Evolution of gene function and regulatory control after whole-genome duplication: Comparative analyses in vertebrates

Karin S. Kassahn,[1] Vinh T. Dang,[1] Simon J. Wilkins,[2] Andrew C. Perkins,[2] and Mark A. Ragan[1,3]

[1] *The University of Queensland, Institute for Molecular Bioscience and ARC Centre of Excellence in Bioinformatics, Brisbane, QLD 4072, Australia;* [2] *The University of Queensland, Institute for Molecular Bioscience and Australian Zebrafish Phenomics Facility, Brisbane, QLD 4072, Australia*

The significance of whole-genome duplications (WGD) for vertebrate evolution remains controversial, in part because the mechanisms by which WGD contributed to functional evolution or speciation are still incompletely characterized. Fish genomes provide an ideal context in which to examine the consequences of WGD, because the teleost lineage experienced an additional WGD soon after divergence from tetrapods and because five teleost genomes are available for comparative analysis. Here we present an integrated approach to characterize these post-duplication genomes based on genome-scale synteny, phylogenetic, temporal, and spatial gene expression and on protein sequence data. A minimum of 3%–4% of protein-coding loci have been retained in two copies in each of the five fish genomes, and many of these duplicates are key developmental genes that function as transcription factors or signaling molecules. Almost all duplicate gene pairs we examined have diverged in spatial and/or temporal expression during embryogenesis. A quarter of duplicate pairs have diverged in function via the acquisition of novel protein domains or via changes in the subcellular localization of their encoded proteins. We compared the spatial expression and protein domain architecture of zebrafish WGD-duplicates to those of their single mouse ortholog and found many examples supporting a model of neofunctionalization. WGD-duplicates have acquired novel protein domains more often than have single-copy genes. Post-WGD changes at the gene regulatory level were more common than changes at the protein level. We conclude that the most significant consequence of WGD for vertebrate evolution has been to enable more-specialized regulatory control of development via the acquisition of novel spatiotemporal expression domains. We find limited evidence that reciprocal gene loss led to reproductive isolation and speciation in this lineage.

[Supplemental material is available online at www.genome.org.]

The availability of an ever-increasing number of complete genome sequences has fuelled research into the evolution and function of genomes as a whole. Eukaryotic genomes have been modified over the course of evolution not only by single gene duplications (Ohno 1970; Lynch 2002) but also by several rounds of whole-genome duplication (WGD) (Jaillon et al. 2004; Dehal and Boore 2005), which were typically followed by extensive gene loss. These WGD events would thus have had significant effects on gene regulatory control and protein–protein interactions. Nonetheless, WGD are comparatively common and have been described in plants (Vandepoele et al. 2002), yeast (Kellis et al. 2004), the ancestor of vertebrates (Dehal and Boore 2005), teleost fishes (Jaillon et al. 2004; Le Comber and Smith 2004), and the frog *Xenopus laevis* (Sémon and Wolfe 2008). Furthermore, polyploidy can be artificially induced by heat shock in rainbow trout and common carp, and triploid fish are commonly generated in aquaculture to achieve sterility and thus avoid interbreeding with native fish stocks (Le Comber and Smith 2004). The fact that ploidy levels can be so easily manipulated in teleost fishes and that several rounds of WGD and subsequent gene loss have occurred in vertebrate evolution challenges our experience that knocking-down or altering individual genes can suffice to disrupt normal vertebrate development and function.

Studying the function of post-duplication genomes can thus contribute to our understanding of how genomes evolve as a whole, which components are amenable to change, and by which mechanisms new functions or regulatory control evolve (e.g., Woolfe and Elgar 2007). In terms of biodiversity, loss of alternative copies of a duplicated locus has been suggested to promote within-population mating and to lead to reproductive isolation between populations. Speciation dynamics and gene loss patterns in polyploid yeast, for example, provide strong support for the "divergent resolution" hypothesis of speciation (Wong et al. 2002; Scannell et al. 2006). There is some evidence that reciprocal gene loss after WGD could have also contributed to the radiation of teleost fishes (Sémon and Wolfe 2007). These fishes experienced a WGD event during their early evolution, some 305–450 million years (Myr) ago (Amores et al. 1998; Christoffels et al. 2004; Hoegg et al. 2004; Vandepoele et al. 2004). Today, teleost fishes constitute the most speciose vertebrate lineage, with over 22,000 extant species (Taylor et al. 2003). The last WGD event has thus often been implicated as a driver for the radiation and diversification of this lineage (Amores et al. 1998; Meyer and Schartl 1999), although others have questioned the significance of this WGD for generating species diversity (e.g., Robinson-Rechavi et al. 2001).

[3] **Corresponding author.**
E-mail m.ragan@imb.uq.edu.au; fax 61-7-3346-2101.

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.086827.108.

19:1404–1418 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

The teleost-specific WGD offers great potential for understanding the evolution of this lineage as well as for understanding vertebrate genome evolution and function more generally. However, to date there have been no systematic, genome-scale studies investigating which genes have been retained in duplicate in different teleost lineages. Evolutionary theory predicts that most gene duplicates would rapidly become nonfunctional and lost (Force et al. 1999). For example, gene retention after WGD in the pufferfishes *Tetraodon nigroviridis* and *Takifugu rubripes* may be as low as 1%–5% (Aparicio et al. 2002; Jaillon et al. 2004), although there has been debate regarding these estimates (Brunet et al. 2006). Analysis of individual gene families in the zebrafish *Danio rerio* suggested that up to 20% of gene duplicates may have been retained from the last teleost-specific WGD event (Postlethwait et al. 2000, 2004; Woods et al. 2005). Previous studies have not been able to determine the proportion of retained duplicate genes, because these studies either were limited in the number of families investigated or did not distinguish between gene duplicates derived by WGD and those derived by gene-specific duplication events thereafter. Nevertheless, vertebrate genomes have been shown to contain a large number of anciently duplicated genes, many of which are expected to have originated by WGD (Blomme et al. 2006; Brunet et al. 2006).

In this study, we performed comparative genome analyses, including gene order (synteny) and phylogenetic analyses, in *D. rerio, T. rubripes, T. nigroviridis*, medaka (*Oryzias latipes*), and stickleback (*Gasterosteus aculeatus*) to identify gene duplicates retained from the last, teleost-specific WGD. We show that a minimum of 3%–4% of protein-coding genes have been retained in duplicate in each of the five fishes. Almost all *D. rerio* duplicate gene pairs examined here differed in spatiotemporal expression during embryogenesis, suggesting significant changes in gene regulatory control after WGD. The observed expression data support a model of neofunctionalization (Lynch et al. 2001), with many duplicates having acquired novel expression domains after duplication, although the signatures indicative of neofunctionalization are also easier to detect than those of other evolutionary fates, such as subfunctionalization. A quarter of duplicate pairs encode proteins with different protein domain architecture and/or subcellular localization, suggesting functional differences between their protein products. Our assessment of changes in regulatory control versus changes in protein sequence indicates that WGD primarily led to increased specialization of gene regulatory control of development, although some functional variation in coding sequence was observed. These data shed new light on the impact of WGD on vertebrate genome evolution and on how these post-duplication genomes have evolved new functionalities.

Results

Identifying sister chromosome regions based on conserved gene order

Paralogs derived from WGD are expected to be located in chromosome regions of shared ancestry. To help identify such regions, we first determined fish–human gene homology relationships by exhaustive, "all-against-all," sequence similarity searches, saving all matches with an *E*-value $<1 \times 10^{-3}$, assuming that true homologs will have more-significant *E*-values (McLysaght et al. 2002; Christoffels et al. 2004). Some 20,300 fish proteins had a match in the human proteome using this threshold, while some 18,500 human proteins had a match in each of the five fish proteomes (Table 1). Approximately 11,200 fish–human protein pairs were reciprocal best hits in each of these comparisons (Table 1). Fish–human and human–fish unidirectional and reciprocal best hits were used to build the initial gene (positional) homology matrices and to identify collinear regions in the genomes of fish and human, followed by a search for additional homologs that map to the identified syntenic regions (Fig. 1). Due to the lack of a physical genomic map for *T. rubripes*, synteny analyses could not be performed in this species. Approximately 9300 positional homologs, namely, genes that share significant sequence similarity as well as conserved gene order and chromosome location across genomes, were identified per fish–human genome comparison (Table 1). Some 2100 human gene loci had positional homologs in two fish genome regions, suggesting that these fish loci are duplicate loci retained from WGD (Table 1). Synteny maps

Table 1. **Comparative analyses of the genomes and proteomes of five teleost fishes**

		Sequence similarity search (SSEARCH)			Synteny analyses		Phylogenetic analyses		
	Protein-coding genes	Fish–human best hits	Human–fish best hits	Reciprocal best hits	Anchor-points	Retained duplicates	Proteins in all Ensembl trees	Proteins in *D. rerio* subtrees	Retained duplicates
Danio rerio	21,322	20,002	18,460	10,715	8038	1753	19,775	4842	1318
Oryzias latipes	20,131	17,975	18,462	11,265	9459	2128	18,711	5185	1436
Gasterosteus aculeatus	20,791	19,279	18,539	11,884	10,075	2274	19,223	4901	1669
Tetraodon nigroviridis	27,918	23,940	18,437	10,717	9455	2294	21,652	5116	1398
Takifugu rubripes	21,880	20,444	18,485	11,211	NA	NA	20,702	5168	1525

Protein-coding gene sequences were taken from Ensembl v48. The number of unidirectional and reciprocal best hits were determined using the exhaustive Smith-Waterman algorithm implemented in the SSEARCH sequence alignment software (Pearson 1995). Only protein match pairs with *E*-values $<1 \times 10^{-3}$ are shown. Synteny analyses were performed between the human and four fish genomes using the i-ADHoRe software (Vandepoele et al. 2002; Simillion et al. 2004), with a minimum of three anchor points required to define a collinear genomic region. The number of human–fish homologs that map to such collinear regions is shown. Where a single human genomic position had anchor points on two distinct fish chromosomes, the fish homologs were retrieved as potential gene duplicates retained from the teleost-specific whole-genome duplication. Note that due to the lack of a physical genomic map, synteny analyses could not be performed for *T. rubripes*. The last three columns describe the number of proteins represented in the 27,308 Ensembl protein family trees (v48) and in the *D. rerio* subtrees that were identified in this study. The number of retained duplicate predictions was based on tree topologies consistent with an origin by WGD.
NA, Not available.

