The "fish specific" Hox cluster duplication is coincident with the origin of teleosts

Research Article

Karen D Crow[1†], Peter F. Stadler[2], Vincent J. Lynch[1],
Chris Amemiya[3], and Günter P. Wagner[1]

Yale University
Department of Ecology and Evolutionary Biology
165 Prospect Street
New Haven, CT 06511

[1] Yale University
Department of Ecology and Evolutionary Biology
165 Prospect Street
New Haven, CT 06511

[2] Bioinformatik, Institut für Informatik
Universität Leipzig, D-04107 Leipzig, Germany

[3] Benaroya Research Institute at Virginia Mason
1201 Ninth Avenue
Seattle, WA 98101-2795

[†] Corresponding author:
Karen D. Crow
Yale University
Department of Ecology and Evolutionary Biology
165 Prospect Street
New Haven, CT 06511

Phone: 203 432 9999
Fax: 203 432 3870
e-mail: karen.crow@yale.edu

Keywords: Hox cluster, genome duplication, ray-finned fishes, Darwinian selection

Running Title: Hox cluster duplication in teleosts

Abstract

The Hox gene complement of zebrafish, medaka, and fugu differs from that of other gnathostome vertebrates. These fishes have seven to eight Hox clusters compared to the four Hox clusters described in sarcopterygians and shark. The clusters in different teleost lineages are orthologous, implying that a "fish specific" Hox cluster duplication has occurred in the stem lineage leading to the most recent common ancestor of zebrafish and fugu. The timing of this event, however, is unknown. To address this question, we sequenced four Hox genes from taxa representing basal actinopterygian and teleost lineages, and compared them to known sequences from shark, coelacanth, zebrafish and other teleosts. The resulting gene genealogies suggest that the "fish specific" Hox cluster duplication occurred coincident with the origin of crown group teleosts. In addition, we obtained evidence for an independent Hox cluster duplication in the sturgeon lineage (Acipenserifornes). Finally, results from *HoxA11* and *HoxB5* suggest that duplicated Hox genes have experienced strong diversifying selection immediately after the duplication event. Taken together, these results support the notion that the duplicated Hox genes of teleosts were causally relevant to adaptive evolution during the initial teleost radiation.

Introduction

Hox genes encode transcription factors associated with specification of axial patterning and the development of other characters like appendages and organ systems, and are homologous to the homeotic gene clusters observed in *Drosophila* (McGinnis and Krumlauf 1992; Schubert, Nieseltstruwe, and Gruss 1993). In vertebrates Hox genes are arranged into highly organized clusters with conservation of gene order, intergenic distances, and associated non coding sequences (Holland et al. 1994; Ruddle et al. 1994; Chiu et al. 2002; Prohaska et al. 2004). Because they play a key role in determination of body plan morphology, it has been widely assumed that they play a key role in the evolution of diverse metazoan body plans. The increased complexity of body plans that has accompanied the evolution of higher vertebrates is a phenomenon of intense interest and paramount importance (Martinez and Amemiya 2002). A particularly intriguing problem is understanding the role of Hox cluster duplications in the evolution of vertebrates (Holland et al. 1994; Malaga-Trillo and Meyer 2001; Wagner, Amemiya, and Ruddle 2003; Prohaska and Stadler 2004). The vertebrates are composed of four major groups of organisms including the agnathans (jawless fishes), the chondrichthyans (cartilaginous fishes), the sarcopterygians (lobe-finned fishes and tetrapods), and the actinopterygians (ray-finned fishes). The latter group, the ray-finned fishes, is comprised of approximately 24,000 species, 97% of which are teleosts (Nelson 1994). Teleosts are the most successful and diverse vertebrate group, and are characterized by remarkable variation in morphology, behavioral, and physiological adaptations.

The duplication of genes and entire genomes are believed to be important mechanisms underlying morphological variation and functional innovation (Ohno 1970;

Taylor, Van de Peer, and Meyer 2001; Wagner 2001). Hox clusters have undergone several rounds of duplication throughout vertebrate evolution. All non-vertebrates investigated to date including the cephalochordate *Branchiostoma* (formerly *Amphioxus*) exhibit a single Hox cluster (Garciafernandez and Holland 1994; reviewed in Martinez and Amemiya 2002). However, gnathostomes have experienced two rounds of genome duplication believed to have produced the four canonical Hox clusters of most gnathostomes, referred to as the "HoxA", "HoxB", "HoxC", and "HoxD" clusters. A subset of ray-finned fishes is known to have undergone a third round of Hox cluster duplication (Amores et al. 1998; Postlethwait et al. 1999; Naruse et al. 2000), and exhibit seven to eight clusters referred to as "Aa" and "Ab" etc. Both phylogeny and synteny data suggest that the lineage leading to the common ancestor of zebrafish and pufferfish experienced a large-scale gene or genome duplication event with subsequent marked but variable gene losses (Taylor et al. 2003; Vandepoele et al. 2004). For example the pufferfish *Takifugu rubripes* (Percomorpha) has only one HoxC cluster (Amores et al. 2004), while the zebrafish (Ostariophysi) appears to have only one HoxD cluster (Amores et al. 1998). This has been described as the "fish specific genome duplication" (Amores et al. 1998; Wittbrodt, Meyer, and Schartl 1998; Ohno 1999; Taylor et al. 2001; Taylor et al. 2003; Van de Peer, Taylor, and Meyer 2003; Vandepoele et al. 2004) and is supported by the occurrence of several other teleost specific duplicate genes (i.e. paralogs found in one or more teleosts, but not it tetrapods, Chiang et al. 2001; Lister, Close, and Raible 2001; Merritt and Quattro 2001; Kao and Lee 2002; Merrit and Quattro 2003; Winkler et al. 2003).

While the evidence supporting the Hox cluster duplication in ray-finned fishes is clear, it is not known when that duplication took place in the evolution of ray-finned fishes. Most studies have been based on comparisons between teleosts (e.g. zebrafish, *Danio rerio* and pufferfish, *Takifugu rubripes*), and sarcopterygians (lungfish, human, other tetrapods). These two lineages diverged approximately 450 mya (Kumar and Hedges 1998; Hedges and Kumar 2003). A molecular clock estimate of this duplication event, based on the comparison of paralog genes from *Takifugu rubripes*, suggests a duplication date of approximately 320 Mio years (Vandepoele et al. 2004). However, it is not known in which stem lineage the duplication event occurred because basal actinopterygian and basal teleost taxa have not been characterized with respect to Hox cluster number and orthology. Previous attempts have begun to address this question. For example, the HoxA cluster has been characterized in bichir (Polypteriformes, Chiu et al. 2004) and orthology of *HoxA11* and *HoxA13* has been characterized in the paddlefish (Acipenseriformes, Metscher et al. 2005). Finally, it has been noted that molecular phylogenies of genes (i.e. gene genealogies), not only absolute number of genes, are crucial to determining the duplication history (Furlong and Holland 2002). Our approach has been to construct gene genealogies of Hox genes from all four vertebrate Hox clusters, with a sampling strategy that includes all major basal actinopterygian and teleost clades. Actinopterygians include the bichirs, sturgeons and paddlefish (Acipenseriformes), gars, bowfin, and teleosts. While bichirs share characters with sarcopterygians, most authors consider them to be the most basal actinoptergian (Bartsch and Britz 1997; Bemis, Findeis, and Grande 1997) and recent molecular data confirm this association (Venkatesh, Erdmann, and Brenner 2001; Kikugawa et al. 2004). The teleosts are monophyletic (de Pinna 1996; Inoue et al. 2003), and include the Osteoglossomorpha, Elopomorpha, Clupeomorpha, and the Euteleostei (zebrafish, medaka, and fugu). However, it is unclear which of the remaining basal actinopterygian groups is sister group to the teleosts. The majority of data indicate that the sister group of teleosts is either the bowfin *Amia calva* (Patterson 1973; Schultze and Wiley 1984; Wiley and Schultze 1984; Nelson 1994; Bemis, Findeis, and Grande 1997), or a clade containing *A. calva* (Nelson 1969; Venkatesh, Erdmann, and Brenner 2001; Inoue et al. 2003; Kikugawa et al. 2004, Figure 1). One study suggested that gars might be sister to teleosts, based on jaw articulation (Olsen 1984), but the study did not include the appropriate outgroups (Inoue et al. 2003). In order to test whether the Hox cluster duplication is specific to teleosts, or a more inclusive clade, the sister group of teleosts is important because both the lineage exhibiting the Hox cluster duplication, and its sister group must be evaluated. Therefore it is essential that the bowfin, *Amia calva*, is included in this study because all current phylogenetic hypotheses of basal actinopterygians infer that *Amia* is, or is part of, the sister clade of teleosts.

Materials and Methods

Taxon sampling

Sequences of Hox genes from single individuals representing every major lineage of basal Actinopterygians (with one exception) and basal teleosts were obtained for this study including the following: Polypteriformes (bichir), Acipenseriformes (paddlefish and/or sturgeon), Amiiformes (bowfin), Osteoglossomorpha (goldeye), Elopomorpha (eel), Clupeomorpha (shad; Table 1). The gar (*Lepisostus platostomous*, Semionotiformes) is not included because we were unsuccessful in amplifying our target genes for this taxon. Sequences from three of the more derived euteleost lineages were obtained from public databases including Ostariophysi (zebrafish), Atherinomorpha (medaka), and Percomorpha (pufferfish). The genes from horn shark, *Heterodontus franscici* (Chondrichthyes), and Indonesian coelacanth*, Latimeria menadoensis* (Sarcopterygii), were obtained from sequenced BAC clones and used as outgroups (Table 1).

DNA extraction and polymerase chain reaction (PCR) amplification

Muscle or fin tissue was collected and preserved in 95% ethanol. DNA extraction was performed using the DNeasy Tissue Kit (Qiagen, Inc) according to the manufacturers protocols.

Four *Hox* genes were targeted to represent each of the four vertebrate clusters- *HoxA11*, *HoxB5*, *HoxC11*, and *HoxD4*. These genes were selected based on maximizing the probability of detecting duplicate paralogs, that is duplicate paralogs are known to exist in at least one taxon. Exon 1 sequences were targeted while introns were excluded due to their high variability and ambiguous alignments. Exon 2 sequences, encoding the homeodomain, also were excluded because of their characteristic conservation and associated lack of phylogenetic signal. Degenerate primers for each locus were designed from conserved regions within exon 1 from sequences of coelacanth, zebrafish, and pufferfish (Table 2). PCR amplification was accomplished using 10 to 100 ng of DNA, 0.2 μM each primer and Reddymix (ABgene, Inc.) to obtain a final reaction volume of 50 μl. Amplification cycling profiles were as follows:

(*HoxA11*) 45s at 94°C, 45s at 54°C, and 1 min at 72°C, x35 cycles;

(*HoxB5*) 45s at 94°C, 45s at 46°C, and 1 min at 72°C, x35 cycles;

(*HoxC11*) 45s at 94°C, 45s at 48°C, and 1 min at 72°C, x35 cycles;

(*HoxD4)* 45s at 94°C, 45s at 48°C, and 1 min at 72°C, x35 cycles;

Genes were cloned using the pGEM vector system (Promega). Sequencing was performed in both directions with the vector primers T7 and SP6 on an ABI 3100 Genetic Analyzer (Applied Biosystems, Foster City, CA). Several clones of each gene were sequenced to increase the probability of detecting duplicated paralogs, and to circumvent errors due to PCR. Sequences for all loci were deposited in GenBank under the following accession numbers: XXXX.

