and a GTR+I+G model of evolution, trees were generated for the full-length coding sequence (6,870 bp) of *myo5*, the motor domain, the neck, and the CBD. The parameters used in the MrBayes-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.

For the four alignments we generated, we 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 CBD, so it was only included in those alignments and phylogenetic analyses and not in the neck or full sequence alignments. Similarly, there were other sequences that were missing sequence data for more than 50% of the alignment. These sequences were not included in those specific alignments (fig. 4).

d*N*/d*S* Rates and Identification of Invariant Codons

We determined the evolutionary rate (d*N*/d*S*) using MEGA6. "d*N*" is defined as the ratio of nonsynonymous substitutions per nonsynonymous site; "d*S*" 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. We used MEGA6 to determine the d*N* and d*S* values for each codon in our alignment for a specific clade which generally consisted of 8–10 teleost sequences for a specific *myo5* duplicate. Summing the d*N* and d*S* values for all the codons in our alignment and then dividing d*N* by d*S* allowed us to determine the d*N*/d*S* ratio for each alignment. To quantify the percentage of codons that are invariant and experiencing extreme purifying selection, we counted the number of codons in each of the original four alignments (whole gene, motor domain, neck, and CBD) that had d*N* and d*S* 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 1 and 2).

Selection Tests

We used the Datamonkey server and the HyPhy software package (Kosakovsky Pond et al. 2005; Delport et al. 2010) to test for purifying selection, positive selection, and episodic selection at the codon level and the branch level among the phylogenies we 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. We 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. We then tested our 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

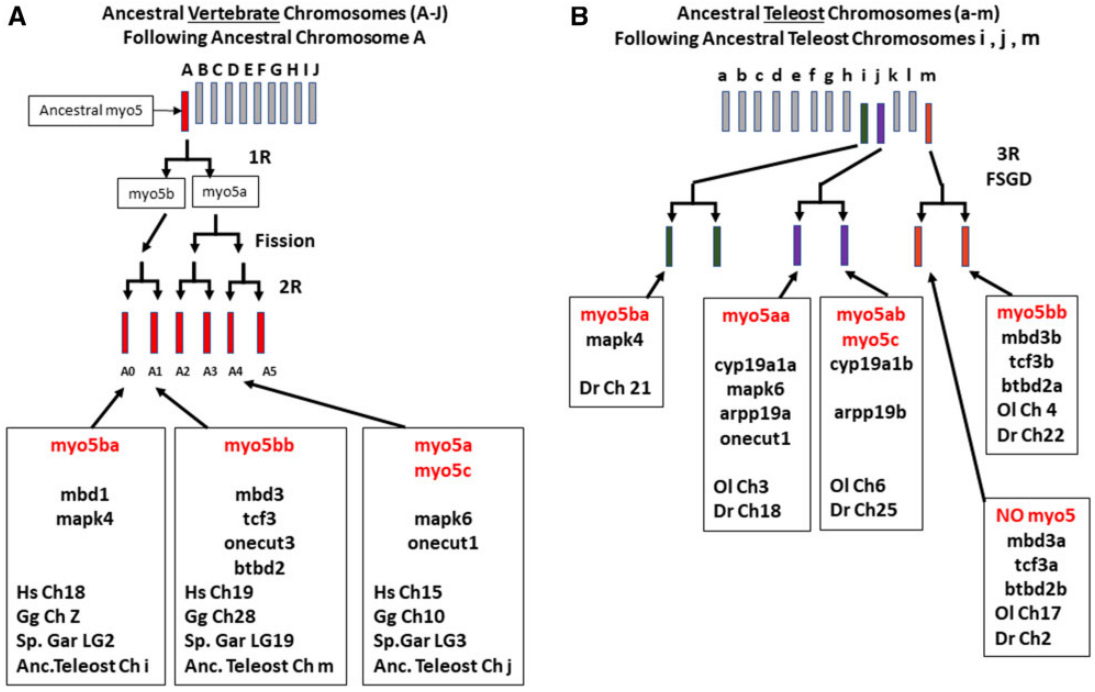


Fig. 2.—Ten ancestral vertebrate proto-chromosomes have been previously described along with thirteen ancestral teleost chromosomes (Nakatani et al. 2007; Bian et al. 2016). All *myo5* genes were traced back to ancestral vertebrate chromosome A (Panel *A*). After two whole rounds of genome duplication and a fission event, six chromosomal fragments (A0–A5) existed. *myo5* genes and select coduplicated genes are shown in the boxed region along with what ancestral chromosome fragment these genes are derived from. The chromosomal location of the genes in various organisms is listed at the bottom of the boxed regions. "Hs" for *H. sapiens*, "Gg" for *G. gallus*, "Sp. Gar" for spotted gar, "Anc. Teleost" for Ancestral teleost, "Ch" for Chromosome, "LG" for 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" for *D. rerio*, "Ol" for *O. latipes*).

Table 1
Teleost *myo5* d*N*/d*S* Values

	Whole Gene	5′ End and ATP Binding		Neck and Actin-Binding		3′ End and Cargo-Binding Domain	
	1,915 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	0.13
myo5ab	0.36	0.12	0.02	0.41	0.14	0.35	0.23
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

Note.—Regions that play a role in functionality (motor domain and CBD) have very low d*N*/d*S* values. d*N*/d*S* values are reflective of codon changes that lead to synonymous (S) or nonsynonymous (N) codons. d*N*/d*S* values are presented for each clade for whole duplicated genes composed of 1,915 codons. d*N*/d*S* values are also presented for smaller regions for each clade which contain the ATP-binding domain, the actin-binding domain and neck region, and the CBD. 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 d*N*/d*S* 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 CBD (10 codons), we 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 10 codons (****) in the CBD as it is unknown what amino acids are involved in this process for the orthologous *myo5c* in human.