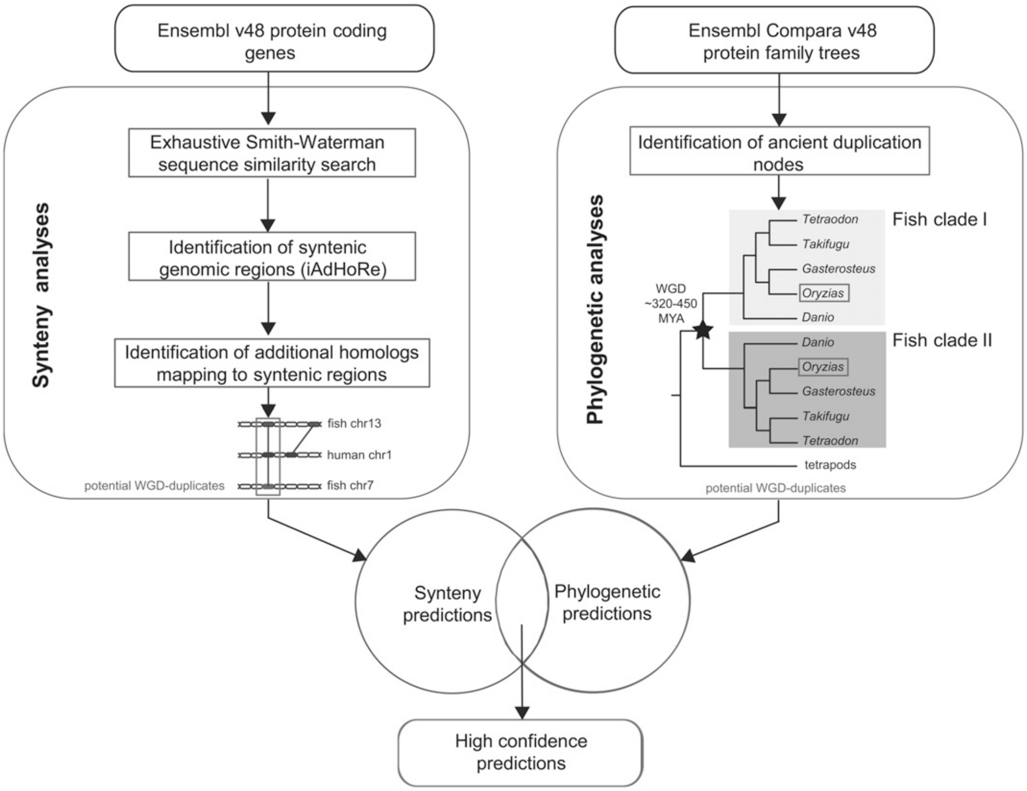

Ensembl v48 protein coding genes
Ensembl Compara v48 protein family trees
Synteny analyses
Exhaustive Smith-Waterman sequence similarity search
Identification of syntenic genomic regions (iAdHoRe)
Identification of additional homologs mapping to syntenic regions
fish chr13
human chr1
fish chr7
potential WGD-duplicates
Phylogenetic analyses
Identification of ancient duplication nodes
Tetraodon
Takifugu
Gasterosteus
Oryzias
Danio
Fish clade I
Fish clade II
WGD ~320-450 MYA
tetrapods
potential WGD-duplicates
Synteny predictions
Phylogenetic predictions
High confidence predictions

Figure 1. Computational analysis pipeline for the identification of gene duplicates retained from whole-genome duplication in five teleost fish genomes. Synteny and phylogenetic analyses provide independent lines of evidence for origin by WGD. The level of support in the synteny and phylogenetic analyses was used to categorize WGD-duplicates. Duplicates of "high confidence" map to sister chromosome regions of shared ancestry and have tree topologies consistent with an origin by WGD.

for each fish–human genome comparison illustrating the positional homologs identified in this study are available in Supplemental Figure 1. The syntenic map between *T. nigroviridis* and human corresponds largely with the one by Jaillon et al. (2004).

Protein family tree topologies consistent with origin by WGD

As a second line of evidence for origin by WGD, we used phylogenetic approaches and the protein family trees of Ensembl Compara v48. Given that *D. rerio* is the most basal fish taxon in our comparisons (Metscher and Ahlberg 1999), we searched for duplication nodes that predate the split between *D. rerio* and the other fish taxa (Fig. 1). Approximately 20,000 proteins per fish species were represented in the Ensembl v48 protein families, corresponding to ~89% of all protein-coding genes in these taxa. Some 1500 gene pairs in each of the five fish genomes showed evidence for origin by WGD based on the topology of the protein family tree (Table 1). Combining the results from the synteny and phylogenetic analyses, in each of the five fish taxa, we identified some 680 duplicates, corresponding to 3%–4% of protein-coding loci, with strong support for origin by WGD (Table 2; for the complete list, see Supplemental Table 1). Duplicates showing strong support for origin by WGD in zebrafish were used in computational and experimental analyses to characterize their present functions.

Table 2. Combined results from synteny and phylogenetic analyses

	Gene pairs with phylogenetic and synteny support	Gene pairs with phylogenetic support and synteny support in related fish species	Total no. of gene pairs with strong support for origin by WGD
Danio rerio	288	327	615
Oryzias latipes	469	203	672
Gasterosteus aculeatus	518	257	775
Tetraodon nigroviridis	422	228	650
Takifugu rubripes	NA	702	702
Average	424	343	683

Paralogs with support for origin by WGD in phylogenetic and synteny analyses were considered to show strong support for origin by WGD.
NA, Not available.

Presence of homologs in basal metazoan lineages and yeast

We expected that retained duplicates would be enriched for "vertebrate innovations," namely, genes that arose only in early vertebrate evolution, and that such duplicates would underlie diversification in this lineage. To test this hypothesis, we looked for homologs in basal metazoa and yeast. Of the 12,533 protein families in fish, 754 contained a WGD-duplicate. Families containing WGD-duplicates were significantly larger than the average

Table 3. **Comparison of family sizes for all Ensembl protein families containing fish sequences, with those containing a WGD-duplicate**

	All fish-containing Ensembl protein families	Ensembl protein families containing WGD-duplicate predictions
Total no. of families	12,533	754
Average family size (±SEM)[a]	38.1 ± 0.5	128.6 ± 4.8
Percentage of families with *Ciona* sequences	39.0	65.9
Percentage of families with insect sequences	36.2	64.1
Percentage of families with worm sequences	28.6	53.1
Percentage of families with yeast sequences	14.9	23.5
Average number of *Ciona* sequences	1.1 ± 0.02	3.0 ± 0.2
Average number of insect sequences	1.6 ± 0.03	4.6 ± 0.3
Average number of worm sequences	0.5 ± 0.01	1.2 ± 0.1
Average number of yeast sequences	0.2 ± 0.01	0.5 ± 0.1

The percentage of families containing homologs in basal metazoan taxa and yeast is indicated for both groups of families. *Ciona* refers to *Ciona intestinalis* and *Ciona savignyi* sequences, two basal chordates; insects include *Anopheles gambiae*, *Aedes aegypti*, and *Drosophila melanogaster*; worm refers to *Caenorhabditis elegans*; and yeast to *Saccharomyces cerevisiae*.

[a]Significant difference in means ($P < 0.0001$, df = 13285, $t = 36.64$).

fish protein family (Table 3). We then determined how many of these families also contained sequences in *Ciona intestinalis* or *Ciona savignyi*, two invertebrate chordates; the insects *Aedes aegypti*, *Anopheles gambiae*, or *Drosophila melanogaster*; the worm *Caenorhabditis elegans*; or the yeast *Saccharomyces cerevisiae*. Protein families that contained a WGD-duplicate were more likely to also contain invertebrate chordate, insect, worm, or yeast sequences than families that did not contain a WGD-duplicate (Table 3). For example, 66% of protein families containing a WGD-duplicate also included a *Ciona* sequence, while only 39% of all fish-containing protein families contained *Ciona* sequences. The corresponding values for families containing a homolog in yeast were 23% and 15%, respectively. To exclude the possibility that WGD-duplicates preferentially belong to ancient metazoan protein families simply because WGD-duplicates tend to belong to families of greater size, we examined the relationship between family size and taxonomic representation among the sequences in the family. The average number of invertebrate sequences is approximately threefold greater in families that contain WGD-duplicates than in families that do not (Table 3). Nevertheless, the percentage of families containing invertebrate sequences was consistently higher for families containing a WGD-duplicate, suggesting that the probability of a protein family containing invertebrate sequences was not purely a matter of family size.

Loss of duplicate gene copies in different teleost lineages and reciprocal gene loss

For each of the 754 protein families containing a WGD-duplicate, we inferred the number of gene losses since WGD along different teleost branches (Fig. 2). Given the uncertainties regarding the existence of a monophyletic clade "Smegmamorpha" (NCBI taxonomy vs. Metscher and Ahlberg 1999; Miya et al. 2003; Kawahara et al. 2008), gene losses were mapped onto two plausible tree topologies. Given that *D. rerio* is the most basal taxon in our comparisons, one of the *D. rerio* duplicate gene copies was arbitrarily designated the reference point against which the presence of the locus in the other fish taxa and in the sister clade was assessed. We marked all instances where a fish taxon or clade was inferred to have lost a copy of the locus, taking into account both plausible species tree topologies (Fig. 2). There was no evidence to suggest that different teleost species had re-tained a significantly different number of duplicate gene copies (based on studentized residuals, there were no outliers in the group). Among these gene losses, we identified 154 instances where two teleost species had lost alternative copies of the same locus, so-called "reciprocal gene losses" (not marked in Fig. 2). Of these 154 reciprocal gene loss events and assuming the existence of a clade Smegmamorpha, only 10 events were consistent with the loss having occurred at the time of species divergence (labeled RL in Fig. 2A). Assuming that *O. latipes* is basal to a clade containing *G. aculeatus* and the Tetraodontiformes, even fewer (seven) reciprocal gene loss events were consistent with the loss having occurred at the time of species divergence (labeled RL in Fig. 2B).