Sequence analysis

Sequences were aligned using the Clustal V algorithm (Higgins, Bleasby, and Fuchs 1992), implemented by the software MegAlign (DNASTAR, Inc.). Gene genealogies were assessed by maximum parsimony (MP), neighbor joining (NJ), Bayesian inference (BPP) implemented by the software packages PAUP (version 4.0, Swofford 1998) and MrBayes (version 2.1, Huelsenbeck and Ronquist 2001). The most parsimonious trees were obtained using a heuristic search. Statistical confidence in nodes was evaluated using 2000 bootstrap replicates (Felsenstein 1985; Hedges 1992; Hillis and Bull 1993). For Bayesian analyses models of evolution were estimated by MrModeltest (Nylander 2002) and statistical confidence in nodes was evaluated by posterior probabilities. Statistical support for nodes will be reported as (BPP, MP, NJ) unless otherwise specified. Stationarity of tree likelihood, sampled every 100 cycles, was consistently achieved after 100.000 (of one million) generations and all sampled trees preceeding stationarity were discarded (i.e. 10% of the data). Analyses were started from random trees and repeated several times for each locus to confirm that convergence had been achieved (Larget and Simon 1999; Huelsenbeck and Ronquist 2001). Alternative topologies were evaluated by posterior probability via filtering post-stationarity trees, where the number of trees consistent with the constraint divided by the total number of trees represents the posterior probability of the hypothesis.

Topologies were further evaluated using a split decomposition method that allows conflicting phylogenies to be simultaneously visualized. Therefore conflicting data are not forced on one unique topology, but rather are depicted as networks. These phylogenetic networks were computed using the neighbor-net method (Bryant and Moulton 2004), as implemented in the SplitsTree package (Huson 1998). This method is a generalization of the well-known neighbor-joining algorithm (Saitou and Nei 1987). The HKY85 distance transformation was used for all loci and statistical confidence in splits was expressed as percent bootstrap support from 1000 replicates.

Finally, we employed the quartet mapping technique of Nieselt-Struwe and von Haeseler (2001) using our own implementation, quartm, which is available from www.bioinf.uni-leipzig.de/Software/. In this technique the sequences are partitioned into four groups including an outgroup, "a" paralog group, "b" paralog group, and the sequence of interest. The quartet mapping directly tests the support for a particular split of interest without regard for the detailed structure within all four potential topologies, effectively reducing noise (Stadler et al. 2004).

Tests for relative rates of evolution and selection

Banch lengths were estimated by the software package HyPhy (version .99beta, Kosakovsky Pond, Frost, and Muse 2004) using the codon model of Goldman and Yang (1994). We tested for selection in specific lineages by estimating the non-synonymous to synonymous substitution rate ratio ($d_N/d_S=\omega$), using codon based maximum likelihood models of sequence evolution (Goldman and Yang 1994) implemented in the software package PAML (version 3.14, Yang 1997).

Estimating the time between specific nodes

Because some inter-nodes were estimated to have acquired no or very few synonymous substitutions, we were interested to obtain a rough estimate of the time between successive nodes. To accomplish this we used a likelihood model for the expected number of synonymous substitutions given either zero or a small number of synonymous substitutions in a number of equally long branches (Appendix A). If there are zero synonymous substitutions in *k* branches of equal length the likelihood function is a strictly decreasing function of the temporal branch length. A maximum likelihood estimate would thus suggest a zero branch length. Because we did detect non-synonymous substitutions along the same branch this is an unreasonable result. We instead used the median of the likelihood function as an estimate (Appendix A)

$$T_{\text{median}} = \frac{-\ln 0.5}{k\mu S}$$

where μ is the per nucleotide mutation rate, *k* is the number of branches with zero synonymous substitutions and *S* is the average number of synonymous sites in the sequences compared. For *x*>0 substitutions in *k* equally long branches the likelihood function is mono-modal and we can use a conventional maximum likelihood approach:

$$T_{\max L} = \frac{x}{k\mu S}$$

The estimates from the median likelihood method and the maximum likelihood approach are in reasonable agreement (see below).

Results

Sequences

Sequences were obtained for 14 taxa, representing one chondrichthyan, one sarcopterygian, all extant lineages of basal Actinopterygians (except Semionotiformes), and three euteleosts (Table 1). Twenty clones were sequenced for each locus resulting in one to eleven replicate sequences for each gene detected (Table 3). Only sequences that could be aligned unambiguously were included, and sequences spanning indels were excluded. Of a total 2496 bp sequenced, 1593 bp were considered for further analyses including 438 bp for *HoxA11*, 423 bp for *HoxB5* (Table 3), 435 bp for *HoxC11*, and 297 bp for *HoxD*. Finally, the *HoxC11* locus was ultimately excluded because we were unable to obtain sequences from taxa representing the chondrichthyan (shark) and basal actonopterygian lineages (bichir, paddlefish and sturgeon), and only a truncated sequence was found for the bowfin, *Amia calva.* However, it should be noted that only one *HoxC11* paralog was detected for all of the four teleost taxa investigated-goldeye, tarpon, eel, and shad. The zebrafish, *Danio rerio,* exhibits two *HoxC11* paralogs, but medaka and pufferfish are known to have only one HoxC cluster. Therefore it is possible that our inability to detect sequences for basal taxa, and duplicate paralogs for teleosts is due to secondary loss.

Plots of transitions and transversions versus genetic distance for each locus indicated sequences were not saturated (data not shown). Observed transition to transversion ratios varied from 1.47 – 2.49 (Table 4), and these data were incorporated in model selection for likelihood analyses.

Gene trees-HoxA11

Our data set of *HoxA11* genes contains 18 sequences. Ten of these were previously published including shark, coelacanth, bichir, paddlefish , zebrafish (Table 1) or extracted from genome sequence data bases (fugu, medaka). These data contain two known paralogs for zebrafish, fugu and medaka and single orthologs for shark, coelacanth, bichir and paddlefish. In accordance with this pattern our data contain a single sequence for the basal actinopterygians, the pallid sturgeon, *Scaphirhynchus albus*, and the bowfin, *Amia calva*; and two distinct sequences for the teleost species, goldeye, *Hiodon alosoides,* and American eel*, Anguila rostrata*. We found only one sequence for the remaining two teleosts investigated, the tarpon, *Megalops atlanticus*, and the shad, *Drososoma cepedianum*. Sequences were validated as *HoxA11* genes and tentatively identified as orthologous to known *HoxA11* paralogs based on BLAST alignments and gene tree topology.

The maximum likelihood tree estimated from *HoxA11* sequences, which was identical to the consensus tree estimated from Bayesian methods, is consistent with accepted features of ray-finned fish phylogeny including the basal position of the bichir lineage, the close affiliation of paddlefish and sturgeon and the monophyletic character of teleost lineages (Figure 2). Our data infer the bowfin as the sister taxon to teleosts and independent of the sturgeon clade (100/63/-). This topology is consistent with the recent nuclear phylogeny proposed by Kikugawa et al. (2004), but varies from the topology proposed by Inoue et al. (2003) based on mitochondrial genomes, who suggested that the holosteans (gars and bowfin) and the Acipenseriformes (paddlefish and sturgeon) form a clade which is sister to teleosts (Figure 1). All duplicated *HoxA11* paralog sequences are grouped in a well supported clade (100/98/100) indicating that the bowfin lineage diverged prior to the duplication event that generated the paralog genes found in teleosts (Figure 2). The unrooted topology from neighbor nets indicate a significant split between duplicated and unduplicated taxa with a bootstrap value of 99.9 percent (Figure 3). Finally, quartet mapping plots provide further support that *Amia calva* diverged prior to the duplication that gave rise to the paralogs in zebrafish and fugu (Figure 4).

To determine if the topology inferred from the *HoxA11* data were significantly different from two alternative hypotheses with respect to the timing of the duplication, we evaluated the posterior probabilities of the following hypotheses: (1) *Amia calva* diverged after the *Hox* cluster duplication and the gene detected is associated with the "a" or "b" paralog group; and (2) *Hiodon alosoides*, belonging to the most basal teleost lineage, is independently duplicated and the paralogs detected are not associated with the known zebrafish "a" and "b" paralog clades. The topology inferred from the *HoxA11* data is significantly different from both alternative hypotheses, which exhibited posterior probabilities of <1/900 (i.e. none of the post-stationarity trees from the Bayesian analysis were consistent with these hypotheses). Therefore both alternative hypotheses were rejected and we are left with the conclusion that the bowfin *HoxA11* gene is not orthologous to either of the duplicated zebrafish paralog genes.

Within the clade of teleost genes there are two well resolved clades grouping the known "a" and "b" paralogs with the new sequences found in this study (including the single genes detected for the goldeye and eel). The affiliation of these sequences with known paralogs identifies them as orthologs of the known "a" and "b" paralogs of *HoxA11*. Support for the two paralog clades is strong in the Bayesian analysis, but lacking in the MP and NJ analyses. This lack of resolution for distinct "a" and "b" clades is reflected in the neighbor net from the splitstree analysis shown in Figure 3. The overall topology inferred from *HoxA11* sequences indicates that a single duplication event occurred prior to the most recent common ancestor of teleosts and after the divergence from the bowfin lineage.

Sequence evolution-HoxA11

The total branch lengths of the teleost gene lineages are consistently longer than those of basal actinopterygian lineages. We compared the non-synonymous substitution rate of each teleost gene to the teleost sister taxon, *Amia calva*, using the sturgeon sequence as outgroup and found that all teleost lineages evolve significantly faster (Table 4). The probability that this consistent asymmetry would occur by chance is $2.4\text{x}10^{-4}$. Increased rates of evolution are expected if gene lineages are derived from gene duplication events (Lynch and Conery 2000; Kondrashov et al. 2002; Conant and Wagner 2003; Wagner et al. 2005).

An analysis of substitution rates using maximum likelihood revealed surprising clues about the relative timing and the evolutionary forces acting after the gene duplication. We tested for selection by using codon-based maximum likelihood models of sequence evolution (Goldman and Yang 1994; Yang 1997) to estimate the non-synonymous to synonymous substitution rate ratio ($d_N/d_S=\omega$) and evaluate specific lineages and amino acid sites under positive selection. We used the "ancient fish" phylogeny of basal actinopterygian fishes proposed by Inoue et al. (2003, Figure 1a) based on mitochondrial genomes as the input tree, with the rearrangement of holosteans proposed by Kikugawa et al. (2004, Figure 1b). Initial branch lengths were estimated using a one-ratio model (d_N/d_S same for all branches) to establish a null model for comparison in more complex analyses.

The one ratio model shows that the average non-synonymous to synonymous ratio (ω) is about 0.14, indicating moderately strong stabilizing selection averaged over all lineages. The two ratio model in which the two post-duplication branches have a different rate from the rest of the tree reveals evidence for strong directional selection immediately following the duplication event.

