2

Duplicated Gene Evolution Following Whole-Genome Duplication in Teleost Fish

Baocheng Guo[1,2,3], Andreas Wagner[1,2] and Shunping He[3*]
[1] *Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich*
[2] *The Swiss Institute of Bioinformatics, Quartier Sorge-Batiment Genopode, Lausanne*
[3]*Fish Phylogenetics and Biogeography Group, Key Laboratory of Aquatic Biodiversity and Conservation, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan*
[1,2]*Switzerland*
[3]*PR China*

1. Introduction

Gene and genome duplication have been thought to play an import part during evolution since the 1930s (Bridges 1936; Stephens 1951; Ohno 1970) . Ohno (1970) proposed that the increased complexity and genome size of vertebrates has resulted from two rounds (2R) of whole genome duplication (WGD) in early vertebrate evolution, which provided raw materials for the evolutionary diversification of vertebrates. Recent genomic sequence data provide substantial evidence for the abundance of duplicated genes in many organisms. Extensive comparative genomics studies have demonstrated that teleost fish experienced another round of genome duplication, the so-called fish-specific genome duplication (FSGD) (Amores et al. 1998; Taylor et al. 2003; Meyer and Van de Peer 2005). Because the timing of this WGD and the radiation of teleost species approximately coincided, it has been suggested that the large number (about 27,000 species—more than half of all vertebrate species (Nelson, 2006)) of teleosts and their tremendous morphological diversity might be causally related to the FSGD event (Amores et al. 1998; Taylor et al. 2001; Taylor et al. 2003; Christoffels et al. 2004; Hoegg et al. 2004; Vandepoele et al. 2004). Semon and Wolfe (2007) showed thousands of genes that remained duplicated When Tetraodon and zebrafish diverged underwent reciprocal loss subsequently in these two species may have been associated with reproductive isolation between teleosts and eventually contributed to teleost diversification. A study in yeast demonstrated that speciation of polyploid yeasts may be associated with reciprocal gene loss at duplicated loci (Scannell et al. 2006). Thus, speciation accompanied by differential retention and loss of duplicated genes after genome duplication may be a powerful lineage-splitting force (Lynch and Conery 2000).
For two reasons, teleost fish represent an excellent model system to study the retention and loss of duplicated genes as well as their evolutionary trajectory following whole-genome

* Corresponding author

duplication. First, many duplicated genes that resulted from the FSGD event were preserved in teleost genomes. Second, five teleost genomes have been sequenced and more teleost genomes are being sequenced. Here, we investigate retention, loss, and molecular evolution of duplicate genes after the FSGD in five available teleost geomes that include the genomes of zebrafish *Danio rerio*, stickleback *Gasterosteus aculeatus*, medaka *Oryzias latipes*, Takifugu *Takifugu rubripes*, and Tetraodon *Tetraodon nigroviridis*.

2. Identifying duplicated genes that resulted from the FSGD event throughout the teleost genomes

We obtained 23,155 gene families from the database HOMOLENS version 4 (ftp://pbil.univ-lyon1.fr/databases/homolens4.php) (Penel et al. 2009), which is based on the Ensembl release 49. We chose HOMOLENS, because it allowed us to reliably retrieve sets of orthologous genes for our evolutionary analysis. HOMOLENS is devoted to metazoan genomes from Ensembl and contains gene families from complete animal genomes found in Ensembl. HOMOLENS has the same architecture as HOVERGEN (Duret et al. 1994), in which genes are organized in families and include precalculated alignments and phylogenies. In HOMOLENS 4, alignments are computed using MUSCLE (Edgar 2004) with default parameters; phylogenetic trees are computed with PHYML, using the JTT amino acid substitution model (Jones et al. 1992). Phylogenies are computed based on conserved blocks of the alignments selected with Gblocks (Castresana 2000). Each phylogenetic tree is reconciled with a species tree using the program RAP (Dufayard et al. 2005), which, combined with the tree pattern search functionality, allows detection of ancient gene duplications or selection of orthologous genes (Penel et al. 2009). Several studies on duplicated gene evolution have been performed with data retrieved from HOMOLENS (Brunet et al. 2006; Studer et al. 2008).