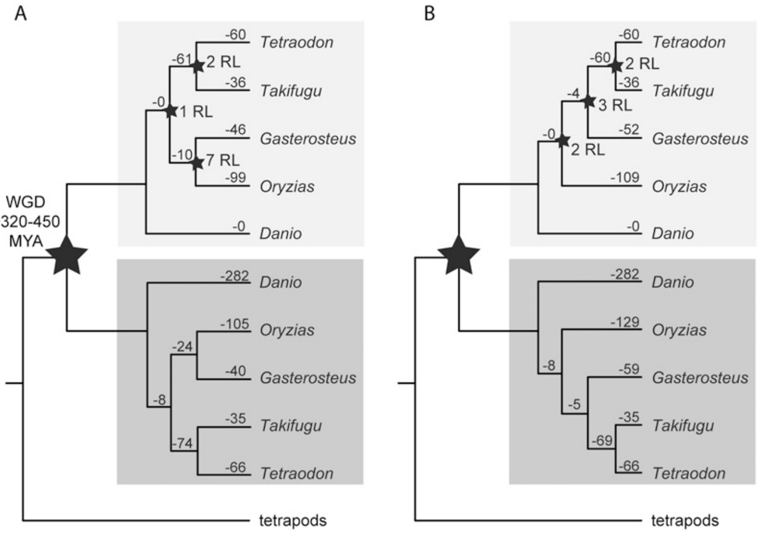


Figure 2. Inferred gene losses along different teleost lineages. For each of the 754 protein families containing a WGD-duplicate, one of the *D. rerio* duplicate gene copies was arbitrarily designated the reference point against which the presence of the locus in the other fish taxa and in the sister clade was assessed. We marked all instances where a fish taxon or clade was inferred to have lost a copy of the locus, taking into account two plausible species tree topologies: (*A*) the tree topology as represented in the NCBI taxonomy, which supports a monophyletic clade Smegmamorpha containing *Oryzias latipes* and *Gasterosteus aculeatus*, or (*B*) the tree topology supported by mitogenomic analyses, which resolves *O. latipes* as immediately basal to a clade containing *G. aculeatus* and the Tetraodontiformes. Instances where two teleost species had lost alternative copies of the same locus were counted as "reciprocal gene losses" (data not shown). Reciprocal gene losses that are consistent with the loss having occurred at the time of species divergence are marked RL.

Enrichment of Gene Ontology (GO) categories among WGD-duplicates

To test whether the probability of duplicate gene retention was related to gene function, we performed gene set enrichment analysis in *D. rerio*. Of the 615 duplicate gene pairs in this species, 674 loci had GO annotations in the Zebrafish Information Network (ZFIN) gene association file (Sprague et al. 2006), and we compared their GO terms to the complete set of GO annotations in ZFIN. Ninety-seven GO terms were significantly enriched among WGD-duplicates with adjusted $P < 0.1$ (Table 4; for a complete list of significant GO terms, see Supplemental Table 2). Enriched terms included, for example, "calcium ion transport," "transcription" and "transcription factor activity," "integrin-mediated signaling pathway," and "growth factor activity" (Table 4; Supplemental Table 2). Genes annotated with function in "calcium ion transport" included, for example, the annexin genes *anxa1a/anxa1b* and *anxa3a/anxa3b*, the ATPase genes *atp2a2a/atp2a2b*, and the calbindin 2 genes *calb2/calb2l*. Many other genes retained in duplicate in zebrafish have annotated functions during development and include, for example, *fzd8a* and *fzd8b* and *otx1* and *otx1lb*.

Expression localization of WGD-duplicates

Having characterized the types of genes retained in duplicate from the last, teleost-specific WGD, we assessed the extent to which these duplicates have acquired different functional roles as indicated by differences in the spatial domains of expression. For *D. rerio*, the ZFIN gene expression database is a comprehensive public resource of mRNA in situ hybridization and RT-PCR expression data (Sprague et al. 2006). A total of 97 WGD-gene pairs have expression localization data available in ZFIN, encompassing 949 individual expression observations (Supplemental Table 3). We categorized the expression patterns of duplicate gene copies as being the same, partially overlapping, or different and as being spatially restricted or ubiquitous throughout the animal (Table 5; Supplemental Table 3). For example, the WGD-duplicates *fzd8a* and *fzd8b*, which encode *wnt* signaling receptors, localize to different anatomical regions during the segmentation and pharyngula stages (Table 5; Supplemental Table 3). Approximately 65% of all expression observations listed distinct or only partially overlapping expression localizations for duplicate gene copies, while 5% described the same expression localization with both copies being spatially restricted and 30% described nonspatially restricted expression for both gene copies. Of the 97 WGD-gene pairs, 87% differed in expression localization during at least one developmental stage, while only 13% shared the same expression domain during all developmental stages investigated thus far. The former value may overestimate the true percentage of gene pairs with similar expression localizations, as some gene pairs may differ at developmental stages or under conditions not examined to date. Of these 13% of gene pairs with common expression localization, the majority (62%) were expressed in a nonspatially restricted manner throughout the animal. The probability of detecting differences in expression localization was not affected by the number of expression observations, as five out of the six gene pairs with the greatest number of expression observations (>20) showed the same expression localization (data not shown) and no other trends in the data suggested such an effect.

During the early stages of development (the zygote, cleavage, and blastula stages), 97% of WGD-duplicates shared the same expression localization. From gastrula to juvenile stages, the majority of WGD-duplicates (~73%) showed either distinct or only partially overlapping expression domains. The percentage of gene duplicates with distinct expression domains was greatest (~25%) during the segmentation and pharyngula stages (Supplemental Table 3).

Temporal expression of WGD-duplicates during embryogenesis

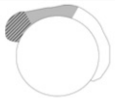
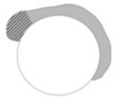
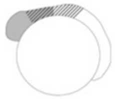
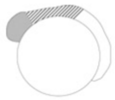
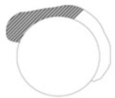
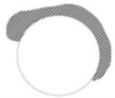
WGD-duplicates may differ not only in the spatial domain of expression but also in their temporal profile of expression during embryogenesis, especially since these loci are enriched for transcription factors and signaling genes with important functions in development. To assess this possibility, we examined two microarray time-course experiments of zebrafish embryogenesis (Mathavan et al. 2005; S Wilkins, M Kerr, M Köppen, B Gardiner, D Taylor, C Simons, M Landsberg, S Grimmond, C Heisenberg, and A Perkins, in prep.). Using stringent sequence

Table 4. A selection of significant Gene Ontology terms identified by gene set enrichment analysis comparing the Gene Ontology annotations of *Danio rerio* genes retained from whole-genome duplication to the complete set of *D. rerio* GO annotations in ZFIN

GO ID	GO name	GO subontology	Count among WGD-duplicates (674)	Count among all ZFIN GO annotations (13,571)	Adjusted *P*-value (FDR)
GO:0006816	**Calcium ion transport**	**B**	**37**	**208**	$\mathbf{4.20 \times 10^{-14}}$
GO:0016021	Integral to membrane	C	136	1618	1.49×10^{-9}
GO:0004859	Phospholipase inhibitor activity	M	6	7	3.70×10^{-6}
GO:0050789	Regulation of biological process	B	116	1575	1.49×10^{-4}
GO:0008083	Growth factor activity	M	15	81	3.08×10^{-4}
GO:0006006	**Glucose metabolic process**	**B**	**11**	**51**	**0.001**
GO:0007229	Integrin-mediated signaling pathway	B	7	23	0.003
GO:0003700	Transcription factor activity	M	51	622	0.005
GO:0019219	**Regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process**	**B**	**74**	**992**	**0.006**
GO:0048731	System development	B	50	649	0.024

P values were adjusted using the method of Benjamini and Hochberg (1995) for controlling the false discovery rate. B refers to biological process; M, molecular function; and C, cellular component in the GO ontology. For a complete list of significant GO terms, see Supplemental Table 2. GO terms that are representative of a cluster of GO terms that share annotation in the same set of genes are in bold. For example, the GO cluster "regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process" includes the terms "regulation of transcription," "transcription," and "regulation of gene expression."

Table 5. **Expression localization of duplicate gene pairs derived from whole-genome duplication in *Danio rerio***

Type of expression localization	Schematic of expression localization	Count	Expression example				
			Gene A	Expression localization (A)	Gene B	Expression localization (B)	Developmental stage
Expression of B subset of A, A and B spatially restricted		94	*fabp11*	Retina, vein, lens	*fabp11l*	Retina	Pharyngula (prim-5)
Expression of B subset of A, A unrestricted		189	*esrrgl*	Whole organism	*esrrg*	Tail bud	Segmentation (1-4 somites to 10-13 somites)
A and B partial overlap, A and B spatially restricted		146	*gpx1a*	Neuromast, retina, gut, liver, lateral line system, optic tectum	*gpx1b*	Neuromast, otic vesicle	Pharyngula (high-pec)
A and B different		191	*fzd8a*	Pharyngeal arch 3-7 skeleton, forebrain, neural tube, ventral mesoderm,hindbrain	*fzd8b*	Telencephalon, diencephalon	Segmentation (14-19 somites)
A and B same, A and B spatially restricted		43	*arr3*	Epiphysis	*arr3l*	Epiphysis	Segmentation (20-25 somites) to pharyngula (prim-25)
A and B same, unrestricted		286	*nedd8*	Whole organism	*nedd8l*	Whole organism	Zygote to hatching (pec-fin)

One example per category was selected to illustrate the type of expression localization. The expression domain of duplicate gene A is gray-shaded, the one of duplicate gene B hatched. Expression data were taken from the ZFIN database (Sprague et al. 2006). In total, 949 expression observations of WGD-duplicate gene pairs in the same developmental stage are represented in ZFIN.

comparisons, we identified 67 gene pairs that were represented by microarray probes that discriminated between duplicate gene copies. Only four of the 67 gene pairs showed significant coregulation across embryogenesis based on significant Pearson correlation coefficients and significance thresholds determined from random probe sets following the approach of Blanc and Wolfe (2004), while the remaining 63 showed differences in the temporal expression of the two gene copies (Fig. 3; Supplemental Table 4). Of these latter 63, five gene pairs showed inverse expression profiles with Pearson correlation coefficients exceeding the significance threshold (Fig. 3; Supplemental Table 4).

Spatiotemporal expression of WGD-duplicates

For two gene pairs with distinct temporal expression profiles (*zgc:73223* and *g12*; *fbxl14* and *fbxl14a*), no spatial expression data were available in ZFIN. To assess whether these gene pairs also differed in their spatial domains of expression, we performed in situ hybridization experiments using established methods (Supplemental Methods; Wilkins et al. 2008). Both gene pairs showed evidence for spatial differences in expression at 24 hours post-fertilization (hpf) (Supplemental Fig. 2), although further experiments are required to fully characterize the expression domains of these loci. For 16 additional gene pairs with temporal expression data, spatial expression data were available in ZFIN. In all 18 cases in which both sources of expression data were available, gene pairs differed in temporal and/or spatial expression with 15 gene pairs differing in both (Supplemental Tables 3, 4).