The model estimates 9.2 non-synonymous substitutions in the stem of the “a” paralog clade and 6.7 for the stem of the “b” paralog clade. In neither branch are any synonymous substitutions estimated to have occurred (G=10.45, $P<0.001$ based on chi-square approximation). Very similar estimates also are recovered from the free ratio model. In addition, the free ratio model suggests two additional episodes of strong directional selection along internal branches. In the “a” clade 4.6 non-synonymous substitutions and no synonymous substitutions are estimated after the divergence of Osteoglossomorphs and before the most recent common ancestor of Elopomorphs and the more derived teleosts. In the “b” clade, 18 non-synonymous substitutions and no synonymous substitutions were estimated to have occurred in the stem lineage of the more derived teleosts after the divergence of Elopomorphs. Note that after the initial period of simultaneous divergence following the duplication, these episodes of strong directional selection did not occur at the same time in the two paralog groups, indicating that the paralogs experience adaptive evolution differentially after duplication.

The absence of synonymous substitutions mapped to the post duplication branches suggests that the time between the duplication event and the divergence of the most basal extant teleost lineage, the mooneyes (Osteoglossomorpha), was very short. Assuming a standard eukaryotic per nucleotide mutation rate of 10^{-9} (Graur and Li, 2000) and a Poisson process for synonymous substitutions, it is possible to estimate the time between the most recent common ancestor of the teleosts and the duplication event. A median likelihood method (see Appendix A) yields an estimate of 3.5 Mio years. This is a very short time compared to the molecular clock estimate of the duplication event of 320 Mio years ago (Vandepoele et al. 2004). Hence it is likely that the post duplication branches occupy only about 2% of the time since the duplication. Furthermore these results imply that our ability to detect distinct “a” and “b” clades for *HoxA11* rests entirely on those non-synonymous substitutions caused by strong directional selection following the duplication.

HoxB5

For the *HoxB5* analysis we considered 21sequences including 14 reported here for the first time. These sequences include known *HoxB5* paralogs for zebrafish, fugu and medaka, and a single copy gene from shark. We report new single sequences for coelacanth, bichir, and bowfin among the basal actinopterygian fish lineages. Two copies of *HoxB5* were found in the sturgeon, and all the basal teleost taxa examined: goldeye, tarpon, eel, and shad. The teleost genes were provisionally assigned to a teleost paralog group based on BLAST alignments and gene tree topology. We call these paralogs α and β, suggesting orthology to the zebrafish "a" and "b" paralogs respectively, but emphasize that this assignment is preliminary.

The *HoxB5* maximum likelihood tree, which was identical to the consensus tree of 900 post stationarity trees from the Bayesian analysis, indicates that the two sequences obtained from the sturgeon, *Scaphirhynchus albus*, along with the paddlefish, *Polyodon spathula*, form a well supported clade (100/100/100) suggesting that this duplication was independent of the one creating the teleost paralogs (Figure 5). Interestingly, the sturgeon *HoxB5*-1 sequence was associated with the single sequence obtained from the paddlefish with support values of 100/89/95, indicating that the gene duplication in the sturgeon lineage occurred before the divergence of paddlefishes and sturgeons. Thus, it is possible that paddlefish has an additional *HoxB5* copy or an entire second HoxB-cluster that went undetected, or has lost one copy of the HoxB cluster. The phylogenetic position of the bowfin inferred from the *HoxB5* sequences was consistent with the hypothesis that *Amia calva* is the sister taxon of teleosts with high levels of support (100/83/79, Figure 5). This topology indicates that teleosts form a monophyletic clade but support for this clade was mixed (100/-/-), and this lack of consistency is reflected in the *HoxB5* neighbor net (Figure 6). In order to test whether the bowfin gene sequenced in our study could be orthologous to one of the known teleost duplicates, we evaluated the alternative hypothesis that the bowfin *HoxB5* sequence is associated with the *HoxB5a* clade or the *HoxB5b* clade. These hypotheses were rejected because they exhibited posterior probabilities of <1/900 (i.e. none of the post-stationarity trees from the Bayesian analysis were consistent with either hypothesis). Quartet mapping provided no resolution as to whether the bowfin *HoxB5* sequence was associated with duplicated paralogs or unduplicated genes. We conclude that the balance of evidence supports the notion that the bowfin lineage diverged prior to the duplication of the teleost *HoxB5* genes. The topology of the *HoxB5* gene lineages is well resolved within the teleosts. There are two well supported clades uniting each of the two new shad *HoxB5* sequences with one of the known *HoxB5* zebrafish paralog genes with support values of (100/95/98) and (100/95/96) for the "a" paralog and "b" paralog clades respectively. The latter clade is joined by the fugu and medaka *HoxB5b* paralog clade. These results indicate that the two shad sequences are orthologous to the known duplicated teleost *HoxB5* paralogs. There are also well supported clades uniting the eel and tarpon "α" paralog and "β" paralog sequences with (71/67/88) and (100/77/92) support respectively, indicating that the duplication resulting in these paralogs occurred prior to the split of the eel and tarpon lineages, i.e. prior to the most recent common ancestor of all extant elopomorphs. The affiliation of the goldeye *HoxB5* sequences remains uncertain. Still, there is no evidence that these genes were independently duplicated in the *Hiodon* lineage. (i.e. none of the post-stationarity trees from the Bayesian analysis were consistent with the hypotheses that (1) the two *Hiodon alosoides* paralogs were independently duplicated or (2) the two *Hiodon alosoides* paralogs were not associated with the known zebrafish “a” and “b” paralog clades). Hence within the clade of teleost gene lineages there are six highly supported clades, four of which are clearly associated with one of the known *HoxB5a/b* paralogs. Finally, a resolved “b” paralog clade was recovered in the Bayesian analysis, but with limited support (BPP=63). Overall, the *HoxB5* topology is consistent with the hypothesis that both *HoxA11* and *HoxB5* were duplicated at the same time-before the most recent common ancestor of teleosts.

Sequence evolution-HoxB5

To further test whether the bowfin *HoxB5* gene could be duplicated, we compared the rate of non-synonymous substitutions of bowfin *HoxB5* with that of the teleost sequences using the codon-based model by Goldman and Yang (1994). All of the 12 teleost sequences used in this comparison exhibit a higher estimated rate of non-synonymous substitutions than the bowfin *HoxB5* sequence. Again, the probability that this consistent asymmetry is observed by chance is $2.4x10^{-4}$. Of these comparisons, only those with the tarpon and eel sequences are not significant using individual p-values. With the Bonferroni correction for multiple comparisons the comparison to zebrafish *HoxB5a* sequence also was not significant (Table 4). All estimated rates for teleost sequences are higher than that estimated for the bowfin sequence, and seven of the twelve comparisons are statistically significant. In order to see whether this rate difference can be attributed to a difference between teleosts and more basal fish lineages (i.e. phylogenetic vs. post-duplication rate acceleration) we compared the rate of bowfin *HoxB5* with the two independently duplicated sturgeon *HoxB5* sequences. The estimated rate of sequence evolution for the duplicated sturgeon genes also is higher than the bowfin sequence and these differences are significant in the amino acid model (P=0.025 and 0.033 respectively), but not in the Goldman/Yang (1994) codon model. This indicates that, for these data, rate acceleration is associated with gene duplication and not with taxon or lineage. Overall the rate comparisons support the conclusion that the bowfin *HoxB5* gene is not derived from a gene duplication in the actinopterygian lineage.

In tests for selection in post duplication lineages, the one ratio model for the *HoxB5* data set estimates an average ω value of 0.17, similar to that estimated for *HoxA11*, indicating that these genes are generally evolving under stabilizing selection. The two-ratio model, where the two post-duplication branches can evolve at a different rate than the rest of the tree, indicates strong selection in the “b” paralog branch with 3.4 non-synonymous substitutions and no synonymous substitutions. While these values indicate strong selection, the amount of evolution along this branch is still less than that found for the *HoxA11* post-duplication branches (eight non-synonymous substitutions on average). On the post-duplication branch for the *HoxB5a* paralog group, 5.3 non-synonymous substitutions and 2.1 synonymous substitutions were estimated, corresponding to an ω=0.8567. This value is fivefold higher than the average ω value of 0.17, but is short of ω>1, which would be necessary to formally demonstrate directional selection. Hence *HoxB5* also experienced higher dN/dS ratios and directional selection, at least in the b-paralog lineage, but the strength of selection appears to have been weaker than in *HoxA11*. Therefore, it is not surprising that the basal node of the paralog group clades of the teleost *HoxB5* genes is not fully resolved.

The low number of synonymous substitutions in the post-duplication branches from *HoxA11* and *HoxB5* were consistent in that post-duplication stem lineages exhibit very short branch lengths. Combining the estimates of the *HoxA11* and *HoxB5* analyses yields *k=4* branches of the same length, in terms of absolute time, and a total of two substitutions with an average number of synonymous sites of *S=103*. Assuming the standard eukaryote per nucleotide mutation rate of 10^{-9} (Graur and Li, 2000), we obtain a maximum likelihood estimate of the time between the cluster duplication (see Appendix) and the most recent common ancestor of teleosts of 5 Mio years. This value is roughly consistent with the median likelihood estimate based on the *HoxA11* data alone of 3.4 Mio years. Therefore, the duplication occurred shortly before the divergence of Osteoglossomorphs (the most basal teleost lineage) from the other Teleosts, leaving relatively little time for the build up of phylogenetic signal for the duplication event.

HoxD4

The situation with the HoxD clusters of ray-finned fishes is considerably more complicated than for HoxA and HoxB clusters. In teleosts the number of HoxD clusters is variable, with zebrafish having one HoxD cluster, while medaka and pufferfish exhibit two HoxD clusters (Amores et al. 1998; Naruse et al. 2000; Amores et al. 2004). The phylogenetic relationships among these clusters is unclear. Most of the evidence indicating that first order paralog Hox clusters in teleosts were duplicated in one event is derived from information about the HoxA and HoxB clusters (Amores et al., 2004; Prohaska and Stadler 2004). We also experienced difficulty in recovering information from our *HoxD4* sequences when we attempted of analyze all available sequences simultaneously. Therefore, we have resorted to a focused interrogation of the data, with stepwise addition of certain sequences to evaluate specific hypotheses. We are reporting support values from analyses of amino acid residues for this gene, but note that results from analyses of nucleotide data were congruent, unless stated otherwise.

As a first step we analyzed the full exon 1 amino acid sequences of shark, coelacanth, zebrafish, and the euteleosts medaka, fugu and an additional pufferfish, *Sphoeroides nephalus*, to determine whether the zebrafish *HoxD4* gene is in fact orthologous to a euteleost *HoxD4* paralog. In effect this tests whether the two HoxD clusters of euteleosts are independenly duplicated or whether a HoxD cluster was lost in the zebrafish lineage. We will use this data set as a reference against which the new sequences will be aligned. Within the reference data set the zebrafish *HoxD4* gene consistently groups with the pufferfish and medaka *HoxD4a* sequences with support values of (100/64/99). This result is consistent with the results of *HoxD9*, i.e. that the zebrafish *HoxD9* gene is orthologous to *HoxD9a* (Prohaska and Stadler 2004). Therefore, we conclude that the single zebrafish HoxD cluster is a *HoxDa* cluster and that the *HoxDb* cluster was lost in the zebrafish lineage.

Among the taxa investigated here one sequence has been found from bowfin, and shad, and two paralogs from goldeye, tarpon and eel. When the two amino acid sequences of the goldeye are confronted with the reference data set, they associate with the *HoxD4a* (100/86/98) and *HoxD4b* (100/90/100) clades respectively. Hence we conclude that the two paralogs reported here for the goldeye arose prior to the most recent common ancestor of crown group teleosts, consistent with the results from *HoxA11* and *HoxB5*.