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 we used in our 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). We used 8–10 teleost sequences from our alignments to test for selection among the duplicated *myo5* genes using the MEME and REL

Table 2
Percentage of Invariant Codons in Teleost *myo5* Genes

	Total Codons	Invariants/Extreme Purifying Selection % of 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	1,908	16.7
myo5ab full length	1,938	8.0
myo5ba full length	1,904	13.6
myo5bb full length	1,668	8.1
myo5c full length	1,761	11.8

NOTE.—The percentage of invariant codons for each *myo5* clade in teleosts is surprisingly high. The number of codons used in each alignment is 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 with 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 CBD is located. Extreme purifying selection is defined here as $dN = dS = 0$. No substitutions were identified in any of the 3 codon positions for these sites. For the CBD (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.

selection tests. We did this for each teleost duplicated *myo5* gene clade and for the smaller regions within the gene. For example, we used the 5′ end motor domain alignment of nine teleost sequences for the *myo5aa* teleost gene clade and ran the MEME and REL selection tests. Similarly, we tested the neck and CBD for the *myo5aa* teleost clade, and we ran these same selection tests using the comparable domains for the teleost clades which included *myo5ab*, *myo5ba*, and *myo5bb* genes (table 4).

Results

Syntenic Analysis

To determine whether *myo5* duplicates arose through duplication of individual genes, chromosomes or their segments, or entire genomes, we performed syntenic analysis. We found the chromosomal locations for *myo5aa* and *myo5ab* in zebrafish on chromosomes 18 and 25, respectively, and the locus for *myo5c* 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 nonteleost vertebrates (fig. 1). Initial phylogenetic analyses revealed a new *myo5* clade (the *myo5bb* clade), and syntenic analyses provided further support of the presence of this gene along with neighboring genes in teleosts, spotted gar, chicken, duck, turkey, turtle, coelacanth, and shark. This gene 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.

We traced the origin of extant *myo*5 sequences to ancestral vertebrate and teleost chromosomes to further test the findings from our syntenic analysis (fig. 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 (fig. 2*A*). The *myo5ba* genes are linked with the A0 fragment and the *myo5bb* genes are linked with the A1 fragment (fig. 2*A*). Coduplicated genes exist, for example, *mbd1* near *myo5ba* and *mbd3* near *myo5bb*. Additional coduplicated 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 (fig. 2*B*).

We 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 figure 4*A* and B shows 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, we found the *pigo* and *ensab* genes on one side of the *myo5* genes, and we found *mapk4*, *cfap53*, and *atp8b* on the other side of the *myo5* genes.

Phylogenetic and d*N*/d*S* Analyses

To understand the molecular evolution of the *myo5* gene family, phylogenetic analysis was performed using 87 genes from 24 different species (see supplementary table S1, Supplementary Material online for names and genomic database identifiers). Using ClustalW, we produced a final alignment of 6,468 base pairs per gene. Four phylogenetic trees were generated, representing the full-length coding sequence

(fig. 3*A*), the portion encoding the CBD at the 3′ end of the *myo5* gene (fig. 3*B*), the 5′ end of the gene which encodes the motor domain with its highly conserved ATP-binding domain (fig. 3*C*), and the more variable portion of the *myo5* gene which encodes the neck and tail regions (fig. 3*D*). Figure 4*A* shows where the smaller alignments fit within our full-length alignment. The *myo5aa* teleost sequences form a monophyletic clade, and the *myo5ab* teleost sequences form a monophyletic clade (fig. 3*A*, *B,* and D). Separate clades form for the *myo5ba* teleost sequences, the *myo5bb* sequences, and the *myo5c* sequences (fig. 3*A*, *B,* and D). Nonteleost *myo5a* sequences formed a clade sister to a clade which included spotted gar and teleost *myo5aa* and *myo5ab* sequences (fig. 3*A* and D). Nonteleost and teleost *myo5c* sequences formed a monophyletic clade (fig. 3*A*–D); however, tetrapod *myo5b* was for the most part monophyletic with teleost *myo5ba*, but not with teleost *myo5bb* (fig. 3*A*, *B*, and D). 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 fig. 4).

We determined d*N*/d*S* values for each clade and each region of the *myo5* gene family (see fig. 5 and table 1). The d*N*/d*S* ratios for the *myo5ba* and *myo5bb* were higher than the d*N*/d*S* ratios for *myo5aa* and *myo5ab* (fig. 5). For the *myo5a* duplicates (*myo5aa* and *myo5ab*), the percentage increase is higher for the d*N*/d*S* values for the motor domain and the CBD with the largest amount of d*N*/d*S* change taking place in the CBD for the *myo5ab* clade. For the *myo5bb* clade we see a much smaller increase in the d*N*/d*S* rates for the CBD with a 68% increase compared with the 250% increase seen in the *myo5ab* clade. The d*N*/d*S* for the motor domain also increased a relatively small amount (33%) for the *myo5bb* clade compared with the motor domain of the *myo5ab* clade (140%).