Comparison of expression domains to those of the single mammalian ortholog

To assess how the expression of duplicated genes changed after WGD, we compared their spatial expression domains in zebrafish to those of the single mouse ortholog, assuming that the expression domains of the mouse ortholog reflect those of the vertebrate ancestor before the divergence of the tetrapod and teleost lineage. While we currently lack expression data for a ray-finned fish that diverged before the WGD to perform such comparison, the time of independent evolution after the divergence of lobe- and ray-finned fishes and before the teleost WGD may also have contributed to expression variation between mouse and fish, independent of the effects of the teleost-specific WGD. To make this comparison, we chose the time when organogenesis is essentially completed, which in zebrafish occurs between 24 and 48 hpf and in mouse between day 9.5 and 16.5 post-coitum. A total of 47 duplicate gene pairs had corresponding in situ hybridization data in mouse available in the Mouse Genome Informatics (MGI) database (Table 6; Supplemental Table 5; Bult et al. 2008). We excluded nine comparisons from the analysis because transcripts from these loci were expressed in a nonspatially restricted manner in zebrafish or mouse, and we could not exclude the possibility that this was an artefact of poor probe specificity. Similarly, where different sources reported spatially restricted versus ubiquitous expression throughout the whole embryo, the spatially restricted expression was used to represent the locus, assuming again that ubiquitous expression was an artefact of poor probe specificity. Of the remaining 38

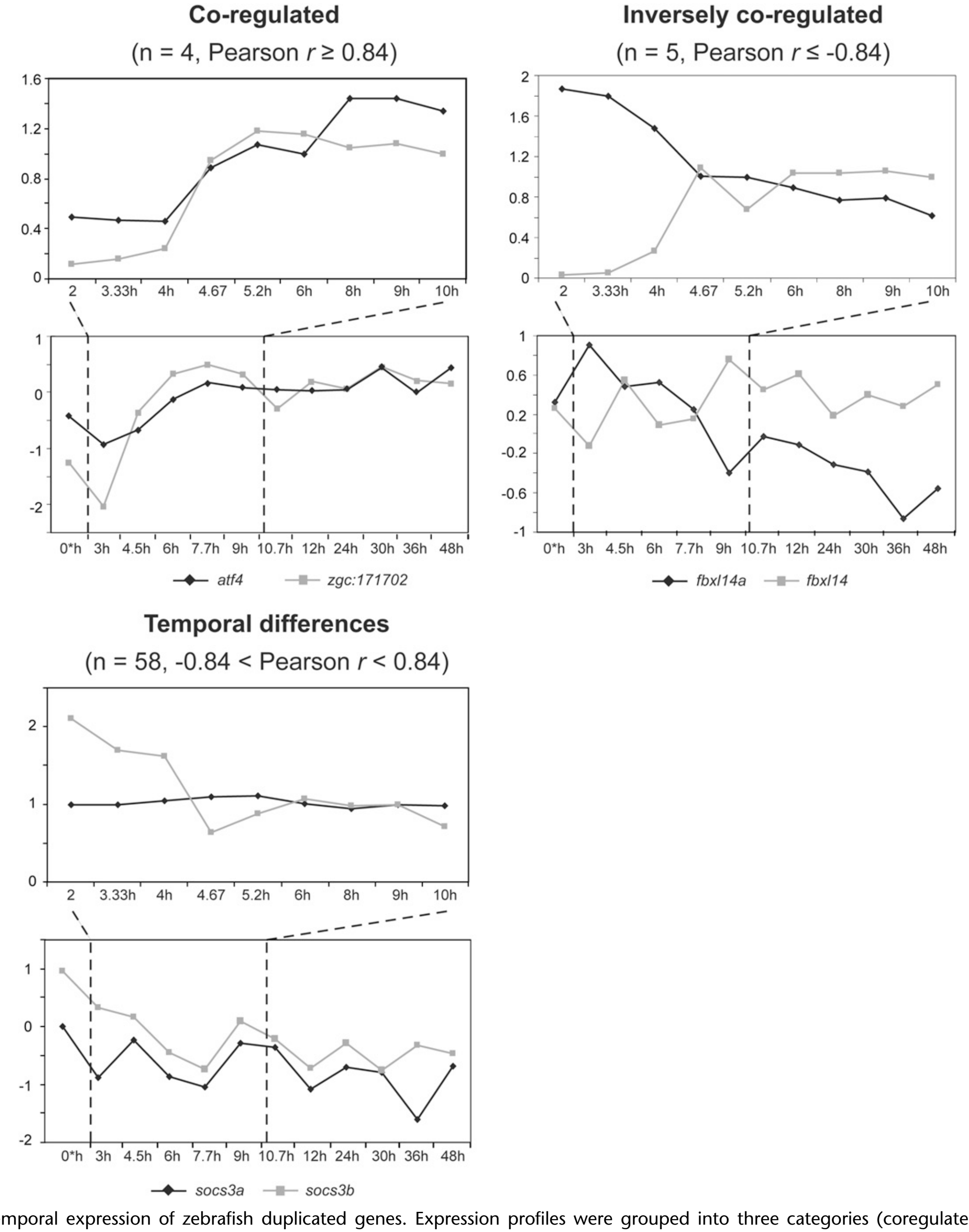


Figure 3. Temporal expression of zebrafish duplicated genes. Expression profiles were grouped into three categories (coregulated, inversely coregulated, temporal differences) based on their Pearson correlation coefficients (*r*) and statistical significance thresholds determined from the distribution of *r* values for 100,000 random probe pairs, following the approach of Blanc and Wolfe (2004). Only one gene pair is shown here to represent each category. In each category, the *top* and *bottom* panels refer to microarray data from Wilkins et al. (S Wilkins, M Kerr, M Köppen, B Gardiner, D Taylor, C Simons, M Landsberg, S Grimmond, C Heisenberg, and A Perkins, in prep.) and Mathavan et al. (2005), respectively, while the dashed lines indicate overlapping time points measured in both data sets. The scale on the *y*-axis differs between the two microarray data sets due to the different normalization methods used by the investigators.

gene pairs, 20 (53%) had novel expression domains not found in the mouse ortholog, supporting a model of neofunctionalization (Table 6). Four gene pairs (11%) had expression domains that were subsets of those of mouse, supporting a model of subfunctionalization, while 10 gene pairs (26%) showed evidence to support both neo- and subfunctionalization. Finally, four zebrafish gene pairs (11%) showed the same expression pattern during this developmental time point, potentially indicating functional redundancy.

Domain architecture of proteins encoded by WGD-duplicates

Besides changes in expression and hence regulatory control, important functional changes may also occur at the protein-coding

Table 6. Comparison of spatial expression domains of 38 zebrafish duplicated genes and their single mouse ortholog

Evolution of expression domains (ED)	Count	Zebrafish locus	Expression domains	Mouse ortholog	Expression domains
Acquisition of novel EDs	20	*anxa2a*	Epidermis, pharynx	*Anxa2*	Telencephalon
		anxa2b	Gut		
Complementary ED subsets	4	*egr2a*	Rhombomeres 3 and 5	*Egr2 (Krox-20)*	Hindbrain, rhombomeres 3 and 5
		egr2b	Hindbrain, rhombomeres 3 and 5		
Mixture of ancestral and novel EDs	10	*pax2a*	Cloacal chamber, eye, forebrain, hindbrain, mesoderm, midbrain, midbrain hindbrain boundary, nervous system, neural tube, optic stalk, otic vesicle, pharyngeal arch 3-7 skeleton, pharyngeal endoderm, proctodeum, pronephros, renal tuble, spinal cord, etc...	*Pax2*	Ear, eye, renal/urinary system, spinal cord
		pax2a	Midbrain hindbrain boundary, otic vesicle		
Retention of same EDs	4	*slit1a*	Forebrain	*Slit1*	Forebrain, liver, retina
		slit1b	Forebrain		

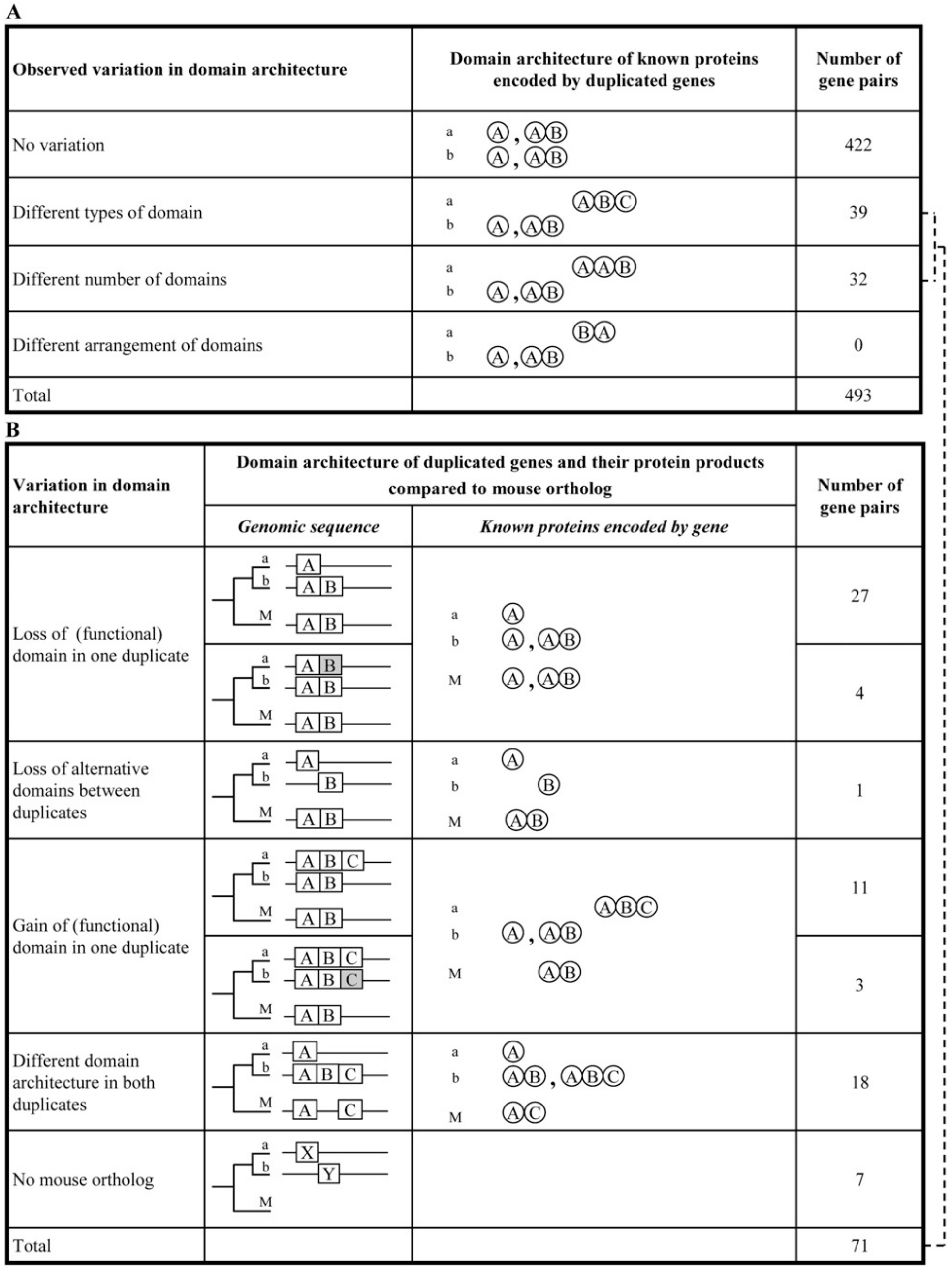
Zebrafish expression data were extracted from ZFIN (Sprague et al. 2006), those of mouse from MGI (Bult et al. 2008). All comparisons were made at a time when organogenesis is completed, which in zebrafish occurs 10–48 h post-fertilization and in mouse 9.5–16.5 d post-coitum. One representative comparison is shown here for each category (for the complete list including primary references, see Supplemental Table 5). Expression domains are given in alphabetical order for ease of comparison.