The bowfin sequence, when confronted with the reference data set, did not provide resolution on the issue of whether the duplication occurred in the stem lineage of teleosts, or before the most recent common ancestor of bowfin and teleosts. In the analysis of amino acid data, the bowfin sequence is associated with the pufferfish and medaka *HoxD4a* sequences with support values of (100/55/63). This contrasts the results from an analysis of nucleotide data in which a well supported *HoxD4a* clade (100/77/66) excludes the bowfin sequence, but the nucleotide alignment did not include “b” paralog sequences. Therefore, we conclude that the bowfin *HoxD4* sequence could not provide consistent evidence for a duplication that predated or postdated the divergence of the bowfin and teleost lineages.

The amino acid sequences of the four genes cloned from the two elopomorph species, tarpon and eel, were also aligned to the reference dataset and analyzed. In all analyses the four elopomorph genes are significantly associated with the *HoxD4a* clade. This pattern suggests that the elopomorph *HoxD4* paralogs described here may have arisen through an independent gene duplication in the stem lineage of elopomorphs, sometime after the *HoxD4a* and *HoxD4b* paralogs diverged. If so, the gene that was duplicated in this event would be orthologous to *HoxD4a*. Note that zebrafish has only a *HoxDa* cluster. Therefore it is possible that the *HoxDb* cluster has been lost independently in different lineages.

Overall we conclude that the *HoxD4* data set is consistent with the scenario deduced from the *HoxA11* and *HoxB5* datasets, but provides less resolution (i.e. that there was a duplication of the Hox clusters prior to the most recent common ancestor of crown group teleosts and after the split of the bowfin/teleost lineages)

Discussion

The timing of the “fish specific Hox cluster duplication,” inferred from gene genealogies of Hox genes from three different clusters, is estimated to have occurred coincidental with the origin of teleosts. More specifically, we found no support for this duplication to have occurred in the basal actinopterygian lineages Polypteriformes, Acipenseriformes, or Holosteans (gars and bowfin, represented here by the bowfin *Amia calva*), but found definitive support for a gen(om)e duplication in the most basal teleost lineage Osteoglossomorpha, and other teleost lineages. This scenario is consistent with the results of a recent study of non-Hox genes which concluded that Sox11 and tyrosinase also were duplicated after the most recent common ancestor of holosteans and prior to the teleost radiation (Hoegg, et al., 2004).

Sequences of Hox genes representing all extant basal lineages of Actinopterygians were included in this study, with the exception of the Semionotiformes (gars). While we did attempt to include the shortnose gar (*Lepisosteus platostomus)* without success, this is not expected to adversely affect our results because it is unlikely that Semionotiformes are the sister group to teleosts. Furthermore, recent molecular data based on non-Hox genes independently indicate that gars do not exhibit the genome duplication exhibited by teleosts, and further corroborate our findings that the genome duplication in actinopterygian fishes is specific to teleosts (Hoegg et al. 2004). Therefore, this is the first study to address the timing of the “fish specific” Hox cluster duplication directly, with the appropriate sampling regime including representatives from all lineages in basal actinopterygians that could be sister to teleosts, and basal teleost lineages.

Our conclusions are drawn from gene genealogies inferred from sequences of *HoxA11*, *HoxB5* and *HoxD4*. The data from *HoxA11* and *HoxB5* are clear and unambiguous with respect to the timing of the duplication. The data from *HoxD4* were less clear, but consistent with the same conclusions. The data from our *HoxC11* sequences were uninformative with respect to the timing of the duplication.

We found two paralogs for most teleost taxa and genes examined, and were able to assign paralog groups with confidence if there was significant support for all analyses (eg. BPP/MP/NJ). The shad (*Dorosoma cepedianum,* Clupeomorpha) is a close relative of the zebrafish (*Danio rerio,* Ostariophysi), therefore Hox gene sequences from these taxa were usually significantly associated. We assigned the one *HoxA11* and both *HoxB5* paralogs in shad with confidence. With respect to *HoxD4* paralog sequences, we were able to infer the association of the zebrafish *HoxD4* sequence with the euteleost *HoxD4a* paralog group, establishing its orthology. In addition, we assigned orthology of shad *HoxD4a* and goldeye *HoxD4a* and *HoxD4b* paralogs with confidence. We were able to tentatively assign orthology (i.e. bootstrap support in one or more analyses) for the following number of *HoxA11* paralogs: goldeye (2), eel (2), and tarpon (1-*HoxA11α);* and the *HoxB5* paralogs of goldeye (*HoxB5β),* eel (2), and tarpon (2). We parsimoniously deduce the orthology of the goldeye *HoxB5α,* based on absence of evidence for independent duplication in that taxon. We uncovered two *HoxD4* paralogs for both tarpon and eel. However, the data indicate that these genes may have arisen from an independent duplication of the *HoxD4a* gene in the common ancestor of elopomorphs. We did not detect sequences in tarpon and eel that were clearly orthologous to *HoxD4b*, and note that, like the zebrafish, the *HoxD4b* cluster may have been secondarily lost.

Relative rates of evolution and tests for selection further supported the timing of the Hox cluster duplication as occurring in the stem lineage of teleosts. Additional insights gained from evolutionary rates included estimates of the time between nodes around the duplication, and an explanation for the variance in strength of signal between loci and analyses.

First, it is well established that duplicated genes experience an increased rate of evolution following duplication (Lynch and Conery 2000; Kondrashov et al. 2002; Conant and Wagner 2003; Wagner et al. 2005). Branch lengths of teleost (duplicated) lineages were consistently longer than basal actinopterygian (non-duplicated) lineages for both *HoxA11* and *HoxB5* based on the number of non-synonymous substitutions. Furthermore we were able to rule out the possibility that the observed rate acceleration is an artifact of common ancestry (i.e. phylogenetic rate acceleration) because an independent *HoxB5* duplication occurred in the acipenseriform lineage, which allowed an additional post-duplication rate comparison. These paralogs (two sturgeon and one paddlefish) also exhibit accelerated rates of evolution compared to non-duplicated gene lineages. Therefore the observed rate acceleration is associated with gene duplication. These data imply that the bowfin genes are not derived from a duplication event during ray finned fish evolution.

Second, the lineages immediately following the duplication in both *HoxA11* and *HoxB5* consistently exhibit very short branch lengths (in terms of absolute time) as indicated by the low number of synonymous substitutions reconstructed on these branches. We estimate that the time between the cluster duplication and the most recent common ancestor of teleosts is approximately 3.4 - 5 Mio years (based on the *HoxA11* data, and *HoxA11* and *HoxB5* data combined, respectively). Therefore, the duplication occurred shortly before the origin teleosts, leaving relatively little time for the build up of phylogenetic signal for the duplication event. Incidentally, there is another internal branch exhibiting no synonymous substitutions in both the *HoxA11* and the *HoxB5* free-ratio analyses. This is the internal branch leading to the most recent common ancestor of elopomorphs and the more derived teleosts. Combining the *HoxA11* and the *HoxB5* results, there are k=4 branches with no synonymous substitutions leading to a median likelihood estimate of 1.7Mio years for this branch. This suggests that the three principal lineages of crown group teleosts-osteoglossomorphs, elopomorphs and the remaining teleosts-originated within approximately 7 Mio years after the Hox cluster duplication. Thus, it is clear that any gene tree reconstruction will exhibit limited signal in the branching order of basal teleost clades unless strong directional selection increased the rate of evolution along those branches.

Finally, positive Darwinian selection in post-duplication lineages can be responsible for functional divergence and innovation (Ohno 1970). However, post duplication selection has been difficult to detect using traditional methods of cumulative dN/dS ratios because the signature of selection is expected to attenuate in 30-50 my due to purifying selection after adaptive evolution (Hughes 1999). To compensate for this limitation, Van de Peer and colleagues (2001) compared 26 duplicated zebrafish genes with mouse orthologs for signs of selection based on radical and conservative amino acid changes in charge or polarity. Few genes showed evidence for selection, but two of three

Hox genes evaluated did exhibit positive Darwinian selection based on cumulative amino acid substitutions resulting in a change in polarity. Recent codon based likelihood models have been developed that allow for variable dN/dS ratios among lineages (Messier and Stewart 1997; Yang 1998). Using these techniques Fares et al. (2003) found evidence for positive Darwinian selection in the stem lineage of *HoxA7* genes of vertebrates (but not *HoxB7*), and in the more recent post duplication branch of *HoxB7b* in the tetraploid *Xenopus*, but pre- and post-duplication lineages were not compared specifically.

We evaluated pre- and post-duplication lineages to see if there was evidence for positive Darwinian selection using the methods of Yang (1998) to detect episodes of selection in different lineages. We found evidence for strong positive selection in the lineages immediately following a duplication event in both *HoxA11* and *HoxB5* in actinopterygian fishes. This is the first evidence explicitly demonstrating directional selection in the lineages immediately following a Hox gene duplication event. This is neither consistent with the neofunctionalization model of paralog retention, which implies an active process of functional adaptation in one paralog but not in the other, nor the subfunctionalization of protein domains or differential expression patterns, as hypothesized by Force et al. (1999) in the duplication-degeneration-complementation (DDC) model. The classical neo-functionalization model assumes that one paralog acquired a novel function while the other paralog preserves the original function. Strong selection on both paralogs immediately following the duplication is inconsistent with this model. On the other hand the DDC model requires the passive build up of degenerative mutations, which is contradicted by the lack of synonymous substitutions in the post duplication branches. Strong directional selection on both paralogs immediately following the duplication has been documented only in gene families where sequence diversity can be directly adaptive, like pathogen resistance genes and olfactory receptor genes (Hughes 1999) but not for transcription factor genes. The evidence for strong directional selection on Hox genes implies a more active role of selection for the maintenance of duplicated Hox clusters.

Is there a correlation between Hox cluster duplication and teleost diversity?

While the increased numbers of *Hox* clusters have not yet been directly linked to evolutionary opportunities for increased body complexity, a major question in Hox cluster evolution is elucidating the causes and effects of increasing cluster number and conserved cluster composition throughout chordate phylogeny. The most widely accepted explanation for the duplication of Hox gene clusters is whole genome duplications coincident with the origin of vertebrates and gnathostomes, and again in the phylogeny of ray-finned fishes (Meyer and Schartl 1999; Prohaska and Stadler 2004). Meyer and Schartl (1999) speculated that the first genome duplication in chordate evolution may have predated the Cambrian explosion, the second early in the Devonian, and later in the Devonian, the genome of ray finned fishes was duplicated for a third time. Many have argued that each of these genome duplications were accompanied by dramatic jumps in morphological complexity, adaptive radiations, and innovations in body design (reviewed in Donoghue and Purnell 2005, in press). Likewise, it has been proposed that increased genomic complexity of fishes, due to multiple rounds of genome duplication, have contributed to their evolutionary success and diversity (Zhou, Cheng, and Tiersch

2001). Basal actinopterygians, referred to as “ancient fish,” including bichirs, sturgeons and paddlefish, gars, and bowfin, are relatively species poor. Together they comprise only 2.4% of extant actinopterygians (567 of approximately 23,700 species). Teleosts are comprised of over 23,000 species (Nelson 1994)-nearly half of all vertebrate species. By all accounts, teleosts are considered a highly successful and diverse group, and here we present evidence that the fish specific Hox cluster duplication is statistically coincidental with the origin of the basal crowngroup teleosts. However, two groups of teleosts have undergone expansive radiations and account for the majority of species richness associated with teleosts-the Ostariophysi and the Perciformes. These groups were not the first to radiate after the gen(om)e duplication, and most basal teleost group, the Osteoglossomorpha, is not characterized by an explosion in species richness. Furthermore, when extinct forms are considered, teleosts do not exhibit greater species diversity than extinct basal actinopterygian lineages, nor is increased complexity of body plans coincident with teleosts (Donoghue and Purnell, in press). Therefore, palaeontological data provide no support for congruence between gen(om)e duplications and body plan complexity or species diversity (Donoghue and Purnell, in press).

