THE ROYAL SOCIETY

doi 10.1098/rstb.2001.0975

Comparative genomics provides evidence for an ancient genome duplication event in fish

John S. Taylor, Yves Van de Peer, Ingo Braasch and Axel Meyer*

Department of Biology, University of Konstanz, 78457, Konstanz, Germany

There are approximately 25 000 species in the division Teleostei and most are believed to have arisen during a relatively short period of time *ca.* 200 Myr ago. The discovery of 'extra' *Hox* gene clusters in zebrafish (*Danio rerio*), medaka (*Oryzias latipes*), and pufferfish (*Fugu rubripes*), has led to the hypothesis that genome duplication provided the genetic raw material necessary for the teleost radiation. We identified 27 groups of orthologous genes which included one gene from man, mouse and chicken, one or two genes from tetraploid *Xenopus* and two genes from zebrafish. A genome duplication in the ancestor of teleost fishes is the most parsimonious explanation for the observations that for 15 of these genes, the two zebrafish orthologues are sister sequences in phylogenies that otherwise match the expected organismal tree, the zebrafish gene pairs appear to have been formed at approximately the same time, and are unlinked. Phylogenies of nine genes differ a little from the tree predicted by the fish-specific genome duplication hypothesis: one tree shows a sister sequence relationship for the zebrafish genes but differs slightly from the expected organismal tree and in eight trees, one zebrafish gene is the sister sequence to a clade which includes the second zebrafish gene and orthologues from *Xenopus*, chicken, mouse and man. For these nine gene trees, deviations from the predictions of the fish-specific genome duplication hypothesis are poorly supported. The two zebrafish orthologues for each of the three remaining genes are tightly linked and are, therefore, unlikely to have been formed during a genome duplication event. We estimated that the unlinked duplicated zebrafish genes are between 300 and 450 Myr. Thus, genome duplication could have provided the genetic raw material for teleost radiation. Alternatively, the loss of different duplicates in different populations (i.e. 'divergent resolution') may have promoted speciation in ancient teleost populations.

Keywords: genome duplication; speciation; phylogenetics; zebrafish (*Danio rerio*); comparative genomics

1. INTRODUCTION

Major transitions, including the evolution of eukaryotes, metazoans, Bilateria and Vertebrata, may have required the genetic raw material provided by gene and/or genome duplications (Ohno 1970; Lundin 1993, 1999; Sidow 1996; Holland 1999; Patel & Prince 2000). Ohno (1970) presented comparative data on genome size and chromosome numbers to support his hypothesis that one or more genome duplications preceded the evolution of vertebrates. Ohno further proposed that the new redundant genes produced by genome duplication evolved new functions that were necessary for vertebrate evolution. The apparent functional connection between duplicate genes and the evolution of vertebrates was more fully asserted by Holland (1992). In mice, paralogues *Hox-1.5* and *Hox-1.6* (renamed *HoxA3* and *HoxA1* respectively—De Robertis 1994) have overlapping expression domains and are at least partially functionally redundant. Holland proposed that overlapping expression domains among paralogous genes (Fitch 1970) delimit the expression domain of their single ancestral gene and that non-overlapping expression domains represent post-duplication gains of function. Holland (1992) also suggested that post-duplication gains of function, particularly in *Hox* genes, facilitated the evolution of vertebrate-specific features such as the control of neural crest cell fate and organogenesis, hindbrain differentiation and otic morphogenesis. It is clear that duplicated genes can evolve previously non-existent functions. Expansion of repetitive regions in one copy of a duplicated pancreatic trypsinogen-like gene produced a gene for antifreeze glycoproteins in Antarctic fish (Cheng & Chen 1999) and mutations in duplicated opsin genes led to the evolution of trichromatic vision in New and Old World primates (Dulai *et al.* 1999). However, the causal link between gene duplication and major evolutionary transitions remains a matter of speculation.

Ohno's hypothesis that big leaps in evolution required the creation of new gene loci with previously non-existent functions emphasized genome duplication via tetraploidy as the mechanism for the production of new genes. Gene number comparisons support this model. Spring (1997) uncovered an average of three orthologous genes in humans for each of 52 *Drosophila* genes and proposed that the additional human genes were produced during two genome duplications. However, Spring's hypothesis, which has recently been referred to as the 'one to four rule' (Ohno 1999) and the '2R' hypothesis (Hughes 1999*a*), remains highly controversial (Hughes 1999*a*; Wang & Gu 2000).