sequence level. To assess this possibility, we first analyzed the protein domain architecture of proteins encoded by WGD-duplicates by searching protein sequences against a database of known protein domain motifs (Pfam-A). Using this strategy, 493 of the 615 gene pairs in zebrafish could be annotated with domain information (Fig. 4). Where the locus is known to be subject to alternative splicing, we compared the domain architecture of all alternatively spliced products. In the majority of cases (422), both duplicate loci encoded proteins with the same protein domain architecture. We found 39 instances where the protein products differed in the types of domains and 32 where they differed in the number of protein domains, but no examples where proteins differed in the arrangesment of domains (Fig. 4). Given the limited sequence coverage of the zebrafish transcriptome, for the 71 gene pairs that showed differences in domain architecture, we investigated whether there was evidence for the "missing" domains at the genomic sequence level, even if the transcripts encoding these domains have not yet been observed. In seven cases, the genomic sequence had the potential to encode for the missing domains, while in the remaining 64 cases we found no evidence of the missing domains at the genomic sequence level (Fig. 4).

Comparing the domain architecture of zebrafish duplicated genes to those of the single mouse ortholog and assuming that the domain architecture of the mouse ortholog represents that of the ancestral locus prior to WGD, we found 31 cases (48%) in which one of the zebrafish duplicated gene loci had lost a functional domain, 14 cases (22%) in which it had gained a novel protein domain, and 18 cases (28%) with a more complex history of domain gain and loss (Fig. 4). In one other case, the two zebrafish loci encoded complementary domain subsets: *zgc:158388* encodes two alternatively spliced proteins with the domain architecture [PH,ArfGap] and ENSDARG00000039386 encodes a single protein with domain architecture [ArfGap,Ank,Ank,Ank,Ank], while the mouse ortholog *Acap3* encodes a single protein of domain architecture [PH,ArfGap,Ank,Ank,Ank,Ank], potentially indicating partitioning of the ancestral gene functions among the two zebrafish gene copies. The complete protein domain annotation in mouse and zebrafish is available as Supplemental Table 6. Given the long time since WGD and the possibility that orthologs may diverge in protein domain architecture for reasons other than the effects of WGD, we also compared the protein domain architecture of one-to-one zebrafish–mouse orthologs. In total, we identified 5878 zebrafish–mouse one-to-one orthologs of which 8.8% differed in the types of domains they encoded, while another 6.5% differed in the number of protein domains they encoded. The remaining 84.7% showed the same protein domain architecture in mouse and zebrafish. In contrast, 11.1% of WGD-duplicates differed from the single mouse ortholog in the types of domains encoded, while 11.7% differed in the number of protein domains. Only 77.2% of WGD-duplicates had the same protein domain architecture as the single mouse ortholog, indicating that more WGD-duplicates have diverged in protein domain architecture than have single-copy genes (χ^2 = 34.3696, degrees of freedom [df] = 1, $P = 4.558 \times 10^{-9}$).

Subcellular localization of proteins encoded by WGD-duplicates

In addition to changes in domain architecture, changes in the subcellular localization of proteins may bear important functional consequences. For example, secreted proteins may act as messengers, while membrane-associated proteins may function as channels or receptors. To predict the subcellular localization of proteins encoded by WGD-duplicates, we used computational methods and the presence or absence of signal peptides and transmembrane domains. A total of 529 WGD-gene pairs could be annotated with information regarding the subcellular localization of the encoded proteins, and in 73 cases, subcellular localization differed between proteins encoded by duplicate gene loci (Fig. 5; Supplemental Table 7). For example, *sema3fa* (ENSDARG 00000011163) encodes a type II membrane protein, while its WGD-duplicate *sema3fb* (ENSDARG00000055373) encodes a soluble intracellular protein (Supplemental Table 7). Semaphorins are important receptors and signaling molecules in neural differentiation (Pasterkamp and Kolodkin 2003; Yu and Moens 2005). The different subcellular localization of *sema3fa* and *sem3fb* suggest



A

Observed variation in domain architecture	Domain architecture of known proteins encoded by duplicated genes	Number of gene pairs
No variation	a Ⓐ, ⒶⒷ b Ⓐ, ⒶⒷ	422
Different types of domain	a ⒶⒷⒸ b Ⓐ,ⒶⒷ	39
Different number of domains	a ⒶⒶⒷ b Ⓐ,ⒶⒷ	32
Different arrangement of domains	a ⒷⒶ b Ⓐ,ⒶⒷ	0
Total		493

B

Variation in domain architecture	Domain architecture of duplicated genes and their protein products compared to mouse ortholog		Number of gene pairs
	Genomic sequence	*Known proteins encoded by gene*	
Loss of (functional) domain in one duplicate	a A b A B M A B	a Ⓐ b Ⓐ, ⒶⒷ M Ⓐ, ⒶⒷ	27
	a A ▨B b A B M A B		4
Loss of alternative domains between duplicates	a A b B M A B	a Ⓐ b Ⓑ M ⒶⒷ	1
Gain of (functional) domain in one duplicate	a A B C b A B M A B	a ⒶⒷⒸ b Ⓐ, ⒶⒷ M ⒶⒷ	11
	a A B C b A B ▨C M A B		3
Different domain architecture in both duplicates	a A b A B C M A C	a Ⓐ b ⒶⒷ, ⒶⒷⒸ M ⒶⒸ	18
No mouse ortholog	a X b Y M		7
Total			71

Figure 4. Comparison of the protein domain architecture of zebrafish WGD-duplicates and those of their single mouse ortholog. (*A*) A total of 493 zebrafish gene pairs could be annotated with protein domain information when searched against the Pfam-A database of known protein domain motifs, with the majority of gene pairs (422) encoding the same types and number of domains. (*B*) The 71 zebrafish gene pairs that differed in domain architecture were compared to their single mouse ortholog to infer whether duplicate gene copies had lost or gained a domain compared with the ancestral locus. In seven comparisons where duplicate zebrafish genes differed in the types of domains encoded, we found evidence for the missing domains at the genomic sequence level, even if the transcripts or proteins encoding these domains have not yet been observed. These are indicated by gray boxes in the diagram. For seven other gene pairs, there were no orthologs in mouse.

that the proteins encoded by this duplicate gene pair carry out fundamentally different functions.

Changes in regulatory control versus changes in protein-coding sequence

Finally, we investigated the relative importance of changes in regulatory control as indicated by differences in the spatiotemporal expression of WGD-duplicates versus changes in protein function as indicated by differences in the domain architecture or subcellular localization of proteins encoded by WGD-duplicates. In summary, 93% of the 138 gene pairs investigated differed in spatial and/or temporal expression (Table 7). In contrast, only 24% of 545 gene pairs encoded proteins that differed in domain architecture and/or subcellular localization (Table 7). Of the 134 gene pairs for which we had expression and protein domain/

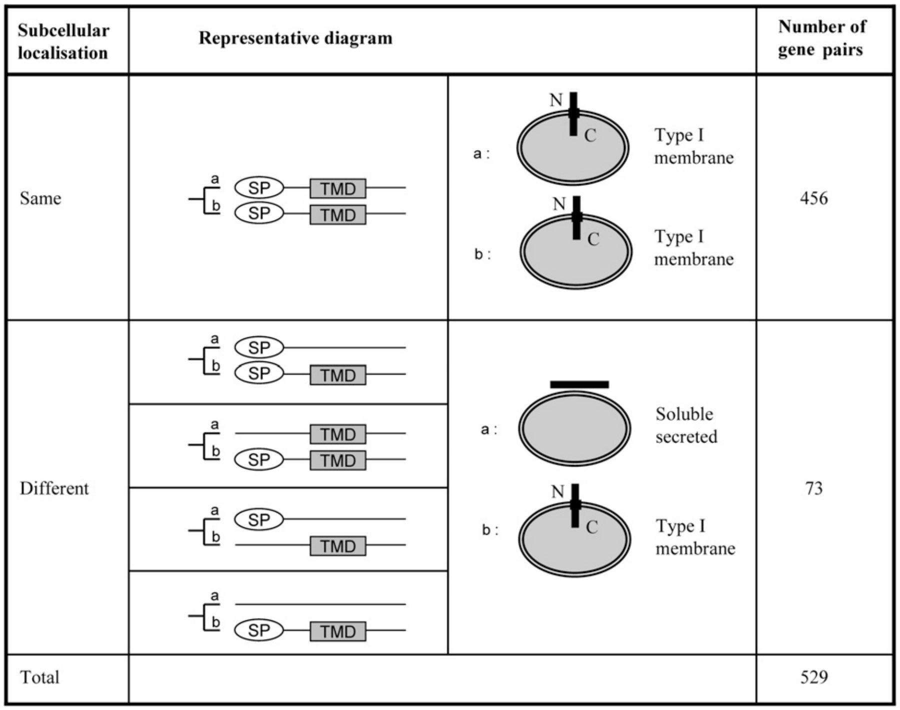


Figure 5. Comparison of the subcellular localization of proteins encoded by zebrafish WGD-duplicates. Subcellular localization was assessed using computational methods, SignalP (Bendtsen et al. 2004) and TMHMM (Krogh et al. 2001). 14% of WGD-duplicates encode proteins that differ in subcellular localization.

localization data, 24 gene pairs (18%) differed in both. This proportion is slightly lower than one would expect if gene pairs differed in regulatory control and protein domain/localization according to their individual proportions (0.93 × 0.24 = 0.22, or 22%), indicating that after some 400 Myr of evolution, duplicates retained after WGD either differ in regulatory control or in protein function, but not usually in both. Comparison to the single mouse ortholog showed that many zebrafish gene pairs have acquired novel expression localizations and protein domain architectures, potentially indicating that neofunctionalization has been a more common evolutionary fate than subfunctionalization or redundancy. Our differential ability to identify these alternate evolutionary fates may, however, have biased our results in favor of neofunctionalization.