To summarize, the argument for a correlation between Hox cluster number and the origin of higher complexity or diversity is not supported. Invertebrates exhibit a greater variety of body plans, and far greater diversity in species richness than any vertebrate group, yet exhibit, at most, single Hox cluster (Carroll 1995). Sarcopterygians exhibit greater complexity and diversity than cartilaginous fishes, yet both groups exhibit the same number of Hox clusters (Robinson-Rechavi, Boussau, and Laudet 2004). And while actinopterygians exhibit more Hox clusters than sarcopterygians, and greater species diversity, it has been argued that a zebrafish is not more complex than a mouse (Bruce et al. 2001). Finally, as stated before, fishes with 7-8 Hox clusters (i.e. teleosts) do not exhibit greater species diversity than fishes with fewer Hox clusters (i.e. basal actinopterygians) when extinct forms are considered. Still, it is clear that Hox cluster duplication and gene retention has played a prominent role in the evolution of vertebrates, but that role is yet to be fully characterized (Wagner, Amemiya, and Ruddle 2003).

The scenario that emerges from the present study is a clear estimate of the phylogenetic timing of the "fish specific" Hox cluster duplication, immediately followed by a period of directional selection on some of these genes. Relatively quickly thereafter lineages diverge, which then continue to exhibit increased rates of evolution compared to non-duplicated lineages. This corresponds to a post duplication "window of evolvability" due to relaxed constraint that has been previously postulated (Wagner, Amemiya, and Ruddle 2003) and is supported by the pattern and frequency of transposable elements in vertebrate and invertebrate Hox clusters (Fried, Prohaska, and Stadler 2004). This pattern may explain the weak phylogenetic signal in gene lineages immediately following a duplication, observed in this study for *HoxD4* and others (Robinson-Rechavi et al. 2001; Hoegg et al. 2004) because resolution can only be expected if the focal gene was subject to strong directional selection, like *HoxA11*.

Is it possible that the occurrence of duplicated Hox clusters, or entire genomes, is associated with a decreased probability of extinction via functional redundancy, post duplication increased rates of evolution, directional selection, and adaptation? Our observation is that these processes are measurable and have been shown to be associated with gen(om)e duplications. And, remarkably, estimates of gen(om)e duplications in vertebrates are preceded by multiple extinct lineages resulting in pre-duplication gaps in extant taxa (illustrated, but not pointed out, in Donoghue and Purnell, in press).

We propose a model in which gen(om)e duplication in vertebrates is associated with increased evolvability, which in turn contributes to reduced probabilities of extinction and, eventually, potential for diversification that is associated with a gen(om)e duplication in vertebrate lineages. Gen(om)e duplication initially provides the genetic redundancy necessary to confer robustness against null mutations (Gu et al. 2003) and possibly other deleterious effects of mutations, while opening a window of relaxed constraint and increased rates of evolution (Wagner, Amemiya, and Ruddle 2003; Fried, Prohaska, and Stadler 2004). This may provide the opportunity for genetic variability to accrue, which would be necessary for directional selection, adaptation, and functional innovation to occur. This is particularly true in a gene family which exhibits strong stasis over long periods of phylogenetic time (compare Shark and human HoxA clusters, Chiu et al. 2002). We note that the time necessary for these evolutionary processes to unfold and contribute to species richness would not predict immediate explosive radiations or jumps in phenotypic complexity. Rather, adaptive evolution in one or both paralogs would result in the build up of co-adapted gene complexes, which form the basis for the evolution of reproductive isolation via Dobzhansky-Muller incompatibilities (Dobzhansky 1936; Muller 1942; reviewed in Orr 1995). Lynch and colleagues recognized the significance of genomic redundancies due to gen(om)e duplication as a powerful substrate for the origin of genomic incompatibilities in isolated populations (Lynch and Conery 2000; Lynch and Force 2000), as first noted by Werth and Windham (1991). Hox gene clusters are particularly likely to be affected by this because Hox genes exhibit colinearity, and form co-adapted gene complexes with shared regulatory and non-coding sequences. Finally, few genes that are associated with speciation have been characterized, however, the emerging theme is that "speciation genes" are under positive Darwinian selection (Orr, Masly, and Presgraves 2004). These aspects of Hox genes may have been under emphasized in their correlation with species diversity and the evolution of complexity.

Acknowledgements

Financial support for this paper was provided by the National Science Foundation, grant IBN#0321470, and intramural research funds from Yale University. PFS is supported by Bioinformatics Initiative of the Deutsche Forschungsgemeinschaft, grant BIZ-6/1-2. The medaka sequences were provided freely by the National Institute of Genetics and the University of Tokyo for use in this publication only. Tissue samples of *Scaphirhynchus albus, Amia calva, Hiodon alosoides*, *Megalops atlanticus*, *Anguilla rostrata* and *Dorosoma cepedianum* were provided by Guillermo Orti.

Literature Cited

Amores, A., A. Force, Y. L. Yan, L. Joly, C. Amemiya, A. Fritz, R. K. Ho, J. Langeland, V. Prince, Y. L. Wang, M. Westerfield, M. Ekker, and J. H. Postlethwait. 1998. Zebrafish hox clusters and vertebrate genome evolution. Science **282**:1711-1714.

Amores, A., T. Suzuki, Y. L. Yan, J. Pomeroy, A. Singer, C. Amemiya, and J. H. Postlethwait. 2004. Developmental roles of pufferfish Hox clusters and genome evolution in ray-fin fish. Genome Research **14**:1-10.

Bartsch, P., and R. Britz. 1997. A single micropyle in the eggs of the most basal living actinopterygian fish, Polypterus (Actinopterygii, Polypteriformes). Journal of Zoology **241**:589-592.

Bemis, W. E., E. K. Findeis, and L. Grande. 1997. An overview of Acipenseriformes. Environmental Biology of Fishes **48**:25-72.

Bruce, A. E., A. C. Oates, V. E. Prince, and R. K. Ho. 2001. Additional hox clusters in the zebrafish: divergent expression patterns belie equivalent activities of duplicate hoxB5 genes. Evol Dev **3**:127-144.

Bryant, D., and V. Moulton. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Molecular Biology and Evolution **21**:255-265.

Carroll, S. B. 1995. Homeotic genes and the evolution of arthropods and chordates. Nature **376**:479-485.

Chiang, E. F., C. I. Pai, M. Wyatt, Y. L. Yan, J. Postlethwait, and B. Chung. 2001. Two sox9 genes on duplicated zebrafish chromosomes: expression of similar transcription activators in distinct sites. Dev Biol **231**:149-163.

Chiu, C. H., C. Amemiya, K. Dewar, C. B. Kim, F. H. Ruddle, and G. P. Wagner. 2002. Molecular evolution of the HoxA cluster in the three major gnathostome lineages. Proceedings of the National Academy of Sciences of the United States of America **99**:5492-5497.

Chiu, C. H., K. Dewar, G. P. Wagner, K. Takahashi, F. Ruddle, C. Ledje, P. Bartsch, J. L. Scemama, E. Stellwag, C. Fried, S. J. Prohaska, P. F. Stadler, and C. T. Amemiya. 2004. Bichir HoxA cluster sequence reveals surprising trends in ray-finned fish genomic evolution. Genome Research **14**:11-17.

Conant, G. C., and A. Wagner. 2003. Asymmetric sequence divergence of duplicate genes. Genome Research **13**:2052-2058.

de Pinna, M. C. C. 1996. Teleostean monophyly *in* M. L. J. Stiassny, L. R. Parenti, and G. D. Johnson, eds. Interrelationships of fishes. Academic Press, San Diego.

Dobzhansky, T. 1936. Studies on hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. Genetics **21**:113-135.

Donoghue, P. C., and M. A. Purnell. 2005. Gene duplication, extinction, and vertebrate evolution. Trends in Ecology and Evolution **in press**.

Fares, M. A., D. Bezemer, A. Moya, and I. Marin. 2003. Selection on coding regions determined Hox7 genes evolution. Molecular Biology and Evolution **20**:2104-2112.

Felsenstein, J. 1985. Confidence-limits on phylogenies - an approach using the bootstrap. Evolution **39**:783-791.

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151**:1531-1545.

Fried, C., S. J. Prohaska, and P. F. Stadler. 2004. Exclusion of repetitive DNA elements from gnathostome Hox clusters. J Exp Zoolog B Mol Dev Evol **302**:165-173.

Furlong, R. F., and P. W. Holland. 2002. Were vertebrates octoploid? Philos Trans R Soc Lond B Biol Sci **357**:531-544.

Garciafernandez, J., and P. W. H. Holland. 1994. Archetypal Organization of the Amphioxus Hox Gene-Cluster. Nature **370**:563-566.

Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution **11**:725-736.

Gu, Z., L. M. Steinmetz, X. Gu, C. Scharfe, R. W. Davis, and W. H. Li. 2003. Role of duplicate genes in genetic robustness against null mutations. Nature **421**:63-66.

Hedges, S. B. 1992. The number of replications needed for accurate estimation of the bootstrap-p value in phylogenetic studies. Molecular Biology and Evolution **9**:366-369.

Hedges, S. B., and S. Kumar. 2003. Genomic clocks and evolutionary timescales. Trends in Genetics **19**:200-206.

Higgins, D. G., A. J. Bleasby, and R. Fuchs. 1992. CLUSTAL-V - Improved software for multiple sequence alignment. Computer applications in the biosciences **8**:189-191.

Hillis, D. M., and J. J. Bull. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Systematic Biology **42**:182-192.

Hoegg, S., H. Brinkmann, J. S. Taylor, and A. Meyer. 2004. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. Journal of Molecular Evolution **59**:190-203.

Holland, P. W., J. Garcia-Fernandez, N. A. Williams, and A. Sidow. 1994. Gene duplications and the origins of vertebrate development. Dev Suppl:125-133.

Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17**:754-755.

Hughes, A. L. 1999. Adaptive evolution of genes and genomes. Oxford University Press, New York.

Huson, D. H. 1998. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics **14**:68-73.

Inoue, J. G., M. Miya, K. Tsukamoto, and M. Nishida. 2003. Basal actinopterygian relationships: a mitogenomic perspective on the phylogeny of the "ancient fish". Molecular Phylogenetics and Evolution **26**:110-120.

Kao, H. W., and S. C. Lee. 2002. Phosphoglucose isomerases of hagfish, zebrafish, gray mullet, toad, and snake, with reference to the evolution of the genes in vertebrates. Molecular Biology and Evolution **19**:367-374.

Kikugawa, K., K. Katoh, S. Kuraku, H. Sakurai, O. Ishida, N. Iwabe, and T. Miyata. 2004. Basal jawed vertebrate phylogeny inferred from multiple nuclear DNA-coded genes. BMC Biol **2**:3.

Kondrashov, F. A., I. B. Rogozin, Y. I. Wolf, and E. V. Koonin. 2002. Selection in the evolution of gene duplications. Genome Biol **3**:RESEARCH0008.