We employed a topology-based method to identify duplicated genes that resulted from the FSGD event in the five teleost genomes we study. Briefly, if two teleosts have been subject to the same whole genome duplication event, a gene *X* that has been duplicated in this event and retained in both genomes, should form two gene lineages ''*Xa*'' and ''*Xb*'' (Figure 1A). We identified gene trees with the topology shown in Figure 1A using the TreePattern functionality (Dufayard et al. 2005) of the FamFetch client for HOMOLENS. We required duplicated genes to exist in at least two species to increase the likelihood that they result from the FSGD event (Figure 1B). In total, we identified 1,500 gene families with duplicated genes in this way.

3. Differential retention and loss of duplicated genes during teleost diversification

The most common fate of a duplicated gene is nonfunctionalization (pseudogenization). After a whole genome duplication event, many genes share this fate, so that a genome's gene content may only appear be slightly increased long after the duplication (Wolfe and Shields 1997; Jaillon et al. 2004). Our data suggest that only 3.3 percent (zebrafish) to 7.2 percent (Takifugu) of genes in current teleost genomes result from the FSGD event (Table 1). These percentages are lower than the 13 percent of retained duplicates in yeast (Wolfe and Shields 1997). One possible reason for this difference might lie in our topology-based method to identify likely FSGD duplicates (Figure 1), which enforces duplicated genes to exist in at least

two teleost genomes. Thus, our method would overlook duplicated genes that result from the FSGD and that are retained in only one teleost genome. While we cannot exclude this possibility, we note that our observations are consistent with a genome-wide study of Tetraodon, in which Jaillon et al. (2004) showed that up to 3 percent of duplicated genes may have been retained since the FSGD event. One plausible explanation of the difference in duplicated gene retention between teleost and yeast may come from the different ages of the genome duplication event. In addition, Kassahn et al. (2009) suggested that a minimum of 3 to 4 percent of protein-coding loci have been retained in two copies in each of the five model

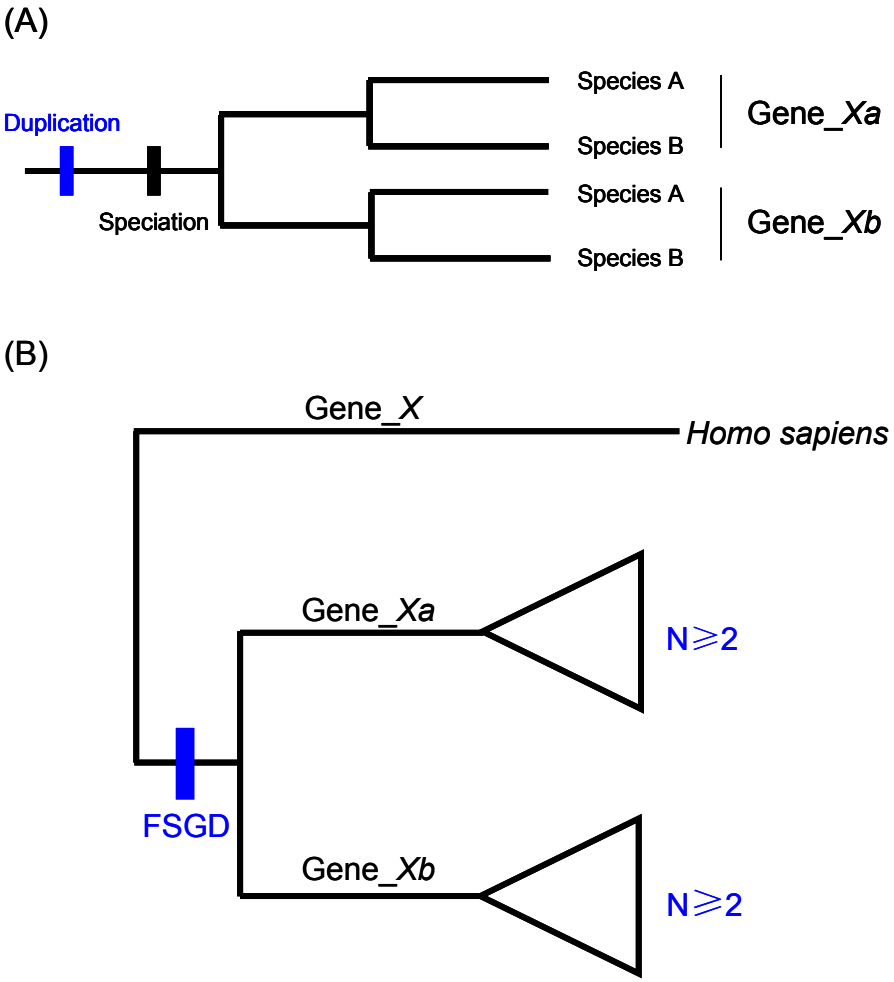

Fig. 1. (A) Expected phylogenetic relationship of duplicated gene *Xa* and *Xb* in two related species A and B when speciation occurred after the duplication event; (B) Tree topology we used for duplicated gene identification in the database HOMLENS 4. 'N ≥ 2' means that duplicated gene pairs must exist at least in two species to increase the likelihood that the duplicated genes actually resulted from the FSGD event.