Discussion

The functional significance of duplicate gene loci, including those derived by WGD in teleosts, has received much attention recently (MacCarthy and Bergman 2007; Wapinski et al. 2007; Conant and Wolfe 2008; Kleinjan et al. 2008). Several models have been proposed for the evolution of genes after duplication, including models describing the evolution of new functions (neofunctionalization), the partitioning of ancestral functions (subfunctionalization), or a combination of both (subneofunctionalization) (Lynch and Force 2000; He and Zhang 2005; Roth et al. 2007). Most duplicates are assumed to diverge in function, thus avoiding genetic redundancy (Blanc and Wolfe 2004; Sharma et al. 2006; Kleinjan et al. 2008), but some duplicates can be subject to selection for increased gene dosage (Conant and Wolfe 2008). Functional divergence may occur through changes in the substrate of an enzyme, the binding partners of a protein or the response to protein binding, the subcellular localization, and/or the spatial or temporal expression of the locus. Despite the significance of studying the functional divergence of WGD-duplicates for our understanding of genome and gene function evolution, WGD-duplicates in vertebrates are still poorly characterized.

Teleost fishes provide a unique opportunity to investigate the evolution of gene function after genome duplication in vertebrates, as five genomes are available for comparative analyses, extensive expression data are available for *D. rerio,* and both *D. rerio* and *O. latipes* are important model organisms for human developmental biology. In *D. rerio* a number of duplicated developmental genes—such as *eng1a* and *eng1b*, identified as WGD-duplicates in this study, and *sox9a* and *sox9b*—have already been shown to have partitioned subfunctions compared with their human orthologs (Force et al. 1999; Cresko et al. 2003). However, it remains unknown whether subfunction-partitioning is a common mechanism underlying the retention of gene duplicates in teleosts.

Table 7. **Number of zebrafish gene pairs derived from WGD showing changes in regulatory control versus changes in protein function**

	Regulatory control			Protein function		
	Spatial expression	Temporal expression	Spatial or temporal expression	Domain architecture (DA)	Subcellular localization (SL)	DA or SL
Same	12 (13%)	4 (6%)	10 (7%)	422 (86%)	456 (86%)	416 (76%)
Different	80 (87%)	63 (94%)	128 (93%)	71 (14%)	73 (14%)	129 (24%)
Total	92	67	138	493	529	545

Spatial expression data were obtained in this study or from ZFIN, temporal expression data were obtained from Wilkins et al. (S Wilkins, M Kerr, M Köppen, B Gardiner, D Taylor, C Simons, M Landsberg, S Grimmond, C Heisenberg, and A Perkins, in prep.) and Mathavan et al. (2005), domain architecture of proteins encoded by WGD-duplicates was determined using InterProScan against the Pfam-A database, and subcellular localization was determined using SignalP (Bendtsen et al. 2004) and TMHMM (Krogh et al. 2001). Only 7% of gene pairs with spatial and/or temporal expression were identical in expression, while 76% of gene pairs with domain architecture and/or subcellular localization information were identical in protein annotation.

Thus, we set out to identify as many WGD-duplicates as possible to allow investigation of their functional evolution after duplication. We took special care to distinguish WGD-duplicates from paralogs of other origin, because, based on evidence from yeast, such duplicates are subject to different evolutionary constraints (Wapinski et al. 2007). Given the long time since the teleost-specific WGD, teleost genomes are expected to have undergone multiple rearrangements (Kasahara et al. 2007) and are likely to have lost many anciently duplicated loci. To overcome these challenges, we have used two independent approaches, based on conserved gene order (synteny) and phylogenetic inference. Using a comparative approach and gene order information from multiple related genomes, following an approach similar to Simillion et al. (2004) and Van de Peer (2004), we were able to map approximately half of all protein-coding genes in fish to colinear chromosomal regions with human; using a single fish genome recovered significantly fewer collinear genomic regions and 8%–14% fewer inter-genome positional homologs (data not shown). Paralogous gene copies that mapped to sister chromosome regions in fish were retrieved as potential WGD-duplicates.

However, given the possibility that paralogs may locate to chromosomal sister regions by a process other than shared ancestry, for example, due to genomic rearrangements, it was important to test the co-orthology of potential fish WGD-duplicates using phylogenetic methods. To exclude gene duplicates whose tree topologies were consistent with origin by WGD, but which originated by gene-specific duplication after WGD instead, we required support for origin by WGD in both the synteny and phylogenetic analyses. This strategy identified some 680 gene pairs, or 3%–4% of protein-coding gene loci in each of the five fish genomes, and represents the first genome-scale estimate of the number of gene duplicates retained from this ancient WGD event. Previous studies were based either on the analysis of individual gene families or on the presence of paralogs of any type (Aparicio et al. 2002; Jaillon et al. 2004). Our findings represent a minimum estimate of the true number of duplicate genes retained after WGD, as our analyses are likely biased toward identifying more-conserved gene pairs. The quality of current teleost genome annotations also likely impacts on the number of WGD-duplicates identified in such analyses. Complex genomic rearrangements could have precluded identification of collinear regions for some duplicates, and errors with phylogenetic tree inference could have precluded identification of co-orthology relationships for others. For example, we could not recover synteny regions for *sox9a* and *sox9b* despite published literature suggesting that these paralogs have been derived from WGD (Cresko et al. 2003). Thus, we expect that our analyses will underestimate the true number of gene duplicates retained from the last, teleost-specific WGD. However, the approach we have implemented here identifies substantially more WGD-gene pairs than has any previous study, suggesting that we have reduced any potential bias associated with these limitations as much as currently possible.

Acknowledging these limitations, we found no significant differences in the number of gene duplicates retained from WGD in the five teleost species. We mapped reciprocal gene losses describing the loss of alternative copies in different teleost lineages and found that very few were consistent with the loss having occurred at the time of species divergence. These results contrast with those of Sémon and Wolfe (2007), who found many instances of reciprocal losses between *D. rerio* and *T. nigroviridis* and concluded that reciprocal gene loss was an important driver in the radiation of teleosts. Our analysis included three additional teleost genomes, and these additional data allowed us to assign a relative time to the inferred gene losses and thus test whether the time of the loss coincided with species divergence. Our results show that few of these losses coincided with the time of species divergence so that most losses are unlikely to have contributed to reproductive isolation in early teleosts. Nevertheless, simple calculations show that a small number of reciprocal gene losses can significantly reduce "hybrid" fitness and could thus lead to reproductive isolation between incipient species. For example, 25% of the offspring (F_1) of parents with alternative copies of a WGD-duplicated locus would not inherit any copy of the locus. With each additional reciprocal gene loss, F_1 fitness decreases further. It is therefore possible that even just a few of these losses could have been sufficient to drive reproductive isolation and speciation, especially if the loci involved were dosage-sensitive. Nevertheless, our data suggest that the more significant contribution of WGD for teleost radiation was the capacity for functional divergence of post-duplication genomes, potentially allowing exploration of new ecological niches and reducing the risk of extinction as suggested by Crow and Wagner (2006).

Here, we show that the probability for retention after WGD is correlated with gene function as retained duplicates are enriched for function in signaling, transcription, calcium ion transport, and metabolism. A general pattern emerges where genes with function in transcription and regulatory control are preferentially retained after WGD in yeast, plants and vertebrates (Blanc and Wolfe 2004; Davis and Petrov 2005; Blomme et al. 2006). It is possible that these types of genes have biochemical features that make them more amenable to evolving novel functional roles as proposed by Conant and Wolfe (2008). Alternatively, duplication of other types of genes, so-called "duplication-resistant genes," may be prohibited because of immediate detrimental effects of duplication (Conant and Wolfe 2008). We further show that retained duplicate genes are enriched for ancient metazoan or eukaryotic genes, indicating that WGD did not capitalize on vertebrate-specific innovations. Instead, many of the families that contain WGD-duplicates in teleosts already existed during early metazoan evolution, and these families have expanded in vertebrates not only via preferential retention after WGD but also via gene-specific duplications thereafter. These ancient families are thus particularly amenable to duplication, but the characteristics underlying their "duplicability" are yet to be determined.

To understand the functional significance of WGD-duplicates, we first examined the spatiotemporal expression of duplicated gene loci in *D. rerio* and found that almost all gene pairs examined have changed regulatory control since duplication. We also examined the protein products of WGD-duplicates in *D. rerio*. About a quarter of WGD-duplicates differed in domain architecture and/or subcellular localization. We found that duplicates tend to differ in regulatory control or in domain architecture/protein subcellular localization, but not in both. At present, it is not clear why functional adaptation should happen at the regulatory or at the protein level, but not at both. It is possible, however, that these results are biased by our inability to fully characterize the protein functions of WGD-duplicates. Comparison of the spatial expression domains in *D. rerio* to those of the single mouse ortholog suggest that many *D. rerio* duplicates have acquired novel expression domains and that, at the regulatory level, neofunctionalization is more common than subfunctionalization. Similarly, comparing the protein domain architecture of *D. rerio* WGD-duplicates to those of the single mouse ortholog, we found many examples describing the acquisition of a novel protein domain, the loss of a domain in one gene copy, and a complex history of domain loss and gain, but only one example of the loss of complementary domains, potentially indicating that, at the protein level, neofunctionalization may also be the prominent fate for the evolution of WGD-duplicates. However, as discussed by Huminiecki and Wolfe (2004), it is more difficult to meet the conditions to identify subfunctionalization than those for neofunctionalization, potentially introducing an ascertainment bias that would compromise accurate quantification of the relative rates of neo- versus subfunctionalization. In particular, proteins that encode complementary domain subsets may not have been identified as coorthologs in the phylogenetic analyses due to problems with family classification or sequence alignment. At the regulatory level, we may have failed to recognize gene pairs that were preserved by quantitative subfunctionalization, or for stoichiometric reasons, due to insufficient sensitivity of the in situ and microarray analyses. In addition, sub- and neofunctionalization are not necessarily mutually exclusive, and functional divergence can also occur independent of duplication, as demonstrated by the differences in spatial expression of human–mouse one-to-one orthologs (Huminiecki and Wolfe 2004). Furthermore, the mechanism underlying the initial retention of the duplicate gene copy, such as subfunctionalization, may differ from the processes that have shaped the evolution of the gene pair thereafter, such as acquisition of novel functions. We assumed here that the spatial expression domains and domain architecture of the mouse orthologs can be used to infer those of the ancestral loci before the teleost-specific WGD, but we caution that interpreting current expression patterns and protein domain architectures in light of their past evolution can be misleading.