Kosakovsky Pond, S. L., S. D. Frost, and S. V. Muse. 2004. HyPhy: hypothesis testing using phylogenies. Bioinformatics.

Kumar, S., and S. B. Hedges. 1998. A molecular timescale for vertebrate evolution. Nature **392**:917-920.

Larget, B., and D. L. Simon. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Molecular Biology and Evolution **16**:750-759.

Lister, J. A., J. Close, and D. W. Raible. 2001. Duplicate mitf genes in zebrafish: complementary expression and conservation of melanogenic potential. Dev Biol **237**:333-344.

Lynch, M., and J. S. Conery. 2000. The evolutionary fate and consequences of duplicate genes. Science **290**:1151-1155.

Lynch, M., and A. Force. 2000. The origin of interspecific genomic incompatibility via gene duplication. American Naturalist **156**:590-605.

Malaga-Trillo, E., and A. Meyer. 2001. Genome duplication and accelerated evolution of Hox genes and cluster architecture in teleost fishes. American Zoologist **41**:676-686.

Martinez, P., and C. T. Amemiya. 2002. Genomics of the HOX gene cluster. Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology **133**:571-580.

McGinnis, W., and R. Krumlauf. 1992. Homeobox Genes and Axial Patterning. Cell **68**:283-302.

Merrit, T. J., and M. Quattro. 2003. Evolution of the vertebrate cytosolic malate dehydrogenase gene family: duplication and divergence in actinopterygian fish. Journal of Molecular Evolution **56**:265-276.

Merritt, T. J., and J. M. Quattro. 2001. Evidence for a period of directional selection following gene duplication in a neurally expressed locus of triosephosphate isomerase. Genetics **159**:689-697.

Messier, W., and C. B. Stewart. 1997. Episodic adaptive evolution of primate lysozymes. Nature **385**:151-154.

Metscher, B. D., K. Takahashi, K. D. Crow, C. Amemiya, D. F. Nonaka, and G. P. Wagner. 2005. Expression of Hoxa-11 and Hoxa-13 in the pectoral fin of a basal ray finned fish, Polyodon spathula: implications for the origin of tetrapod limbs. Evolution and Development.

Meyer, A., and M. Schartl. 1999. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. Current Opinion in Cell Biology **11**:699-704.

Muller, H. J. 1942. Isolating mechanisms, evolution, and temperature. Biological Symposia **6**:71-125.

Naruse, K., S. Fukamachi, H. Mitani, M. Kondo, T. Matsuoka, S. Kondo, N. Hanamura, Y. Morita, K. Hasegawa, R. Nishigaki, A. Shimada, H. Wada, T. Kusakabe, N. Suzuki, M. Kinoshita, A. Kanamori, T. Terado, H. Kimura, M. Nonaka, and A. Shima. 2000. A detailed linkage map of Medaka, Oryzias latipes: Comparative genomics and genome evolution. Genetics **154**:1773-1784.

Nelson, G. J. 1969. Gill Arches and the phylogeny of fishes, with notes on the classification of vertebrates. Bull. Am. mus. Nat. Hist. **141**:475-552.

Nelson, J. S. 1994. Fishes of the World. J. Wiley, New York.

Nieselt-Struwe, K., and A. von Haeseler. 2001. Quartet-mapping, a generalization of the likelihood-mapping procedure. Molecular Biology and Evolution **18**:1204-1219.

Nylander, J. A. A. 2002. MrModeltest. Program distributed by author, Department of Systematic Zoology, Uppsala University.

Ohno, S. 1970. Evolution by Gene Duplication. Springer-Verlag, New York.

Ohno, S. 1999. Gene duplication and the uniqueness of vertebrate genomes circa 1970-1999. Semin Cell Dev Biol **10**:517-522.

Olsen, P. E. 1984. The skull and pectoral girdle of the parasemionotid fish *Watsonulus eugnathoides* from the early Triassic Sakamena Group of Madagascar, with comments on the relationships of the holostean fishes. Journal of Vertebrate Paleontology **4**:481-499.

Orr, H. A. 1995. The population genetics of speciation: the evolution of hybrid incompatibilities. Genetics **139**:1805-1813.

Orr, H. A., J. P. Masly, and D. C. Presgraves. 2004. Speciation genes. Curr Opin Genet Dev **14**:675-679.

Patterson, C. 1973. Interrelationships of holosteans. In: Greenwood, H., Miles, R.S., Patterson, C. (eds.) Interrelationships of fishes. Zool J. Linn Soc Lond (Suppl 1):233-305.

Postlethwait, J., A. Amores, A. Force, and Y. L. Yan. 1999. The zebrafish genome. Pp. 149-163. Methods in Cell Biology, Vol 60.

Prohaska, S. J., C. Fried, C. Flamm, G. P. Wagner, and P. F. Stadler. 2004. Surveying phylogenetic footprints in large gene clusters: applications to Hox cluster duplications. Molecular Phylogenetics and Evolution **31**:581-604.

Prohaska, S. J., and P. F. Stadler. 2004. The duplication of the Hox gene clusters in teleost fishes. Theory in Biosciences **123**:89-110.

Robinson-Rechavi, M., B. Boussau, and V. Laudet. 2004. Phylogenetic dating and characterization of gene duplications in vertebrates: the cartilaginous fish reference. Molecular Biology and Evolution **21**:580-586.

Robinson-Rechavi, M., O. Marchand, H. Escriva, P. L. Bardet, D. Zelus, S. Hughes, and V. Laudet. 2001. Euteleost fish genomes are characterized by expansion of gene families. Genome Research **11**:781-788.

Ruddle, F. H., J. L. Bartels, K. L. Bentley, C. Kappen, M. T. Murtha, and J. W. Pendleton. 1994. Evolution of Hox genes. Annu Rev Genet **28**:423-442.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution **4**:406-425.

Schubert, F. R., K. Nieseltstruwe, and P. Gruss. 1993. The Antennapedia-Type Homeobox Genes Have Evolved from 3 Precursors Separated Early in Metazoan Evolution. Proceedings of the National Academy of Sciences of the United States of America **90**:143-147.

Schultze, H. P., and E. O. Wiley. 1984. The neopterygian *Amia* as a living fossil. Springer-Verlag, New York.

Stadler, P. F., C. Fried, S. J. Prohaska, W. J. Bailey, B. Y. Misof, F. H. Ruddle, and G. P. Wagner. 2004. Evidence for independent Hox gene duplications in the hagfish lineage: a PCR-based gene inventory of Eptatretus stoutii. Molecular Phylogenetics and Evolution **32**:686-694.

Swofford, D. L. 1998. PAUP* 4.0 - Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Assoc., Sunderland, MA.

Taylor, J. S., I. Braasch, T. Frickey, A. Meyer, and Y. V. de Peer. 2003. Genome duplication, a trait shared by 22,000 species of ray-finned fish. Genome Research **13**:382-390.

Taylor, J. S., Y. Van de Peer, I. Braasch, and A. Meyer. 2001. Comparative genomics provides evidence for an ancient genome duplication event in fish. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences **356**:1661-1679.

Taylor, J. S., Y. Van de Peer, and A. Meyer. 2001. Genome duplication, divergent resolution and speciation. Trends in Genetics **17**:299-301.

Van de Peer, Y., J. S. Taylor, I. Braasch, and A. Meyer. 2001. The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. Journal of Molecular Evolution **53**:436-446.

Van de Peer, Y., J. S. Taylor, and A. Meyer. 2003. Are all fishes ancient polyploids? J Struct Funct Genomics **3**:65-73.

Vandepoele, K., W. De Vos, J. S. Taylor, A. Meyer, and Y. Van de Peer. 2004. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. Proc Natl Acad Sci U S A **101**:1638-1643.

Venkatesh, B., M. V. Erdmann, and S. Brenner. 2001. Molecular synapomorphies resolve evolutionary relationships of extant jawed vertebrates. Proc Natl Acad Sci U S A **98**:11382-11387.

Wagner, A. 2001. Birth and death of duplicated genes in completely sequenced eukaryotes. Trends in Genetics **17**:237-239.

Wagner, G. P., C. Amemiya, and F. Ruddle. 2003. Hox cluster duplications and the opportunity for evolutionary novelties. Proceedings of the National Academy of Sciences of the United States of America **100**:14603-14606.

Wagner, G. P., K. Takahashi, V. Lynch, S. J. Prohaska, C. Fried, P. F. Stadler, and C. Amemiya. 2005. Molecular Evolution of duplicated ray finned fish HoxA clusters: Increased synonymous substitution rate and asymmetrical co-divergence of coding and non-coding sequences. Journal of Molecular Evolution ***in press***.

Werth, C. R., and M. D. Windham. 1991. A Model for Divergent, Allopatric Speciation of Polyploid Pteridophytes Resulting from Silencing of Duplicate-Gene Expression. The American Naturalist **137**:515-526.

Wiley, E. O., and H. P. Schultze. 1984. Family Lepisosteidat (gars) as living fossils. Springer-Verlag, New York.

Winkler, C., M. Schafer, J. Duschl, M. Schartl, and J. N. Volff. 2003. Functional divergence of two zebrafish midkine growth factors following fish-specific gene duplication. Genome Research **13**:1067-1081.

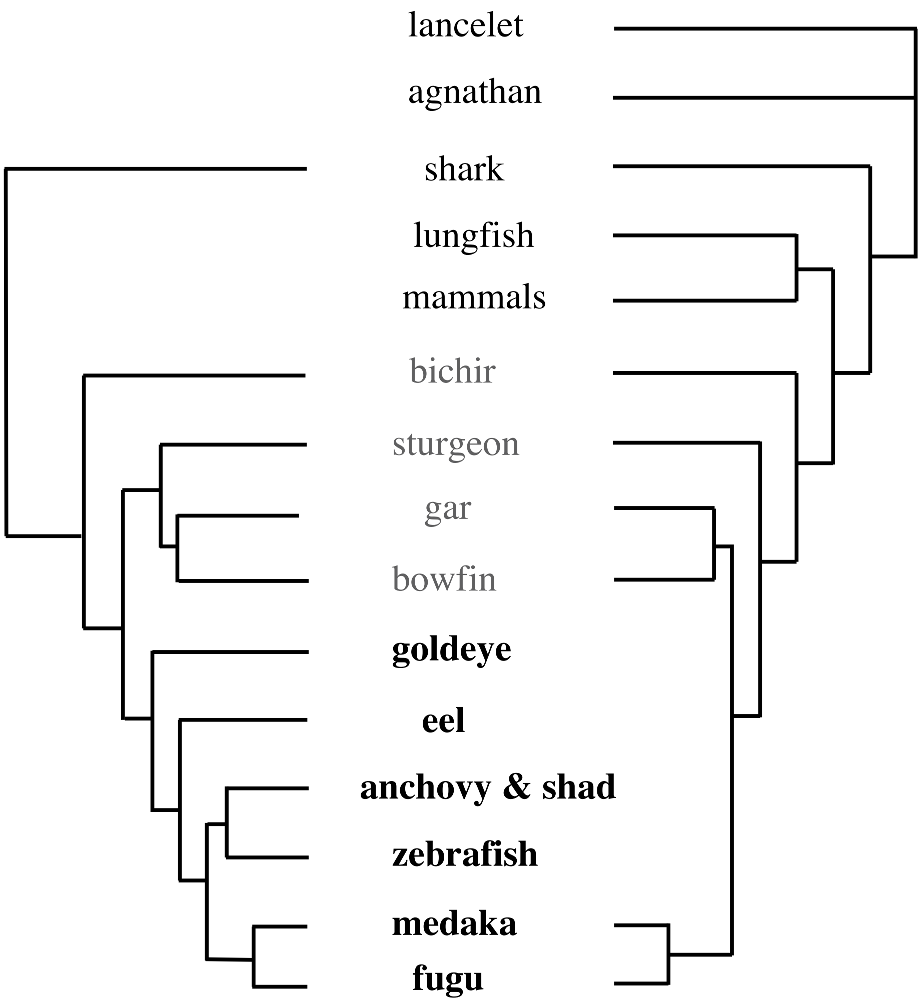
Wittbrodt, J., A. Meyer, and M. Schartl. 1998. More genes in fish? Bioessays **20**:511-515.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci **13**:555-556.

Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Molecular Biology and Evolution **15**:568-573.

Zhou, R. J., H. H. Cheng, and T. R. Tiersch. 2001. Differential genome duplication and fish diversity. Reviews in Fish Biology and Fisheries **11**:331-337.

Figure 1.



1a. Inoue *et al*. 2003

1b. Kikugawa *et al*. 2004

1a. Jun G. Inoue**,** Masaki Miya, Katsumi Tsukamoto and Mutsumi Nishida. 2003. Basal actinopterygian relationships: a mitogenomic perspective on the phylogeny of the "ancient fish". *Molecular Phylogenetics and Evolution*. 26(1)110-120

1b. Kikugawa, Kanae, Kazutaka Katoh, Shigehiro Kuraku, Hiroshi Sakurai, Osamu Ishida, Naoyuki Iwabe, and Takashi Miyata. 2004. Basal jawed vertebrate phylogeny inferred from multiple nuclear DNA-coded genes. *BMC Biology* 2:3

Figure 2.

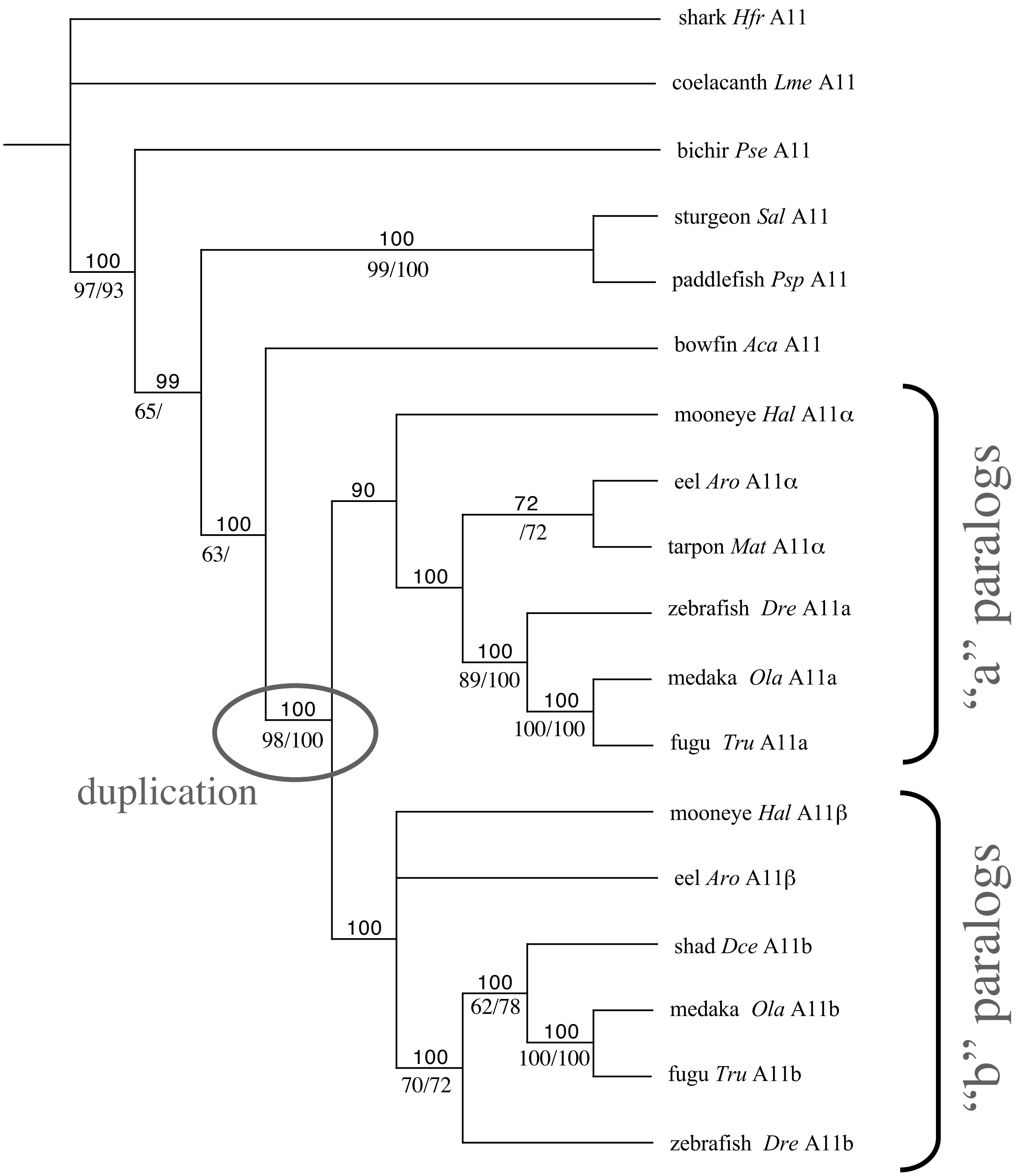

shark Hfr A11
coelacanth Lme A11
bichir Pse A11
sturgeon Sal A11
paddlefish Psp A11
bowfin Aca A11
mooneye Hal A11α
eel Aro A11α
tarpon Mat A11α
zebrafish Dre A11a
medaka Ola A11a
fugu Tru A11a
mooneye Hal A11β
eel Aro A11β
shad Dce A11b
medaka Ola A11b
fugu Tru A11b
zebrafish Dre A11b
100
97/93
99
65/
100
99/100
100
63/
100
98/100
duplication
90
72
/72
100
100
89/100
100
100/100
100
100
70/72
100
62/78
100
100/100
“a” paralogs
“b” paralogs

Figure 3.

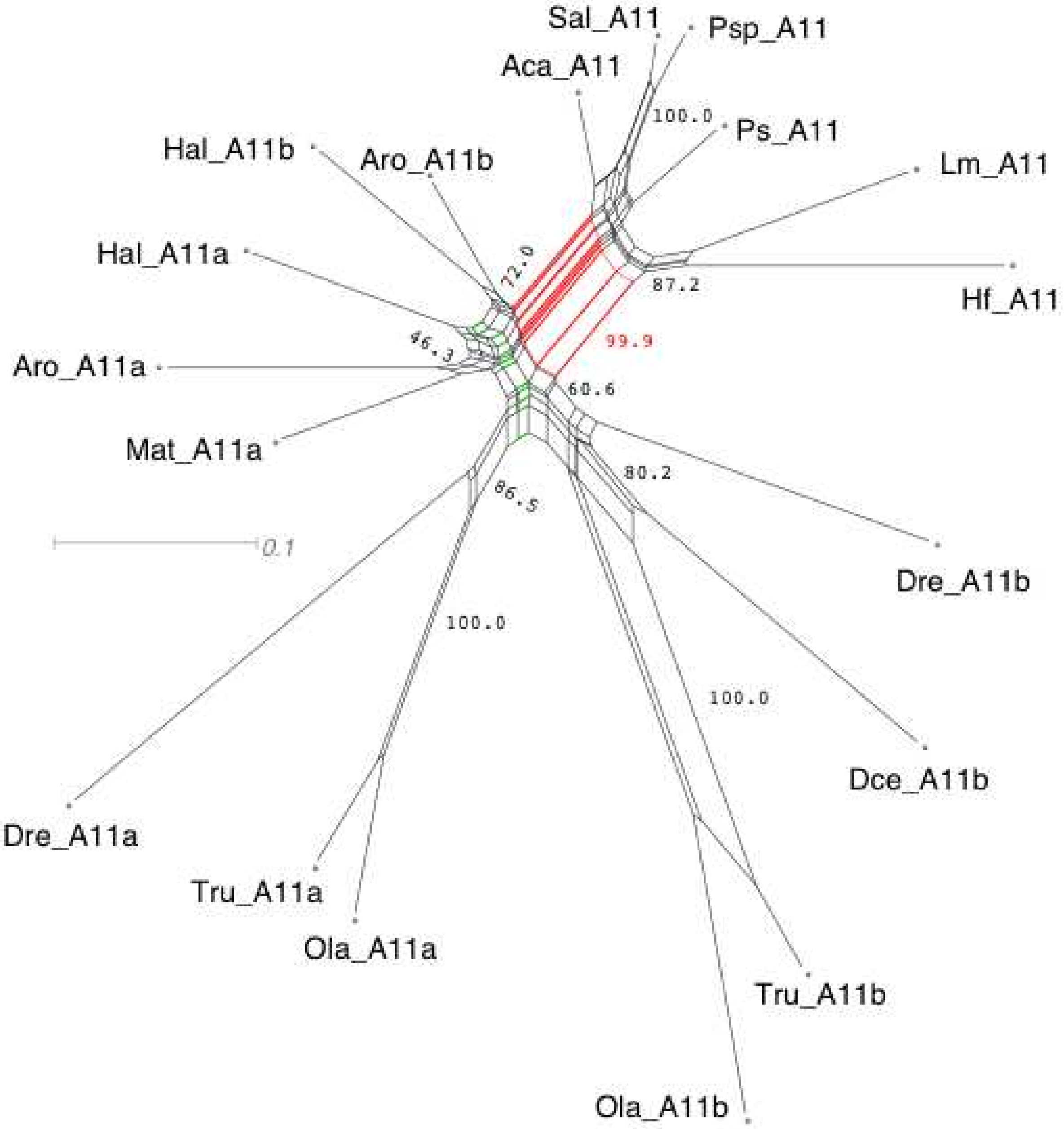

Sal_A11
Psp_A11
Aca_A11
100.0
Ps_A11
Hal_A11b
Aro_A11b
Lm_A11
Hal_A11a
72.0
87.2
Hf_A11
99.9
46.3
Aro_A11a
60.6
Mat_A11a
80.2
86.5
0.1
Dre_A11b
100.0
100.0
Dce_A11b
Dre_A11a
Tru_A11a
Ola_A11a
Tru_A11b
Ola_A11b