In addition to calculating the d*N*/d*S* ratios for each of the whole genes and for specific regions within the *myo5* genes, codon-specific values for d*N* and d*S* for each alignment tested were calculated. For a given region of a *myo5* gene, for example the 5′ end, d*N* and d*S* were calculated by comparing sequences from at least 8 teleost species. The 3′ end where the CBD is encoded evinced far fewer invariant sites in the *myo5ab* clades compared with the *myo5aa* clades. We identified 30.1% of the codons for the teleost *myo5aa* clade 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 8–9 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 2). 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, we compared the d*N*/d*S* 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 (supplementary tables S2 and S3, Supplementary Material online), suggesting these duplicates retain the motor functions related to ATP- and actin-binding. However, there is a significant difference in the CBD when looking at the codon sites linked with functionally important amino acids for MYO5 proteins (supplementary tables S4 and S5, Supplementary Material online). We examined the 10 sites that are linked with RAB11a binding to MYO5b in mammals (Pylypenko 2013), and we found that the d*N*/d*S* values for these ten sites are 0 or mathematically undefined, meaning the value of the denominator (d*S*) equaled 0, highlighting the high conservation for these sites in *myo5ba* in teleosts (supplementary table S5, Supplementary Material online). 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, we specifically selected 21 codons that code for amino acids linked with the functional myosin motor activity for further analysis. The d*N*/d*S* rate for these 21 codons was lower compared with the d*N*/d*S* rate of the entire 5′ region. The average d*N*/d*S* for codons in the 5′ regions for teleost *myo5* genes was 0.08 (table 1), but the average d*N*/d*S* 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 1).

In the myosin head, the aspartate at position 134, D134, is an example of an amino acid that was conserved in all *myo5* sequences analyzed, with the following exceptions: The inferred amino acid sequence from the single *myo5* gene in *Ciona* manifests a D→E change. In *Tetraodon*, there is 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. The cavefish *myo5aa* gene has a premature stop codon which truncates the protein before the CBD is translated, so cavefish may not have a functional *myo5aa* gene.

Another feature of the myosin head's ATP-binding domain is the p-loop, a region of the protein that interacts with the terminal phosphate on ATP (Coureux et al. 2003). Among the

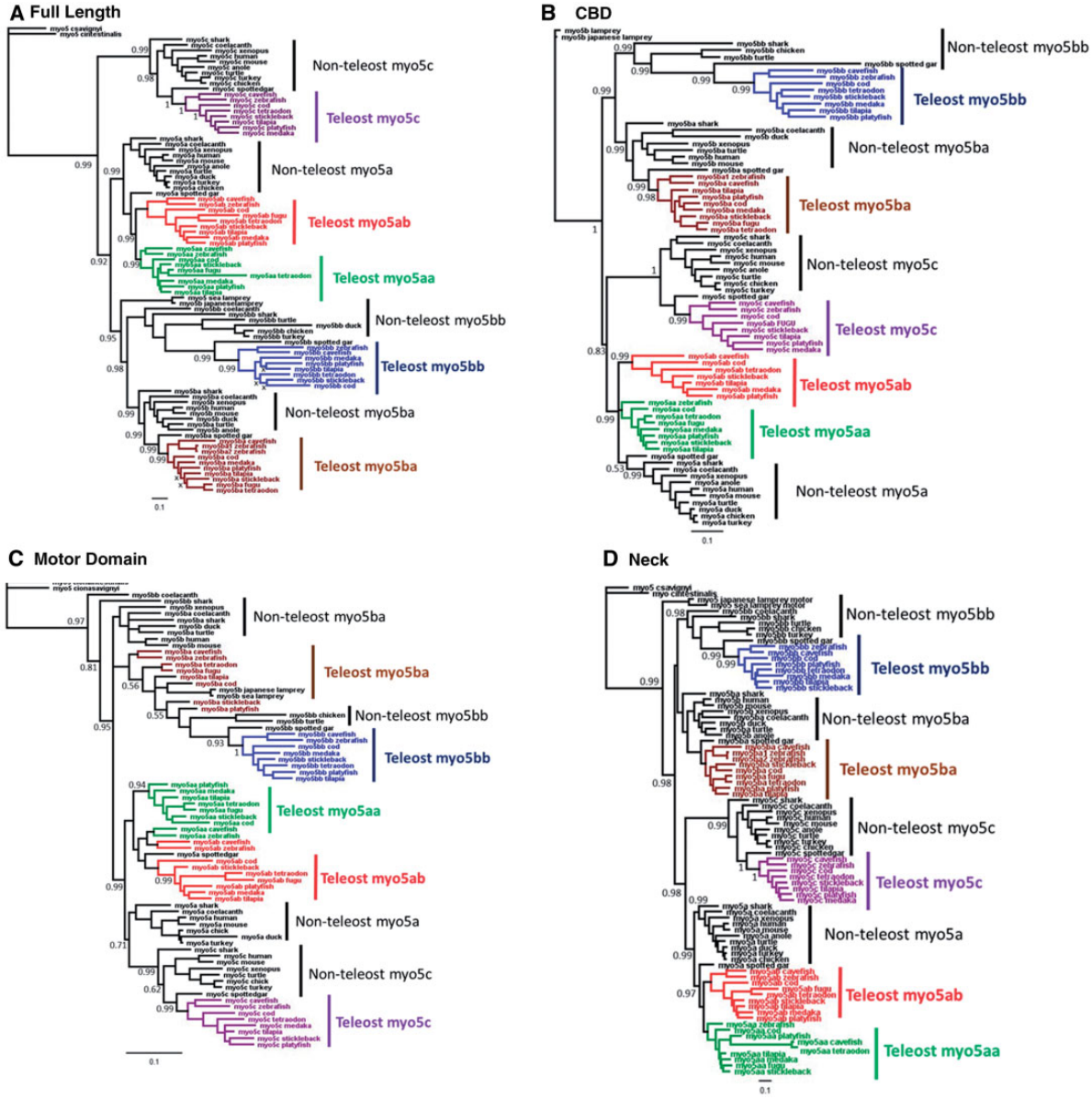
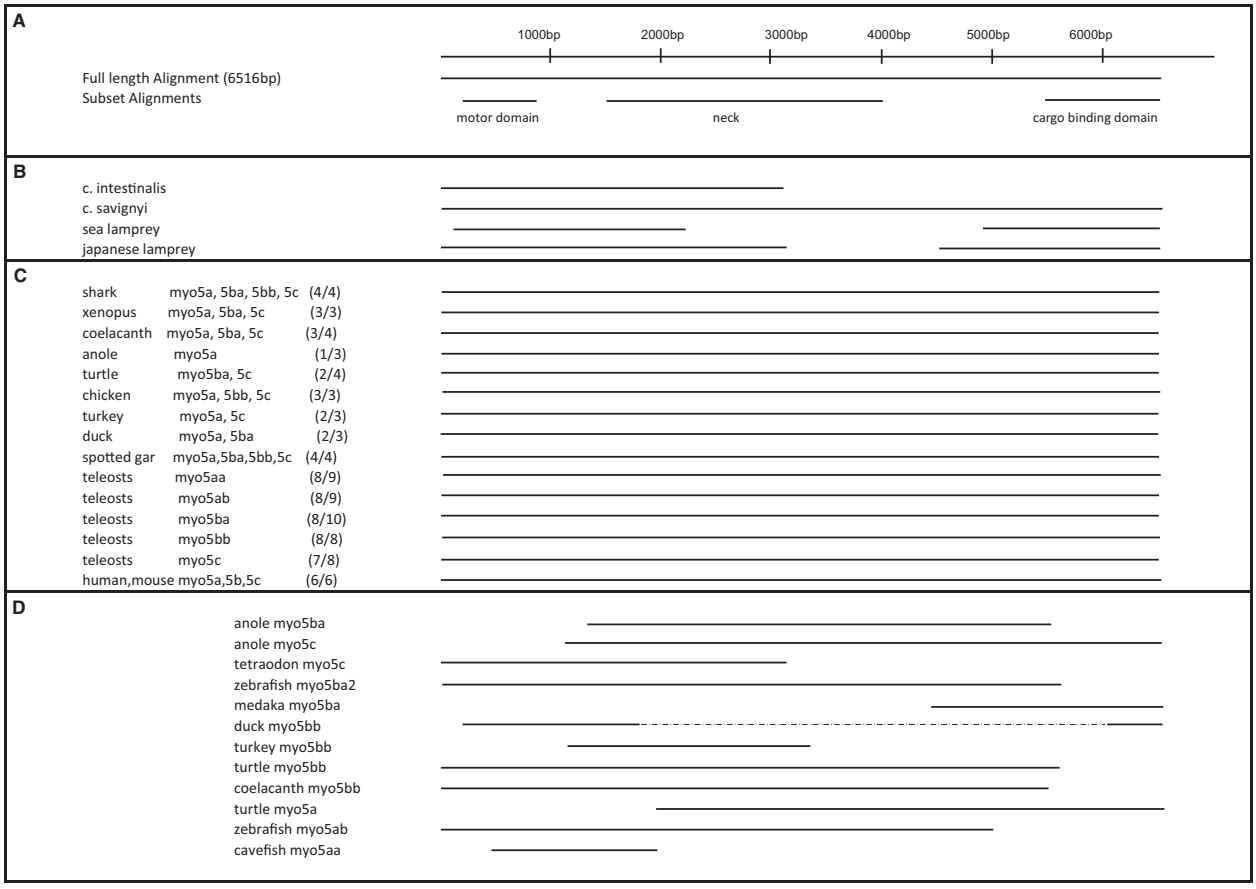