Our analyses demonstrate that most WGD-duplicates have acquired changes in function, either via changes in regulatory control or via changes in protein function. In both cases we found more examples supporting a model of neofunctionalization, or the evolution of novel functionalities, than one of subfunctionalization, or the partitioning of ancestral functions. We also found that changes at the regulatory level were much more common than changes at the protein level. To our knowledge, this is the first study to investigate, on a large scale and in any vertebrate, the spatiotemporal expression and the protein products of WGD-duplicates. Our aim was to develop a more general understanding of the functional significance of duplicates retained from WGD and of the evolutionary opportunities that polyploidization offers. We have presented here a significant attempt to identify gene duplicates retained from WGD in five teleost genomes as well as information illuminating their functional evolution after duplication. We conclude that, in the teleost lineage, the most significant contribution of WGD has been to allow more-specialized regulatory control of development, typically via the acquisition of novel spatial expression domains. These results contribute to our understanding of vertebrate evolution at both the gene and genome level. Finally, a deeper understanding of the differences in gene regulatory control underlying mammals and teleosts, such as we describe here, will be fundamental to the expanding utility of *D. rerio* and *O. latipes* as model organisms for understanding human development and disease.

Methods

Gen(om)e annotations

The sequences and genomic locations of all protein-coding loci for each of the five fish genomes and human were downloaded from Ensembl v48. These included 21,322 protein-coding loci in *D. rerio* (assembly version 7; K Howe, pers. comm.), 20,121 in *O. latipes* (v1), 20,791 in *G. aculeatus* (v1; F Di Palma and K Lindblad-Toh, pers. comm.), 21,880 in *T. rubripes* (v7), and 27,919 in *T. nigroviridis* (v4). Eighty-seven manually curated HOX and cytokine genes were excluded from the *T. nigroviridis* data set. For human, we excluded genes on alternative genome assemblies and mitochondrial genes, resulting in a total of 22,762 protein-coding loci (assembly NCBI 36). The genebuilds for human, *D. rerio*, *O. latipes*, and *G. aculeatus* were produced by Ensembl; that for *T. nigroviridis* by Genoscope; and that for *T. rubripes* by the Fugu Genome Project.

Similarity searches

All-against-all proteome comparisons between human and fish were carried out using the SSEARCH sequence alignment tool (Pearson 1995). Where multiple protein translations exist for a single locus, only the longest protein translation was used in the analysis. To avoid exclusion of potential homologs, all protein pairs with *E*-values $<1 \times 10^{-3}$ were saved.

Synteny mapping between fish and human genomes

Identification of syntenic or collinear genomic regions in fish and human was performed using the i-ADHoRe synteny mapping tool (Vandepoele et al. 2002; Simillion et al. 2004) using four fish genomes (excluding *T. rubripes*) and the human genome as input. The initial gene (positional) homology matrices were built using the genomic location of unidirectional and reciprocal best hits identified by the proteome comparisons described above. Elements that were homologous in the matrix across genomes are referred to as anchor points. We used the following parameter settings in the i-ADHoRe analyses: gap size (40), cluster gap (150), *q*-value (0.8), probability cutoff (0.001), and minimum number of anchor points (3). This analysis provided a series of genomic anchor points, namely, fish–human positional homologs with preserved gene order and chromosome location delineating syntenic regions between human and fish.

We also tried other criteria for inferring protein homologies and building the gene homology matrices. For example, in a different analysis, we selected all those fish–human protein matches with an *E*-value below 1×10^{-5}, alignment coverage of greater than 50%, and sequence identity of greater than 20%. Based on published literature (McLysaght et al. 2002; Christoffels et al. 2004) and the observed distribution of values for alignment cover and percentage ID across the protein space, these values are conservative estimates for retrieving potential homologs between these species. However, when using putative homologs defined in this manner to build the initial gene homology matrices, we obtained fewer genomic anchor points than when using unidirectional and reciprocal best hits only. In addition, we performed synteny analyses using only pairwise comparisons between one fish genome and the human genome. These pairwise comparisons retrieved fewer genomic anchors than the analysis based on the inclusion of all four fish genomes. Thus, the use of four fish genomes provided additional information for recovering transient and distant colinearity relationships that could not be retrieved when performing only pairwise genome comparisons with human, supporting similar findings from plant genome comparisons (Simillion et al. 2004).

Identifying additional anchors between fish and human genomes

For the comparison of synteny between *T. nigroviridis* and human, we were able to benchmark our analyses against previously published chromosomal sister regions (Jaillon et al. 2004). This comparison showed that we had identified fewer positional homologs than did Jaillon et al. (2004), potentially due to the parameter settings used to build the gene homology matrices being too strict. For this reason, we searched for additional positional homologs, using the known anchors identified by i-ADHoRe as a guide. For each protein-coding human locus, we determined whether i-ADHoRe had identified an anchor within a range of 20 genes along the human chromosome location in question. If so, we took these known anchors and searched the fish chromosome for potential homologs within a neighborhood of 40 genes along the fish chromosome location of the known anchor. We repeated this process for each of the fish–human comparisons. All protein pairs with an *E*-value below 1×10^{-5} were used as putative homologs in the search for additional syntenic anchors. This procedure increased the number of genomic anchors by 10%–20% in each of the fish–human genome comparisons.

Where a single human locus had two or more genomic anchors within the fish genome of interest, we extracted these fish–human protein triplets. These fish proteins shared significant sequence similarity with the human protein and were located within genomic regions that shared at least three fish–human homologs in close proximity, suggesting that the two fish proteins were derived from a shared ancestral chromosome by duplication, most likely as a result of the WGD. Many of these human–fish protein pairs were also unidirectional or reciprocal best hits (Supplemental Table 1). This procedure provided us with "synteny predictions" for duplicates retained from WGD (Fig. 1).

Establishing orthology/paralogy relationships using protein family trees

As an independent line of evidence, we used phylogenetic data, in particular the protein family trees of the Ensembl Compara pipeline v48 (available at ftp://ftp.ensembl.org/pub/release-48/emf/ensembl_compara/homologies/), to identify duplicate fish genes likely derived by WGD. We searched all protein family trees for the presence of an ancient duplication node at the base of teleosts, a topology consistent with origin by WGD. We excluded any trees from the analysis that did not contain any tetrapod or *D. rerio* sequences, as well as any trees that did not contain multiple protein sequences of at least one fish taxon. This procedure left us with 4453 of the 27,308 protein family trees. Large families were split into subtrees based on the presence of tetrapod and fish sequences, resulting in a total of 11,488 subtrees. Of these 11,488 subtrees, 522 subtrees did not have a *D. rerio* sequence and 7891 subtrees had only a single sequence per fish taxon (no duplicates). Excluding these subtrees, 3075 subtrees were left in the analysis. In each of these subtrees, we looked for an ancient duplication node basal to the split of *D. rerio* and the other teleosts. This strategy allowed us to identify instances where a fish taxon contained sequences in both clades derived from the ancient duplication. These sequences were considered to show a tree topology consistent with origin by WGD, and were extracted as our set of "phylogenetic duplicate predictions" (Fig. 1).

Gene-set enrichment analysis of WGD-duplicates

For gene-set enrichment analysis, we used the GO annotation file for zebrafish "gene_association.zfin" downloaded on 24/4/08 from the Gene Ontology site at http://www.geneontology.org/GO.current.annotations.shtml. In order to retrieve GO annotation information from the ZFIN gene association file, the Ensembl peptide IDs corresponding to these duplicate gene loci were mapped to ZFIN IDs using the Ensembl BioMart tool http://www.ensembl.org/biomart/martview. Analyses of gene function enrichment were performed using the GOstat software tool (Beissbarth and Speed 2004). The method of Benjamini and Hochberg (1995) for controlling the false discovery rate (FDR) was used to control experiment-wise error rates in the face of multiple testing. GO terms with FDR-adjusted *P*-values < 0.1 were considered to be significantly enriched or under-represented among WGD-duplicates compared with the total set of GO annotations for *D. rerio*. Frequently in GO analysis, the most significant GO terms are annotated in the same set of genes as each gene may have several, similar GO annotations. To account for such clusters of similarly annotated genes, we grouped significant GO terms that were annotated in the same set of genes or where one set of genes was a subset of the other, allowing for a maximum of five genes to differ between the annotations of different GO terms (Supplemental Methods).

Spatial analysis of gene expression

Spatial expression data for *D. rerio* were downloaded from ZFIN (http://www.zfin.org); those for the mouse orthologs, from MGI (http://www.informatics.jax.org/). We designed riboprobes for two gene pairs, *zgc:73223* and *g12* and *fbxl14* and *fbxl14a*, that showed interesting temporal expression profiles but for which no expression data were available in ZFIN or the published literature (Supplemental Methods).

Temporal analysis of gene expression

We compared the temporal expression of WGD-duplicates in *D. rerio* during embryogenesis using two microarray data sets, Wilkins et al. (S Wilkins, M Kerr, M Köppen, B Gardiner, D Taylor, C Simons, M Landsberg, S Grimmond, C Heisenberg, and A Perkins,) and Mathavan et al. (2005). Both expression analyses were performed on the Compugen 16K *D. rerio* oligonucleotide microarray. This array contains 16,399 oligos (65-mers) representing 15,806 unique *D. rerio* gene clusters plus controls. To identify which of our 615 gene pairs in *D. rerio* were represented by unique probes on the microarray, we used the *exonerate* sequence alignment tool (Slater and Birney 2005) to map between zebrafish cDNA sequences (downloaded from Ensembl) and microarray probes, using a score threshold of 262 and allowing a maximum of seven mismatches between transcript and 65-mer probe. This procedure identified 445 genes that were represented by one or more Compugen probes with less than seven mismatches, and these probes did not match any other genes with the same mismatch setting. These 445 genes correspond to 67 gene pairs. The similarity between the expression profiles of duplicate gene copies was assessed using Pearson's correlation coefficient (*r*). To determine a suitable threshold *r*-value below which duplicate gene copies could be considered divergent in expression, we generated 100,000 random probe pairs from each of the two microarray data sets using sampling with replacement. Because two genes represented by a randomly selected probe pair are generally not functionally related and hence not coregulated, any gene pair with *r* greater than 95% of random probe pairs can be considered significantly coregulated at $P = 0.05$, following the same strategy as Blanc and Wolfe (2004).

Protein domain architecture and subcellular localization

To test for the acquisition or loss of a protein domain, all protein sequences encoded by WGD-duplicates in *D. rerio* were scanned against the Pfam-A database (Bateman et al. 2004) using the InterProScan tool (Zdobnov and Apweiler 2001). As a reference for comparison, we also performed Pfam-A protein domain analyses on 5878 zebrafish–mouse orthologs that were identified as one-to-one orthologs using the Ensembl Compara v53 database (list available on request). Protein subcellular localization was assessed using the web-based TMHMM Server v2.0 (Krogh et al. 2001), available at http://www.cbs.dtu.dk/services/TMHMM/, and SignalP-HMM and SignalP-NN applications using the web-based SignalP Server v3.0 (Bendtsen et al. 2004), available at http://www.cbs.dtu.dk/services/SignalP/. Predictions were combined using the schema established by Davis et al. (2006b). For the comparison of protein subcellular localization, we compared gene pairs only if HMM and NN methods resulted in the same signal peptide prediction. We defined subcellular localization by five categories, following the classification of Davis et al. (2006a): (1) soluble intracellular protein, no signal peptide and no transmembrane domains; (2) soluble secreted protein, signal peptide and no transmembrane domains; (3) type I membrane protein, signal peptide and a single transmembrane domain; (4) type II membrane protein, no signal peptide and a single transmembrane domain; and (5) multi-spanning membrane protein, multiple transmembrane domains.