Figure 4

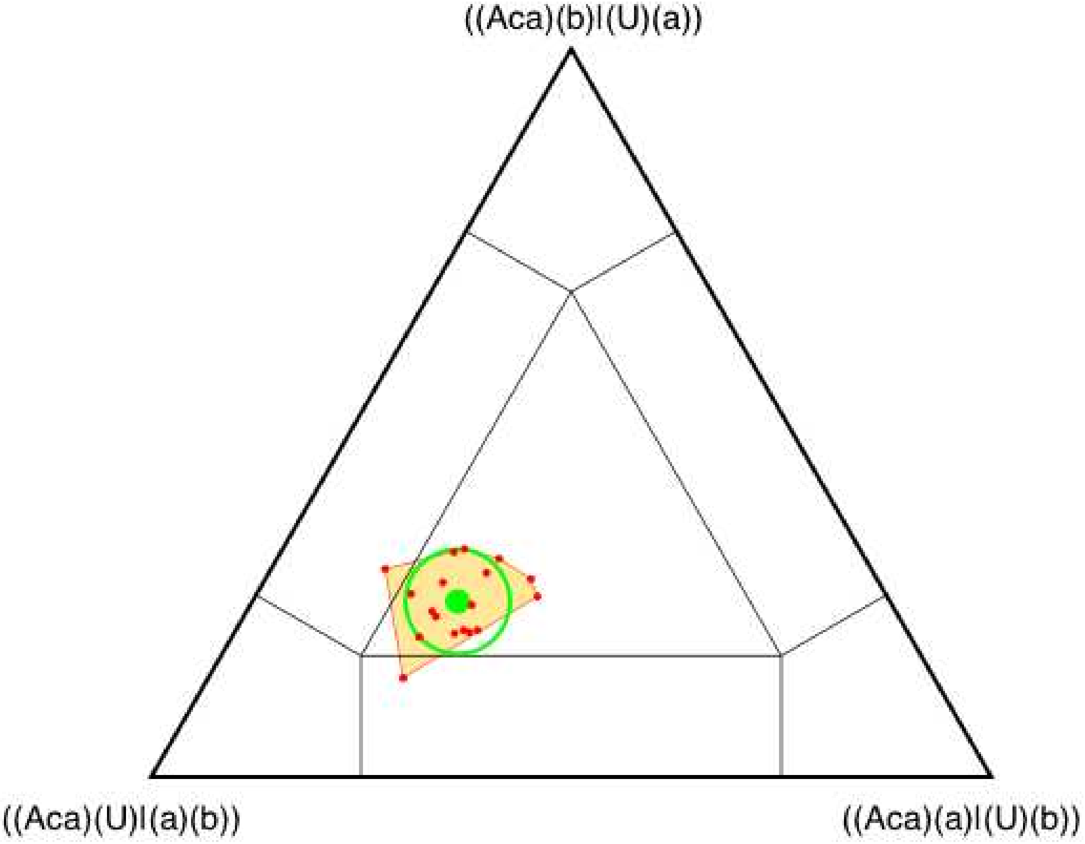


Figure 5.

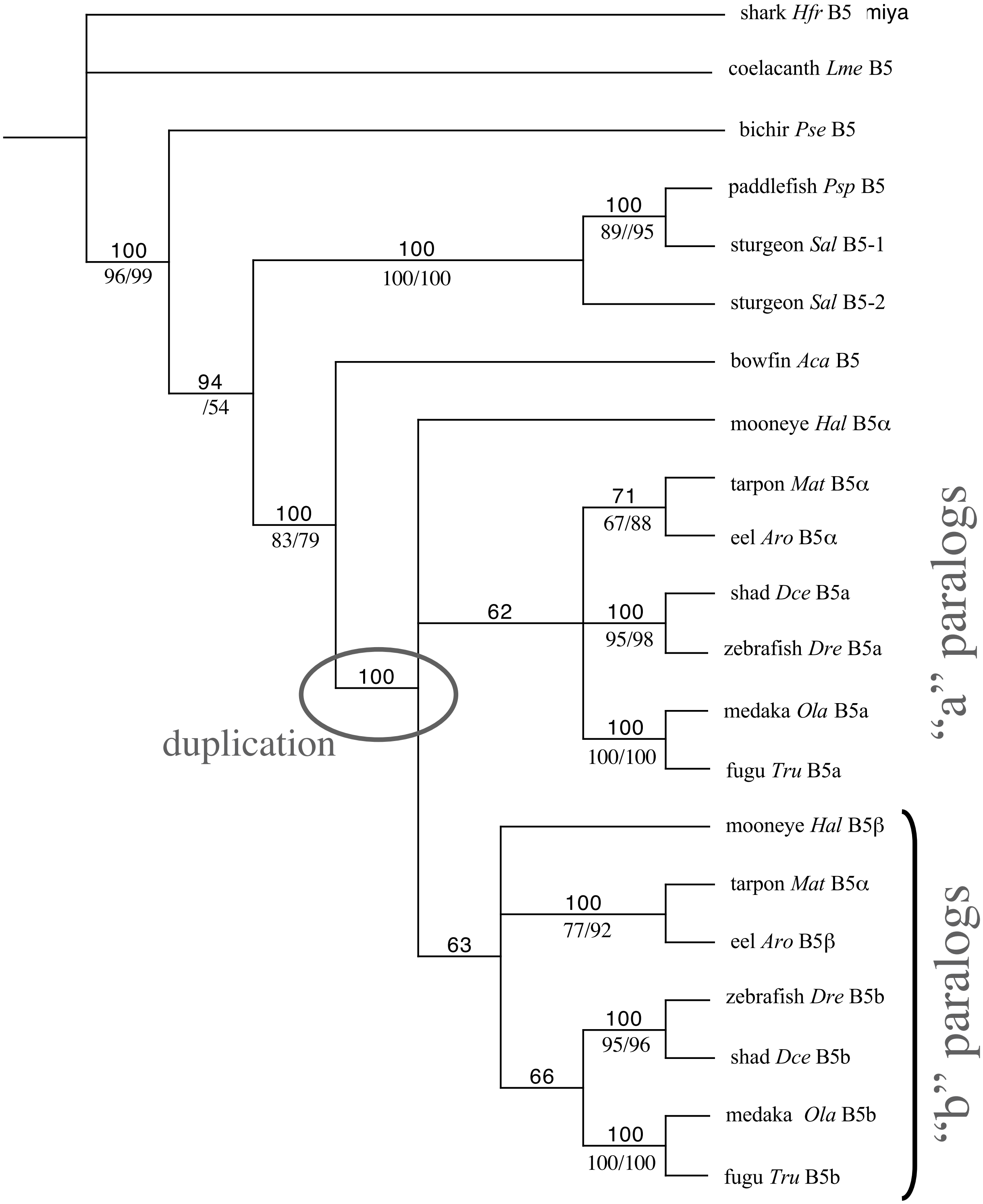
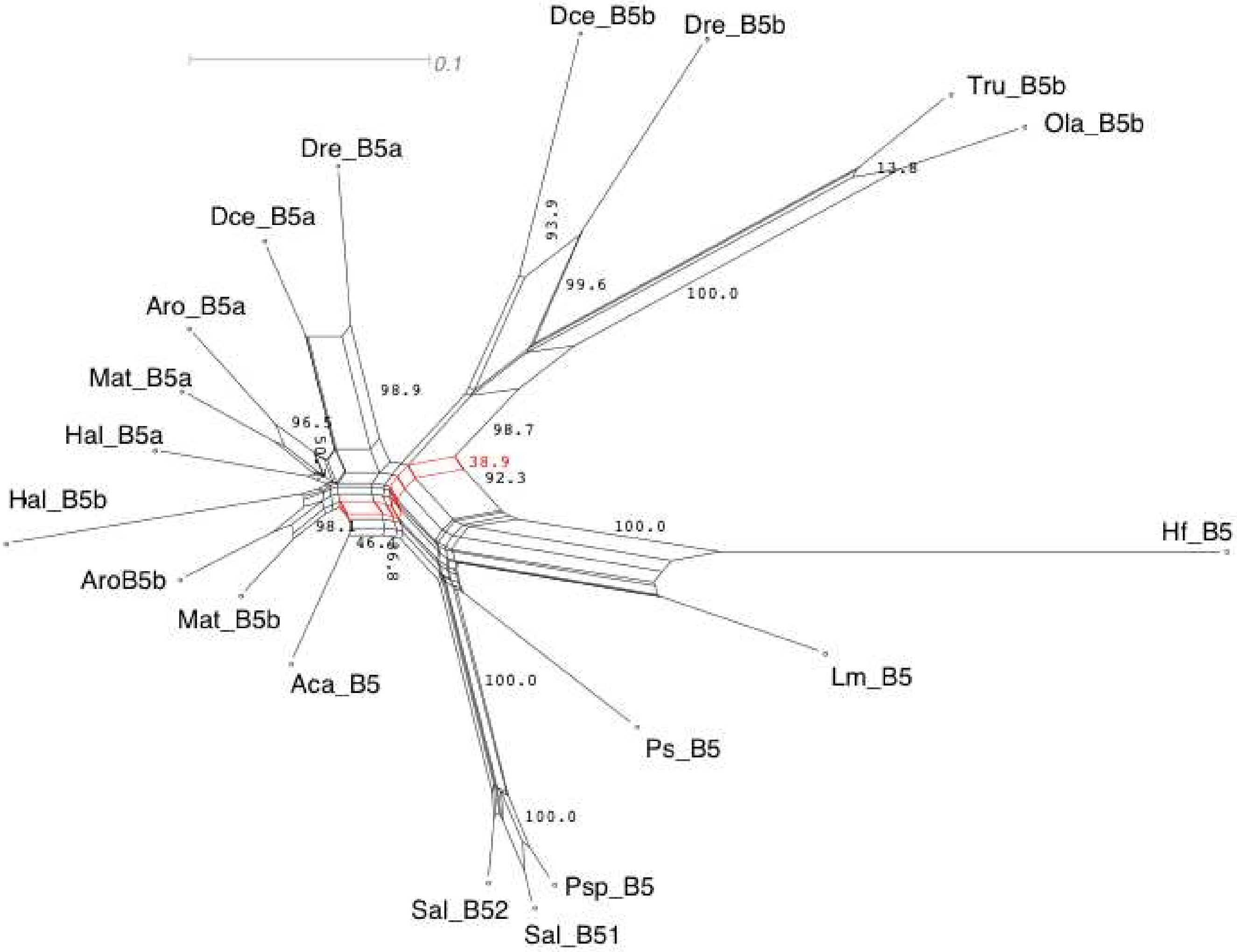

shark *Hfr* B5 miya
coelacanth *Lme* B5
bichir *Pse* B5
paddlefish *Psp* B5
sturgeon *Sal* B5-1
sturgeon *Sal* B5-2
bowfin *Aca* B5
mooneye *Hal* B5α
tarpon *Mat* B5α
eel *Aro* B5α
shad *Dce* B5a
zebrafish *Dre* B5a
medaka *Ola* B5a
fugu *Tru* B5a
mooneye *Hal* B5β
tarpon *Mat* B5α
eel *Aro* B5β
zebrafish *Dre* B5b
shad *Dce* B5b
medaka *Ola* B5b
fugu *Tru* B5b
"a" paralogs
"b" paralogs
duplication
100
96/99
94
/54
100
100/100
100
89//95
100
83/79
100
62
71
67/88
100
95/98
100
100/100
63
100
77/92
66
100
95/96
100
100/100

Figure 6.


0.1
Dce_B5b
Dre_B5b
Tru_B5b
Ola_B5b
13.8
Dre_B5a
Dce_B5a
93.9
99.6
100.0
Aro_B5a
Mat_B5a
98.9
96.5
50.
Hal_B5a
98.7
38.9
92.3
Hal_B5b
100.0
Hf_B5
98.1
46.6
96.8
AroB5b
Mat_B5b
Aca_B5
100.0
Lm_B5
Ps_B5
100.0
Psp_B5
Sal_B52
Sal_B51