Fig. 3.—(*A*) Phylogenetic tree for full-length coding sequence (6,468 bp) of *myo5* using MrBayes 3.1 and a GTR+I+G model of evolution. Teleost *myo5bb* clade shown in blue with an extended branch leading up to the clade. Posterior probability values are provided for some nodes. If not shown, the posterior 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. (*B*) Phylogenetic tree for the CBD (1,035 bp fragment) of *myo5* using MrBayes 3.1 and a GTR+I+G model of evolution. This 1,035 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. (*C*) Phylogenetic tree for the motor domain (a 651 bp fragment) of *myo5* using MrBayes 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. (*D*) Phylogenetic tree for the neck and coiled coil domain (a 2,505 bp fragment) of *myo5* using MrBayes 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 with 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.



* sequences not found in the available genomic databases: (myo5ba-chicken, turkey); (myo5bb-xenopus, anole, mammals, fugu); (myo5c-duck, fugu)

Number in parentheses represent the number of full length myo5 sequences for that species or group of species out of total number of sequences used in this study for that group. For groups in panel C that are incomplete, the missing sequence can be found in panel D along with a representation of what part of the gene sequence is available in the genomic databases. For example, there are 9 teleost myo5aa sequences but only 8 of them have full length sequence data. The other myo5aa sequence for teleosts that is not full length is the cavefish myo5aa gene which has a small fragment of a gene depicted in panel 4 and there is a stop codon present in that gene.

Fig. 4.—Alignment of *myo5* sequences used in this study. Panel *A* shows the full-length alignment size using 87 species along with the three smaller subsets (motor domain, neck, and CBD) that were used for further characterization of the *myo5* gene family. Panel *B* shows the smaller sequences found among lamprey and *Ciona 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 our total number of 87 sequences are truncated or missing some part of the full-length sequence.

highly conserved amino acid residues in the p-loop, which comprises amino acids 163–170 (GESG**A**GKT), the only variation that we see in this region is for the whole *myo5bb* teleost clade, in which the alanine at position 167 has been substituted with a serine, yielding the consensus sequence GESG**S**GKT.

The 742 codons in the neck region show the largest amount of molecular variation with d*N*/d*S* rates ranging from 0.23 for *myo5aa* to 0.41 for *myo5ab* (table 1). When comparing the duplicates for this region, *myo5ab* has a larger d*N*/d*S* value (0.41) 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 with the neck domain except for *myo5aa* (table 1 and supplementary table S3, Supplementary Material online). For *myo5aa*, the d*N*/d*S* value for the 23 codons associated with actin binding is 0.23 but for the other four teleost genes the d*N*/d*S* range is 0.04–0.14.