Acknowledgments

We acknowledge financial support from the ARC Centre of Excellence in Bioinformatics (ARC CE0348221) and an award under the Merit Allocation Scheme on the APAC/NCI National Supercomputing Facility. The microarray work in this article was supported by the NH&MRC Australian Zebrafish Phenomics Facility (AZPF) grant no. 455871. We acknowledge the support of an ARC LIEF grant, LE0347607, which enabled purchase of the Compugen 16k probe set. We thank Cédric Simillion for access to the i-ADHoRe software, Simon Wong for assistance with the identification of WGD-duplicates, Melissa Davis for assistance with the analyses of protein subcellular localization, Andrés Esteban Marcos for help with the visualization of synteny maps, Milena Gongora and Mark Crowe of the SRC Microarray Facility for assistance with annotating the Compugen probe set, Aliesha Griffin for assistance with in situs, Angela Lawton for zebrafish husbandry, and Kim-Anh Lê Cao and Emanuelle Meugnier for statistical advice. We also acknowledge access to the zebrafish genome sequence, which is sequenced and annotated by The Wellcome Trust Sanger Institute and funded by the Wellcome Trust (http://www.sanger.ac.uk/Projects/D_rerio/). We thank the Broad Institute Genome Sequencing Platform and Genome Sequencing and Analysis Program, Federica Di Palma, and Kerstin Lindblad-Toh for making the data for *Gasterosteus aculeatus* available.

References

Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL, et al. 1998. Zebrafish *hox* clusters and vertebrate genome evolution. *Science* **282:** 1711–1714.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia J-m, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, et al. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297:** 1301–1310.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, et al. 2004. The Pfam protein families database. *Nucleic Acids Res* **32:** D138–D141.

Beissbarth T, Speed TP. 2004. GOstat: Find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics* **20:** 1464–1465.

Bendtsen JD, Nielsen H, von Heijne G, Brunak S. 2004. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340:** 783–795.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* **57:** 289–300.

Blanc G, Wolfe KH. 2004. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* **16:** 1679–1691.

Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y. 2006. The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol* **7:** R43. doi: 10.1186/gb-2006-7-5-r43.

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.

Bult C, Eppig JT, Kadin JA, Richardson JE, Blake JA. 2008. The Mouse Genome Database (MGD): Mouse biology and model systems. *Nucleic Acids Res* **36:** D724–D728.

Christoffels A, Koh EGL, 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.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: How duplicated genes find new functions. *Nat Rev Genet* **9:** 938–950.

Cresko WA, Yan YL, Baltrus DA, Amores A, Singer A, Rodriguez-Mari A, Postlethwait JH. 2003. Genome duplication, subfunction partitioning, and lineage divergence: Sox9 in stickleback and zebrafish. *Dev Dyn* **228:** 480–489.

Crow KD, Wagner GP. 2006. What is the role of genome duplication in the evolution of complexity and diversity? *Mol Biol Evol* **23:** 887–892.

Davis JC, Petrov DA. 2005. Do disparate mechanisms of duplication add similar genes to the genome? *Trends Genet* **21:** 548–551.

Davis MJ, Hanson KA, Clark F, Fink JL, Zhang F, Kasukawa T, Kai C, Kawai J, Carninci P, Hayashizaki Y, et al. 2006a. Differential use of signal peptides and membrane domains is a common occurrence in the protein output of transcriptional units. *PLoS Genet* **2:** e46. doi: 10.1371/journal.pgen.0020046.

Davis MJ, Zhang F, Yuan Z, Teasdale RD. 2006b. MemO: A consensus approach to the annotation of a protein's membrane organization. *In Silico Biol* **6:** 387–399.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* **3:** e314. doi: 10.1371/journal.pbio.0030314.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531–1545.

He X, Zhang J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169:** 1157–1164.

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.

Huminiecki L, Wolfe KH. 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Res* **14:** 1870–1879.

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431:** 946–957.

Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K, Kasai Y, et al. 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447:** 714–719.

Kawahara R, Miya M, Mabuchi K, Lavoué S, Inoue JG, Satoh TP, Kawaguchi A, Nishida M. 2008. Interrelationships of the 11 gasterosteiform families (sticklebacks, pipefishes, and their relatives): A new perspective based on whole mitogenome sequences from 75 higher teleosts. *Mol Phylogenet Evol* **46:** 224–236.

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

Kleinjan DA, Bancewicz RM, Gautier P, Dahm R, Schonthaler HB, Damante G, Seawright A, Hever AM, Yeyati PL, van Heyningen V, et al. 2008. Subfunctionalization of duplicated zebrafish *pax6* genes by *cis*-regulatory divergence. *PLoS Genet* **4:** e29. doi: 10.1371/journal.pgen.0040029.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J Mol Biol* **305:** 567–580.

Le Comber SC, Smith C. 2004. Polyploidy in fishes: Patterns and processes. *Biol J Linn Soc* **82:** 431–442.

Lynch M. 2002. Gene duplication and evolution. *Science* **297:** 945–947.

Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154:** 459–473.

Lynch M, O'Hely M, Walsh B, Force A. 2001. The probability of preservation of a newly arisen gene duplicate. *Genetics* **159:** 1789–1804.

MacCarthy T, Bergman A. 2007. The limits of subfunctionalization. *BMC Evol Biol* **7:** 213. doi: 10.1186/1471-2148-7-213.

Mathavan S, Lee SG, Mak A, Miller LD, Murthy KR, Govindarajan KR, Tong Y, Wu YL, Lam SH, Yang H, et al. 2005. Transcriptome analysis of zebrafish embryogenesis using microarrays. *PLoS Genet* **1:** 260–276.

McLysaght A, Hokamp K, Wolfe KH. 2002. Extensive genomic duplication during early chordate evolution. *Nat Genet* **31:** 200–204.

Metscher BD, Ahlberg PE. 1999. Zebrafish in context: Uses of a laboratory model in comparative studies. *Dev Biol* **210:** 1–14.

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

Miya M, Takeshima H, Endo H, Ishiguro NB, Inoue JG, Mukai T, Satoh TP, Yamaguchi M, Kawaguchi A, Mabuchi K, et al. 2003. Major patterns of higher teleostean phylogenies: A new perspective based on 100 complete mitochondrial DNA sequences. *Mol Phylogenet Evol* **26:** 121–138.

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

Pasterkamp RJ, Kolodkin AL. 2003. Semaphorin junction: Making tracks toward neural connectivity. *Curr Opin Neurobiol* **13:** 79–89.

Pearson WR. 1995. Comparison of methods for searching protein sequence databases. *Protein Sci* **4:** 1145–1160.

Postlethwait JH, Woods IG, Ngo-Hazelett P, Yan Y-L, Kelly PD, Chu F, Huang H, Hill-Force A, Talbot WS. 2000. Zebrafish comparative genomics and the origins of vertebrate chromosomes. *Genome Res* **10:** 1890–1902.

Postlethwait J, Amores A, Cresko W, Singer A, Yan Y-L. 2004. Subfunction partitioning, the teleost radiation and the annotation of the human genome. *Trends Genet* **20:** 481–490.

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

Roth C, Rastogi S, Arvestad L, Dittmar K, Light S, Ekman D, Liberles DA. 2007. Evolution after gene duplication: Models, mechanisms, sequences, systems, and organisms. *J Exp Zoolog B Mol Dev Evol* **308B:** 58–73.

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.

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

Sémon M, Wolfe KH. 2008. Preferential subfunctionalization of slow-evolving genes after allopolyploidization in *Xenopus laevis*. *Proc Natl Acad Sci* **105:** 8333–8338.

Sharma MK, Liu R-Z, Thisse C, Thisse B, Denovan-Wright EM, Wright JM. 2006. Hierarchical subfunctionalization of f*abp1a, fabp1b* and *fabp10* tissue-specific expression may account for retention of these duplicated genes in the zebrafish (*Danio rerio*) genome. *FEBS J* **273:** 3216–3229.

Simillion C, Vandepoele K, Saeys Y, Van de Peer Y. 2004. Building genomic profiles for uncovering segmental homology in the twilight zone. *Genome Res* **14:** 1095–1106.

Slater G, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6:** 31. doi: 10.1186/1471-2105-6-31.

Sprague J, Bayraktaroglu L, Clements D, Conlin T, Fashena D, Frazer K, Haendel M, Howe DG, Mani P, Ramachandran S, et al. 2006. The Zebrafish Information Network: The zebrafish model organism database. *Nucleic Acids Res* **34:** D581–D585.

Taylor JS, Braasch I, Frickey T, Meyer A, de Peer YV. 2003. Genome duplication, a trait shared by 22,000 species of ray-finned fish. *Genome Res* **13:** 382–390.

Van de Peer Y. 2004. Computational approaches to unveiling ancient genome duplications. *Nat Rev Genet* **5:** 752–763.

Vandepoele K, Saeys Y, Simillion C, Raes J, Van de Peer Y. 2002. The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. *Genome Res* **12:** 1792–1801.

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* **101:** 1638–1643.

Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449:** 54–61.

Wilkins SJ, Yoong S, Verkade H, Mizoguchi T, Plowman SJ, Hancock JF, Kikuchi Y, Heath JK, Perkins AC. 2008. Mtx2 directs zebrafish morphogenetic movements during epiboly by regulating microfilament formation. *Dev Biol* **314:** 12–22.

Wong S, Butler G, Wolfe KH. 2002. Gene order evolution and paleopolyploidy in hemiascomycete yeasts. *Proc Natl Acad Sci* **99:** 9272–9277.

Woods IG, Wilson C, Friedlander B, Chang P, Reyes DK, Nix R, Kelly PD, Chu F, Postlethwait JH, Talbot WS. 2005. The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Res* **15:** 1307–1314.

Woolfe A, Elgar G. 2007. Comparative genomics using *Fugu* reveals insights into regulatory subfunctionalization. *Genome Biol* **8:** R53.

Yu H-H, Moens CB. 2005. Semaphorin signaling guides cranial neural crest cell migration in zebrafish. *Dev Biol* **280:** 373–385.

Zdobnov EM, Apweiler R. 2001. InterProScan: An integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17:** 847–848.

Received December 19, 2008; accepted in revised form April 22, 2009.