	Number of genes *	Gene families with likely FSGD duplicates		
		FSGD Duplicates	Singleton	Double loss
D. rerio	21,420	731	541	228
G. aculeatus	20,839	681	669	150
O.latipes	19,687	1,162	311	27
T. rubripes	18,709	1,340	148	12
Te. nigroviridis	27,991	1,047	397	56

* Total gene number in each genome, data based on the Ensembl release 49.

Table 1. Summary of different gene retention and loss in the 1,500 duplicated gene families we identified.

fish genomes. The FSGD occurred between 253 and 404 Million years ago (MYA) (Hoegg et al. 2004; Vandepoele et al. 2004), whereas the yeast whole genome duplication may have occurred more recently, between 100 and 150 MYA (Sugino and Innan 2005). More time has elapsed since the FSGD, allowing more duplicate genes to be lost.

Differential retention and loss of duplicated genes is a common phenomenon during speciation after genome duplication. It has been observed in yeast (Scannell et al. 2006) as well as in teleosts (Semon and Wolfe 2007), and is believed to lead to speciation. We thus expected that our dataset would contain many gene families with differential gene retention and loss, as well as fewer families where both copies are retained in all five teleost genomes. Indeed, when we consider all five species together, we observed that 90.4 percent of the 1,500 gene families we identified show differential retention and loss of duplicated genes, and in only 9.6 percent (144 gene families) are both copies retained in all five teleost genomes. Figure 2 and Table 1 show relevant data, broken down by study species. In 45.4 percent to 89.3 percent (depending on the species) of the 1,500 gene families we identified, both duplicates were retained. In 9.9 percent to 44.6 percent of the duplicates (depending on the species), one copy was lost. Our data also indicate that differences in differential gene retention are associated with the phylogenetic position and the relatedness between two teleost species (Figure 2). Taken together, these observations indicate that differential duplicated gene retention and loss are pervasive in teleosts, that the loss of duplicated genes is an ongoing process that has continued for hundreds of million years after the FSGD event, and that this process may be associated with teleost diversification.

We next discuss an illustrative example of differential duplicate gene retention and loss. It involves *Hox* genes, which encode a subclass of homeodomain transcription factors that help determine the anterior–posterior axis of bilaterian animals (McGinnis and Krumlauf 1992). In vertebrates, *Hox* genes have evolved a highly compact organization, where genes are arranged in clusters on chromosomes. *Hox* gene clusters are one of the best-studied systems for assessing gene retention and loss after the FSGD event (Amores et al. 1998; Prohaska and Stadler 2004; Hoegg et al. 2007; Guo et al. 2010), due to their genomic architecture and gene complement variation in teleosts. Seven or eight *Hox* clusters with different complements of

Hox genes exist in extant diploid teleosts. They are a result of the FSGD event, which was followed by loss of some *Hox* gene duplicates. The putative *Hox* cluster complement of the teleost ancestor and the *Hox* clusters of several model teleost species are shown in Figure 3. *Hox* clusters exhibit remarkably different gene complements in different teleost lineages after the FSGD event. Theoretically, 8 *Hox* clusters containing at least 80 *Hox* genes genes may have existed in the ancestor of teleosts after the FSGD event. Up to now, 66 of these *Hox* genes have been found in different teleost species and extant evolutionary diploid teleost usually have 45 to 49 *Hox* genes in their genome (Figure 3). According to the summary of Hoegg et al (2007) (Figure 3), the Ostariophysii have lost seven *Hox* genes since their hypothetical common ancestor with the Neoteleosts; during the evolution of the Neoteleosts eight *Hox* genes were lost; and the pufferfish lineage lost three genes in the common lineage leading to Takifugu and Tetraodon. Some *Hox* genes are specifically preserved in different teleosts, for example, *HoxA1b* has been identified thus far only in the Japanese eel (Guo et al. 2010). At the cluster level, eight *Hox* clusters were retained in basal species such as the Japanese eel (Guo et al. 2010) and the goldeye (Chambers et al. 2009), whereas one *Hox* cluster (C or D) was lost respectively in the Otocephala (Amores et al. 1998) and Euteleostei (Kurosawa et al. 2006). Based on the phylogeny of teleosts, Guo et al. (2010) proposed that the *HoxDb* cluster