For the 319 codons in the 3′ end of the myo5 genes, which include the CBD, the *myo5ab* and *myo5bb* genes have the highest d*N*/d*S* values at 0.35 and 0.32, respectively. The *myo5aa* and *myo5ba* genes are much more conserved in this region with d*N*/d*S* rates of 0.10 and 0.19, respectively. When looking at the 10 codons linked with cargo binding,

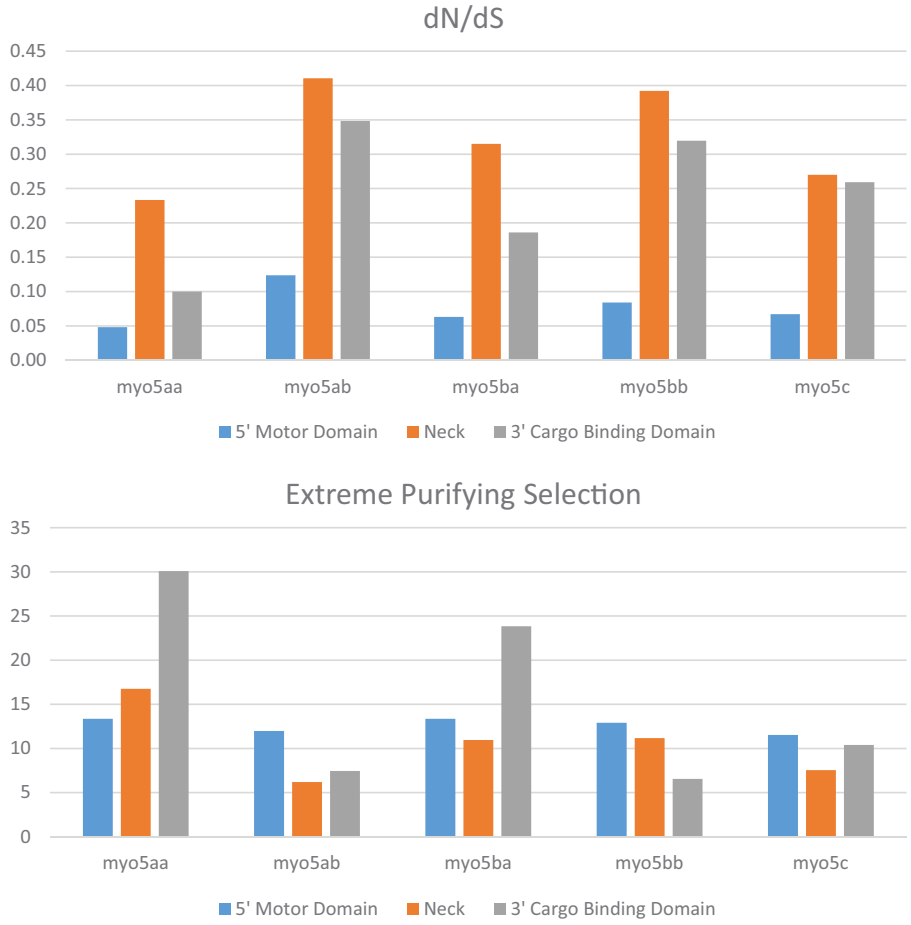


Fig. 5.—d*N*/d*S* rates and percentage of codons that are invariant or under extreme purifying selection for all 5 *myo5* genes in teleosts. We see smaller d*N*/d*S* rates for *myo5aa* compared with *myo5ab* and for *myo5ba* compared with *myo5bb* for all cases using smaller regions of the *myo5* genes. 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* CBD has a much higher percentage of invariant codons compared with the paralogous *myo5bb* CBD, 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.

myo5ba has a d*N*/d*S* rate of 0 and *myo5aa* has a d*N*/d*S* rate of 0.13 (table 1 and supplementary tables S4 and S5, Supplementary Material online).

Selection Test Results

We carried out several selection tests (table 3) accessed from the Datamonkey server and utilizing the HyPhy software package. We used BUSTED to test for selection across our phylogeny and this test revealed that episodic diversifying selection was occurring somewhere in our full-length phylogeny ($P < 0.05$). We specifically selected *myo5b* branches to test as foreground branches, and the remaining branches were considered background branches. Three rate classes (ω_1, ω_2, ω_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 we 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 CBD. In the CBD, we see more episodic diversifying selection in the *myo5bb* clade of teleosts versus the *myo5ba* clade of teleosts. We 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

Table 3
Results from MEME (Mixed Effects Model of Evolution) and REL (Random Effects Likelihood)

			MEME (No. of Sites)	REL (# of Sites)	
	No. of Sequences	Total Codons	$P<0.05$	REL +	REL −
Motor-myo5a	25	217	3	0	180
Motor-myo5aa	9	217	2	0	217
Motor-myo5ab	9	217	0	2	183
Motor-myo5b	22	217	3	0	217
Motor-myo5ba	8	217	1	0	217
Motor-myo5bb	8	217	1	0	217
Motor-myo5c	8	217	1	0	217
Neck-myo5a	25	830	16	0	830
Neck-myo5aa	9	830	7	4	226
Neck-myo5ab	9	830	6	0	830
Neck-myo5b	25	830	26	1	408
Neck-myo5ba	9	831	20	0	242
Neck-myo5bb	8	830	17	0	377
Neck-myo5c	8	830	5	0	96
CBD-myo5a	24	343	3	0	343
CBD-myo5aa	8	343	1	2	132
CBD-myo5ab	8	343	1	1	103
CBD-myo5b	23	343	5	2	180
CBD-myo5ba	9	343	0	0	78
CBD-myo5bb	8	343	5	5	247
CBD-myo5c	14	343	1	0	69

NOTE.—Summary of results from MEME (Mixed Effects Model of Evolution) and REL (Random Effects Likelihood) 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 we see a more episodic diversifying selection (MEME) in the *myo5bb* clade of teleosts versus the *myo5ba* clade of teleosts. We 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. Cells reporting results from 8 to 9 sequences are based solely on teleost sequences. The *myo5c* CBD clade consists of 8 teleost sequences and 6 nonteleost sequences. The clades with 22–25 sequences contain all the teleost sequences in that group (16–18 sequences) plus 6–8 nonteleost sequences.

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 CBD 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 CBD 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 CBD and 103 sites subject to purifying selection (table 3).

A BS-REL (Branch Site-Random Effects Likelihood) test was used on the phylogenies we generated to test for episodic or diversifying selection along branches. We identified episodic selection taking place along the *myo5bb* branch leading up to the ray-finned fish lineage (table 4). On this branch, 20% of the sites in the CBD 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 CBD 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 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 ω_1 and ω_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 ω_1 (dN/dS) = 0.316 for 76% of the sites and $\omega_2 = 80.1$ for 24% of the sites.

Table 4
Branch Site-Random Effects Likelihood (BS-REL) Test Results

	Branch	*P* Value	Positive	Neutral	Purifying
Motor domain	myo5bb clade	0.022	0.08	0.34	0.58
CBD	myo5b clade	0.014	0.13	0.03	0.84
	myo5ba clade	0.014	0.04	0.03	0.93
	myo5bb clade	0.040	0.20	0.26	0.54

NOTE.—Using the BS-REL test through the Datamonkey server, the CBD and motor domain showed evidence of episodic diversifying selection. Twenty percent of the sites along the *myo5bb* CBD 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. The results of the BS-REL test for the myo5bb clade are highlighted.

Discussion

We 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) looked at Myo7a in fish and mentioned that the antibody used is likely recognizing both Myo7a1 and Myo7a2. Here, we highlight 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 nonteleost 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 (fig. 1). One of these *myo5* duplications may be the result of a tandem gene duplication event (TGD) which preceded the divergence of jawed vertebrates; the resulting paralogs are currently referred to as *myo5a* and *myo5c*. The fourth duplication event we identified is specific to teleosts and is likely the result of the teleost- or fish-specific genome duplication event (R3) and 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), we suspect a gene loss took place after the R2 duplication event. Our examination of the syntenic data and our ancestral chromosome mapping support these predictions 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. In the Ensembl genomic database, several of these genes are identified as *myo5b* for nonteleosts 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. Our 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 (see Kuraku 2010, 2013; Qiu et al. 2011 for more details about hidden paralogy).

For each lamprey species, we found a gene that aligns with the 5′ end of our alignments and a second gene that aligns with the 3′ end of our alignments. However, we 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. We identified the *myo5* motor domain on an independent small scaffold without any genes around it. We 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. In addition, the sizes of the exons and introns for Japanese lamprey sequences are comparable with 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 our phylogenetic analysis should have placed the CBD for lamprey in a different *myo5* clade than the lamprey *myo5* motor domains. Interestingly, we see a couple of other examples in our study where there are truncated *myo5* genes. One of the cavefish genes, *myo5aa*, seems to have only a short

sequence covering the motor domain. We identified a short tetraodon *myo5c* sequence that contains the first 3,000 bp in the 5′ region of the *myo5* gene and is missing the CBD. We also identify the *C. intestinalis* sequence to be a short sequence of 2,982 bp, missing the CBD.

In addition to providing a model for the evolutionary history of the *myo5* genes, our syntenic analysis (along with the phylogenetic analysis) has clarified the orthology of the *myo5bb* genes and has helped validate the nomenclature of the duplicated teleost genes because in some cases (namely the *myo5bb* genes) a *myo5* name had not been assigned to all of the *myo5* genes in Ensembl or other genomic data depositories (supplementary table S1, Supplementary Material online). Although we have chosen to retain the gene names used in previous studies, we recognize the potential confusion due to unevenness in the nomenclature, for example *myo5ba* and *myo5bb* representing paralogs that arose prior to the divergence of all jawed vertebrates whereas *myo5aa* and *myo5ab* represent duplicates consequent to the fish-specific genome duplication. In the latter case, *myo5aaa* and *myo5aab* would be names that better reflect the evolutionary history and relationships among the genes, but perhaps introduce unwarranted cumbersomeness.

With respect to the orthology of the *myo5bb* genes, we introduce the idea of the *myo5b* gene family having been duplicated as a result of the R1 genome duplication, an inference supported by the observation that the *myo5bb* gene family is found not only in teleosts but also in spotted gar, coelacanth, shark, turtle, chicken, and turkey.

There are a couple of instances where a duplicated clade might be expected but it appears that those duplicated regions have gone missing over evolutionary time. For example, in figure 2*A* on fragment A5, we would expect there to be a duplicated region structurally similar to what we found on fragment A4. If this were the case, then we would have a *myo5ab/myo5cb* on fragment A5, and the duplicated genes on fragment A4 would be named *myo5aa/myo5ca*.

Phylogenetic Analyses

Our phylogenetic trees presented in figure 3*A* and D show similar topologies that are consistent with the results presented in figures 1 and 2. For the trees in figure 3*A* and D, teleost *myo5aa* genes group in one clade and *myo5ab* genes group in a sister clade consistent with the teleost genome duplication. The divergence of spotted gar *myo5a* prior to the genesis of the two teleost *myo5a* clades is expected and compatible with the current understanding of evolutionary relationships among these taxa. These two phylogenetic trees also support the divergence of a clade including a lobe finned fish along with tetrapods, which is what one would expect. Surprisingly, shark *myo5a* also segregates to this branch. Based on current understandings of the phylogenetic relationships among sharks and other chordates, one would have expected shark *myo5a* to segregate in a branch basal to the teleost and tetrapod groups.