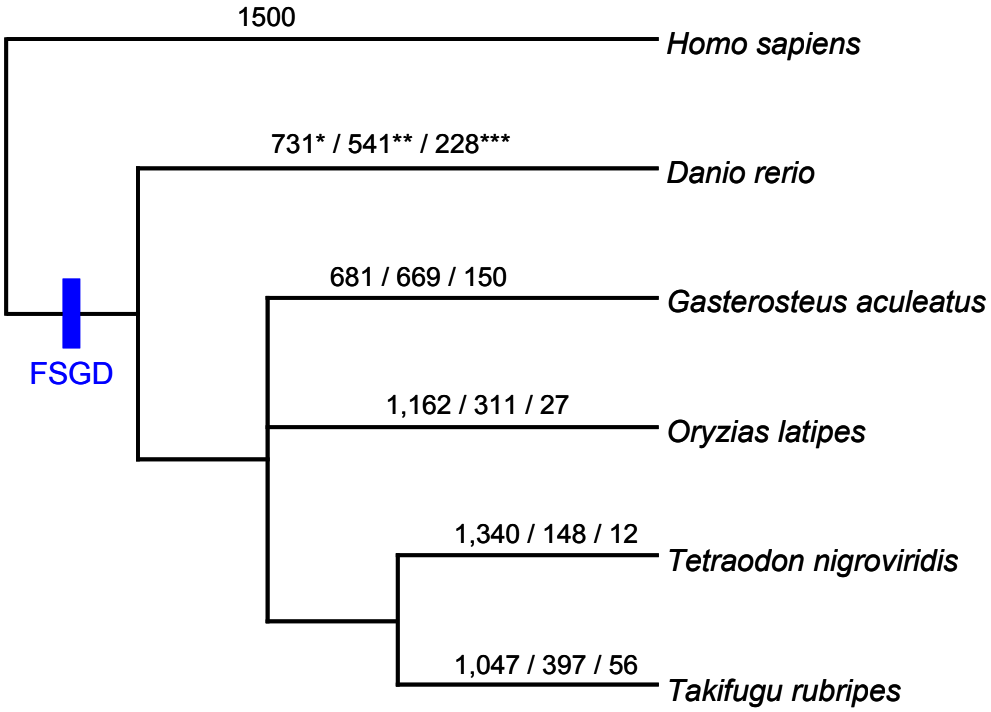

Fig. 2. Differential retention and loss of duplicated genes during teleost diversification. The topology is adopted from (Negrisolo et al. 2010). *: retention of both copies; **: retention of one copy; ***: loss of both copies.