Figure 3*D* indicates the divergence of the *myo5a* gene family and the *myo5c* gene family from a (relatively recent) common ancestor. We believe this tree best represents the actual evolutionary history since the sequences for the neck regions have the most genetic diversity, and therefore analyses of these domains has the greatest power to resolve evolutionary events over geologic time frames. The other trees were based on alignments of the highly conserved motor domain (fig. 3*B*) and the CBD (fig. 3*C*), which led to poorer resolution for the branches in those phylogenies. Figure 3*A* showed a vastly different topology regarding the divergence of the *myo5c* gene family. If the tree in figure 3*A* was the true tree, this would suggest that the *myo5a/myo5c* duplication took place before R1; whereas, if figure 3*D* is the true tree, then it suggests that the *myo5a/myo5c* duplication took place after R2 and before R3.

In figure 3*D*, we also present support for the newly named *myo5ba* clade which includes an expected divergence pattern of shark, followed by bony fish which break up into two sister clades of lobe finned fish and ray finned fish with spotted gar diverging before other teleosts in the ray finned fish lineage. For the newly identified *myo5bb* clade we first see the divergence of coelacanth and shark *myo5bb* genes. These genes do not show the expected pattern of divergence because coelacanth *myo5bb* diverges before shark *myo5bb*. However, after that, we see the divergence of tetrapod *myo5bb* genes and ray-finned fish including spotted gar and teleost *myo5bb* genes.

d*N*/d*S* Analyses

As branch lengths represent the number of substitutions per site, we thought the long branches evident in the *myo5bb* lineage might reflect a large amount of substitutions resulting in amino acid changes (fig. 3*A*–D). Were that the case, our examination of the amino acid sequences encoded by the *myo5bb* genes would be expected to reveal an increase in the d*N*/d*S* 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, we 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 2).

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 d*N*/d*S* values than other codons in the *myo5* genes, indicating these codons were among the most conserved codons in all 9 teleost species and among the duplicated genes (supplementary tables S2–S5, Supplementary Material online). For many of the sites in the *myo5* genes (supplementary tables S2–S5, Supplementary Material online), d*N*/d*S* 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 CBD (fig. 5). The largest percentage difference of invariant codons between two paralogous clades exists between the CBD of the *myo5aa* (30.1%) and *myo5ab* (7.5%) clades and between the CBDs of the *myo5ba* (23.8%) and *myo5bb* (6.6%) clades (table 2). The CBD has previously been characterized as playing a role in lightly or nonpigmented (*dilute*) phenotypes (Nascimento et al.1997), and in this domain sequence conservation mostly persists. The serine residue at position 1,650 has been shown to be a site for phosphorylation by which disassociation of melanosomes from the myosin motor is regulated (Karcher et al. 2001; Pylypenko et al. 2013). It is present in all of Myo5 sequences analyzed with the exception of Myo5ab from *T. nigroviridis*. The basic residues K1706 and K1779 were identified in Li and Nebenführ (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 with the cavefish *myo5aa* gene. The cave dwelling nonpigmented cavefish have a premature stop codon in the *myo5aa* gene (fig. 3*D*), precluding translation of the CBD, and likely preventing the Myo5aa protein from transporting its melanin cargo. Because our sequence data are 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.

Slightly C-terminal of 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*'s nonfunctional 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 with other regions of the *myo5* genes. In comparing the CBD, 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, we 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* CBD 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 CBD 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, we 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, our data imply that these duplications have persisted over evolutionary time with a high degree of conservation at specific sites and suggest that selection continues to operate on the protein products of these genes. This finding raises the question as to why sites are so highly conserved over hundreds of millions of years if these duplicated genes are nonfunctional. Although the possibility exists that one of these duplicated genes has obtained a neofunctional role over evolutionary time, an alternative explanation for our data supports a model of relaxed selective pressure likely due to the redundancy of the duplicated genes.

Using d*N*/d*S* evolutionary rate comparisons, selection tests, and the identification of a high percentage of codons subject to extreme purifying selection, we present data linking the newly identified *myo5bb* clade with a high degree of conservation at functionally important amino acids, suggesting *myo5bb* is a duplicate that has retained function. The relaxed selective pressure on this *myo5bb* family of genes could lead to alternative expression patterns developmentally or within the organisms and possibly leading to a new function for the Myo5bb proteins. The high degree of conservation of specific sites linked with functionality supports an evolutionary pathway leading to relaxed selective pressure at a minimum and

possibly neofunctionalization for the *myo5ab* duplicated genes (found in teleosts only) and *myo5bb* duplicated genes (found in birds, turtle, shark, coelacanth, spotted gar, and teleosts).

We 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 with nonteleosts (Braasch, Brunet, et al. 2009). 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, we suspect that *myo5aa* is carrying out the melanosome shuttling role similar to *myo5a* in nonteleosts, the role of *myo5ab* remains to be determined. In addition, we suspect that *myo5ba* in fish are carrying out the same role as *myo5b* (more accurately, *myo5ba*) in nonteleosts 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, we suspect that the proteins encoded by these genes still have a functional role and we suspect that the new role is related to the variability and positive selection we have identified in the CBD.

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 nonfish 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. We 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) We identified invariant codons to exist throughout our alignments of duplicated genes and at times these invariant codons were clustered in groups of as many as 10 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.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported through funding by the National Science Foundation-Project Flowing Waters awarded to Dr Julie Westerlund, Dr Weston Nowlin, and Dr Tim Bonner (NSF Grant # 742306) and the Department of Biology, Texas State University. We acknowledge two anonymous referees for insightful and constructive criticism.