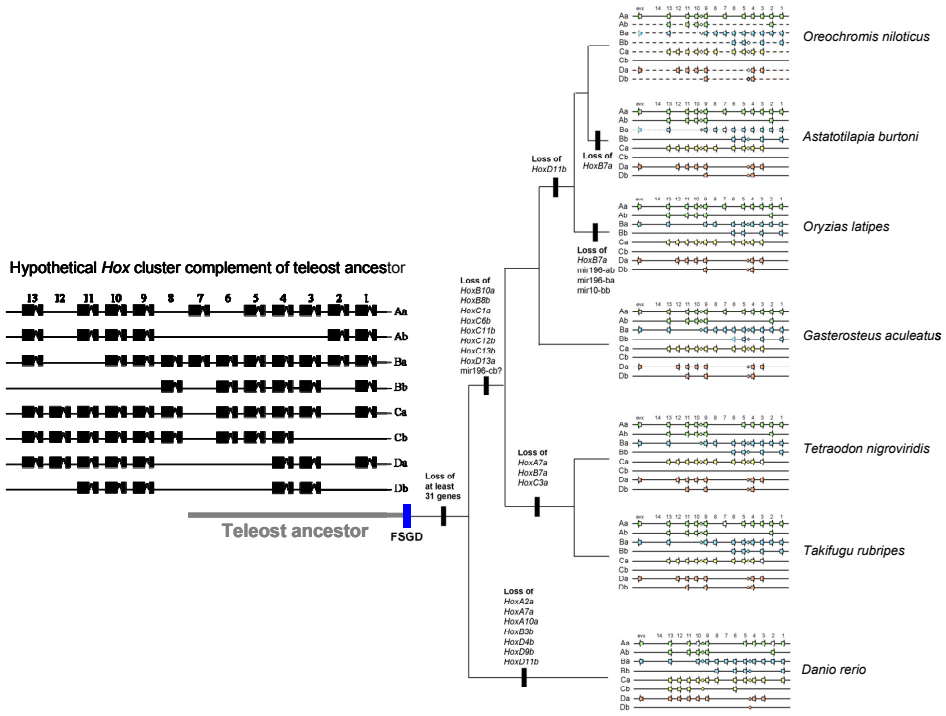

Fig. 3. *Hox* gene clusters, the best-studied examples of differential duplicate gene retention and loss in teleosts. Hypothetical *Hox* clusters of the teleost ancestor (modified from Guo et al. 2010), and *Hox* clusters of teleost model fish species, together with specific gene loss events shown on a phylogenetic tree of select fish species (adapted form Hoegg et al. 2007).

was lost independently in the Otocephala and Euteleostei after the FSGD event. The ongoing process of *Hox* gene loss and retention in teleosts illustrates again that degeneration of functionally important duplicated genes can last for hundreds of millions of years after the FSGD event.

4. Molecular evolution of duplicated genes

We next wished to study patterns of sequence evolution in the 1,500 duplicate gene families we had identified. To this end, we downloaded both nucleic acid and amino acid sequences for genes in these families. For each species, we retained only one gene copy in each duplicated clade (Figure 1B) for further analysis, and discarded all other copies in those gene families where additional duplications have occurred after the FSGD event. We then aligned the amino acid sequences within each gene family with MUSCLE (Edgar

2004), and calculated DNA alignments from protein alignments with RevTrans (Wernersson and Pedersen 2003). The following computations were then done on the new DNA alignments. We estimated the nucleic acid evolutionary distance between fish genes and their human orthologs using the LogDet nucleotide substitution model (Tamura and Kumar 2002) in PHYLIP-3.6b (Felsenstein 2004).

Previous studies show that duplicated genes in yeast often diverge asymmetrically (Kellis et al. 2004), meaning that one copy evolves significantly faster than the other. We asked whether this is also the case for teleost duplicates. To this end, we compared evolutionary distances of duplicated genes with their human orthologs within the 1,500 gene families we had identified. There is indeed evidence for asymmetric evolution between duplicated gene pairs from the FSGD event (Table 2). Average evolutionary distances to the human homologue between members of duplicated gene pairs are significantly different for each of our five teleost species (paired *t*-test: $P < 4.8 \times 10^{-95}$). As all duplicated gene pairs stemming from the FSGD diverged at the same time from their human orthologs, we can directly convert differences between evolutionary distances into differences between evolutionary rates. Taken together, our observations suggest that duplicate genes tend not to accumulate sequence change at the same rate. Our results are consistent with previous works in teleosts (Brunet et al. 2006; Steinke et al. 2006) and yeast (Kellis et al. 2004), and confirm that asymmetric sequence evolution between duplicated genes is a frequent pattern of duplicated gene evolution after a genome duplication event.

	D. rerio	*G. aculeatus*	*O.latipes*	*T. rubripes*	*Te. nigroviridis*
Duplicate_L	0.613 ± 0.243	0.607 ± 0.229	0.621 ± 0.230	0.623 ± 0.229	0.614 ± 0.224
Duplicate_S	0.529 ± 0.213	0.526 ± 0.200	0.536 ± 0.195	0.535 ± 0.195	0.505 ± 0.182
P-value*	4.1×10^{-105}	4.8×10^{-95}	1.9×10^{-165}	7.3×10^{-175}	8.9×10^{-133}

Duplicate_L: duplicated gene in each duplicate pair that has the larger distance to the human orthologue (distances averaged over all duplicate gene families); Duplicate_S: duplicated gene in each duplicate pair that has the smaller distance to the human orthologue (distances averaged over all duplicate gene families). All means are ± one standard deviation.
* paired *t*-test

Table 2. Average evolutionary distances of duplicated genes in five teleost species to their human orthologs.

5. Conclusion

In summary, we used a phylogenetic method to identify 1,500 duplicated gene families in five teleost species that are likely to have resulted from the FSGD event. Only a small fraction of genes in extant teleost genomes have been retained in the FSGD event. Differential retention and loss of duplicated gene is pervasive in the five species we studied, as is illustrated by genes in the teleost *Hox* gene clusters. Sequence analysis suggests that some duplicated genes pairs may evolve asymmetrically. Our work provides a framework for future studies of the evolutionary trajectory of duplicated genes in the teleost genome.

6. Acknowledgement

Support was provided by the National Natural Science Foundation of China (NSFC) to Shunping He.

7. References

Amores A, Force A, Yan YL, et al. 1998. Zebrafish hox clusters and vertebrate genome evolution. *Science* 282:1711-1714.

Bridges CB. 1936. The bar "gene" a duplication. *Science* 83:210-211.

Brunet FG, Crollius HR, Paris M, Aury JM, Gibert P, Jaillon O, Laudet V, Robinson-Rechavi M. 2006. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol* 23:1808-1816.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17:540-552.

Chambers KE, McDaniell R, Raincrow JD, Deshmukh M, Stadler PF, Chiu CH. 2009. Hox cluster duplication in the basal teleost Hiodon alosoides (Osteoglossomorpha). *Theory Biosci* 128:109-120.

Christoffels A, Koh EG, Chia JM, Brenner S, Aparicio S, Venkatesh B. 2004. Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol Biol Evol* 21:1146-1151.

Dufayard JF, Duret L, Penel S, Gouy M, Rechenmann F, Perriere G. 2005. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* 21: 2596-2603.

Duret L, Mouchiroud D, Gouy M. 1994. HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res* 22:2360-2365.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792-1797.

Felsenstein J. 2004. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Guo B, Gan X, He S. 2010. Hox genes of the Japanese eel Anguilla japonica and Hox cluster evolution in teleosts. *J Exp Zool B Mol Dev Evol* 314:135-147.

Hoegg S, Boore JL, Kuehl JV, Meyer A. 2007. Comparative phylogenomic analyses of teleost fish Hox gene clusters: lessons from the cichlid fish Astatotilapia burtoni. *BMC Genomics* 8:317.

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

Jaillon O, Aury JM, Brunet F, et al. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* 431:946-957.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8:275-282.

Kassahn KS, Dang VT, Wilkins SJ, Perkins AC, Ragan MA. 2009. Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. *Genome Res* 19:1404-1418.

Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* 428:617-624.

Kurosawa G, Takamatsu N, Takahashi M, et al. 2006. Organization and structure of hox gene loci in medaka genome and comparison with those of pufferfish and zebrafish genomes. *Gene* 370:75-82.

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

McGinnis W, Krumlauf R. 1992. Homeobox genes and axial patterning. *Cell* 68:283-302.

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

Negrisolo E, Kuhl H, Forcato C, Vitulo N, Reinhardt R, Patarnello T, Bargelloni L. 2010. Different Phylogenomic Approaches to Resolve the Evolutionary Relationships among Model Fish Species. *Mol Biol Evol* 27:2757-2774.

Ohno S. 1970. Evolution by gene duplication. Springer-Verlag, New York.

Penel S, Arigon AM, Dufayard JF, Sertier AS, Daubin V, Duret L, Gouy M, Perriere G. 2009. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics* 10 Suppl 6:S3.

Prohaska SJ, Stadler PF. 2004. The duplication of the Hox gene clusters in teleost fishes. *Theory Biosci* 123:89-110.

Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* 440: 341-345.

Semon M, Wolfe KH. 2007. Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. *Trends Genet* 23:108-112.

Stephens SG. 1951. Possible Significance of Duplication in Evolution. In: M Demerec, editor. Advances in Genetics: Academic Press. p. 247-265.

Studer R, Duret L, Penel S, Robinson-Rechavi M. 2008. Pervasive positive selection on duplicated and non duplicated vertebrate protein coding genes. *Genome Res 18:1393-1402.*

Sugino RP, Innan H. 2005. Estimating the time to the whole-genome duplication and the duration of concerted evolution via gene conversion in yeast. *Genetics* 171: 63-69.

Tamura K, Kumar S. 2002. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. *Mol Biol Evol* 19:1727-1736.

Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y. 2003. Genome duplication, a trait shared by 22000 species of ray-finned fish. *Genome Res* 13:382-390.

Taylor JS, Van de Peer Y, 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* 356:1661-1679.

Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y. 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.

Wernersson R, Pedersen AG. 2003. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res* 31:3537-3539.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387:708-713.