Literature Cited

Anisimova M, Yang Z. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol. 24(5):1219–1228.

Bian C, Hu Y, Ravi V, Kuznetsova IS, Shen X, Mu X, Sun Y, You X, Li J, Li X, Qiu Y, Tay BH, Thevasagayam NM, Komissarov AS, Trifonov V, Kabilov M, Tupikin A, Luo J, Liu Y, Song H, Liu C, Wang X, Gu D, Yang Y, Li W, Polgar G, Fan G, Zeng P, Zhang H, Xiong Z, Tang Z, Peng C, Ruan Z, Yu H, Chen J, Fan M, Huang Y, Wang M, Zhao X, Hu G, Yang H, Wang J, Wang J, Xu X, Song L, Xu G, Xu P, Xu J, O'Brien SJ, OrbûÀn L, Venkatesh B, Shi Q. 2016. The Asian arowana (*Scleropages formosus*) genome provides new insights into the evolution of an early lineage of teleosts. Sci Rep. 6:1–17.

Bielawski JP, Yang Z. 2003. Maximum likelihood methods for detecting adaptive evolution after gene duplication. J Struct Funct Genomics. 3(1/4):201–212.

Bielawski JP, Yang Z. 2004. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. J Mol Evol. 59(1):121–132.

Braasch I, Brunet F, Volff J-N, Schartl M. 2009. Pigmentation pathway evolution after whole-genome duplication in fish. Genome Biol Evol. 1:479–493.

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

Braasch I, Schartl M, Volff J-N. 2007. Evolution of pigment synthesis pathways by gene and genome duplication in fish. BMC Evol Biol. 7(1):74.

Brennecke J, Stark A, Russell RB, Cohen SM. 2005. Principles of microRNA-target recognition. PLoS Biol. 3(3):e85.

Catchen JM, Conery JS, Postlethwait JH. 2009. Automated identification of conserved synteny after whole-genome duplication. Genome Res. 19(8):1497–1505.

Coureux P-D, et al. 2003. A structural state of the myosin V motor without bound nucleotide. Nature 425(6956):419–423.

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

Force A, et al. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151(4):1531–1545.

Hammer JA, Wagner W. 2013. Functions of class V myosins in neurons. J Biol Chem. 288(40):28428–28434.

Hodel C, et al. 2014. Myosin VIIA is a marker for the cone accessory outer segment in zebrafish. Anat Rec. 297(9):1777–1784.

Jaillon O, et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431(7011):946–957.

Karcher RL, et al. 2001. Cell cycle regulation of myosin-V by calcium/calmodulin-dependent protein kinase II. Science 293(5533):1317–1320.

Kosakovsky Pond SL, Frost SDW. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 22(5):1208–1222.

Kosakovsky Pond SL, et al. 2011. A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol. 28(11):3033–3043.
Kosakovsky Pond SL, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 21(5):676–679.
Krek A, et al. 2005. Combinatorial microRNA target predictions. Nat Genet. 37(5):495.
Kuraku S. 2010. Palaeophylogenomics of the vertebrate ancestor—impact of hidden paralogy on hagfish and lamprey gene phylogeny. Integr Comp Biol. 50(1):124–129.
Kuraku S. 2013. Impact of asymmetric gene repertoire between cyclostomes and gnathostomes. Semin Cell Dev Biol. 24:119–127.
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(1):15–20.
Li WH. 1980. Rate of gene silencing at duplicate loci. A theoretical study and interpretation of data from tetraploid fishes. Genetics 95(1):237–258.
Li JF, Nebenführ A. 2008. The tail that wags the dog: the globular tail domain defines the function of myosin V/XI. Traffic 9(3):290–298.
Mellgren EM, Johnson SL. 2005. kitb, a second zebrafish ortholog of mouse Kit. Dev Genes Evol. 215(9):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 (Cambridge, England) 134(6):1081–1090.
Murrell B, et al. 2015. Gene-wide identification of episodic selection. Mol Biol Evol. 32(5):1365–1371.
Murrell B, et al. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 8(7):e1002764.
Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol. 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 Res. 17(9):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 A. Spudich J, editor. Mol Biol Cell. 8(10):1971–1988.
Nei M, Kumar S. 2000. Molecular evolution and phylogenetics. New York, USA: 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. Berlin, Heidelberg (Germany): Springer Berlin Heidelberg.
Pond SLK, Frost SDW. 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21(10):2531–2533.
Putnam NH, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. Nature 453(7198):1064–1071.
Pylypenko O, et al. 2013. Structural basis of myosin V Rab GTPase-dependent cargo recognition. Proc Natl Acad Sci U S A. 110(51):20443–20448.
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 Eng Des Sel. 15(10):783–798.
Rodriguez OC, Cheney RE. 2002. Human myosin-Vc is a novel class V myosin expressed in epithelial cells. J Cell Sci. 115(Pt 5):991–1004.
Sittaramane V, Chandrasekhar A. 2008. Expression of unconventional myosin genes during neuronal development in zebrafish. Gene Expr Patterns. 8(3):161–170.
Smith MD, et al. 2015. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Mol Biol Evol. 32(5):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 Genet. 10: e1004614.
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. J Biol Chem. 282(32):23725–23736.
Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 30(12):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. Philos Trans R Soc Lond B Biol Sci. 356(1414):1661–1679.
Trybus KM. 2008. Myosin V from head to tail. Cell Mol Life Sci. 65(9):1378–1389.
Yang Z, Neilsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol. 19(6):908–917.
Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 22(12):2472–2479.

Associate editor: Davide Pisani