*Author for correspondence (axel.meyer@uni-konstanz.de).

Genome duplication in Actinopterygii (ray-finned fishes) is the focus of this study. The recent discovery of 'extra' *Hox* gene clusters in zebrafish (*Danio rerio*) and pufferfish (*Takifugu rubripes*) led Amores *et al.* (1998) to the conclusion that a chromosome doubling event, probably by whole genome duplication, occurred after the divergence of ray-finned and lobe-finned fishes. *Hox* genes encode DNA-binding proteins and occur in one or more clusters of up to 13 genes per cluster. In Sarcopterygii (a monophyletic group including lobe-finned fishes, amphibians, reptiles, and mammals) there appear to be four *Hox* clusters labelled A, B, C and D with each cluster occurring on a different chromosome. In contrast, zebrafish possess at least seven *Hox* clusters and the pufferfish has two '*Hox A*' clusters (Amores *et al.* 1998; Aparicio 2000). As in sarcopterygians, fish *Hox* clusters occur on different chromosomes. Following Amores *et al.*'s (1998) conclusion that genome duplication was the explanation for the 'extra' *Hox* clusters in fish, Meyer & Schartl (1999) expanded the 'one to four rule' to the 'one to four to eight rule' to account for this additional genome duplication. Teleostei is the most diverse of all vertebrate groups and includes approximately 25 000 species (Nelson 1994). Major teleost lineages are believed to have arisen between *ca.* 100 and 200 Myr ago (Carroll 1997; Lydeard & Roe 1997) and Amores *et al.* (1998) and Meyer & Schartl (1999) proposed that genome duplication facilitated this radiation.

Stellwag (1999) suggested that, with respect to *Hox* cluster number, the zebrafish is not representative of actinopterygians and that the genome duplication proposed by Amores *et al.* (1998) might be limited to only a few derived fish or even the zebrafish lineage alone. This argument was weakened when it was discovered that medaka (*Oryzias latipes*), which is placed in a different teleost superorder than zebrafish, also possess seven *Hox* clusters (Naruse *et al.* 2000). Other criticisms of the teleost genome duplication hypothesis have focused on the fact that *Hox* genes reveal the history of only a small portion of the entire genome. Most fishes have smaller genomes than humans (Ohno 1970; Hinegardner & Rosen 1972). The zebrafish genome is approximately half the size of the human genome (Hinegardner & Rosen 1972). Morizot *et al.* (1991) estimated that the genome of the platyfish (*Xiphophorus*) is five times smaller than the human genome and Elgar *et al.* (1999) estimated that the pufferfish genome is eight times smaller than the human genome. Although genome size and gene content may not be correlated, Elgar *et al.* (1999) suggested that the duplication of *Hox* clusters by regional duplication is easier to reconcile with fish genome size data than genome duplication.

The goal of our study was to use a phylogenetic approach to evaluate the hypothesis that the 'extra' *Hox* genes and the rest of the genome in fishes were produced during a genome duplication in a teleost ancestor rather than by a series of regional duplications. The genome duplication hypothesis makes clear predictions about the number of genes in fishes compared with humans and about the topology of gene trees: a gene tree should match the expected organismal tree but have two zebrafish orthologues for each human gene and the zebrafish orthologues should be sister sequences in a phylogenetic analysis (figure 1). We refer to this predicted topology as the 'duplication topology'. Furthermore, pairs of zebrafish orthologues from different genes should have been formed at the same time and should be unlinked.

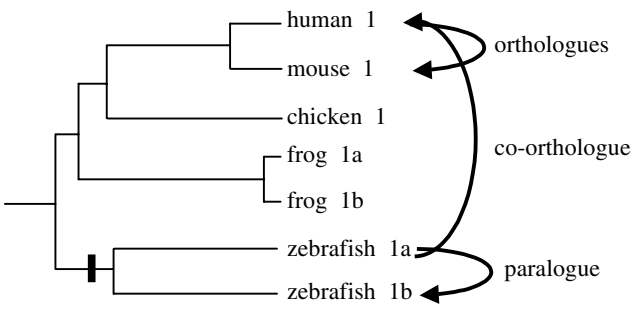

Figure 1. Phylogenetic topology predicted assuming the ancestor of actinopterygian fishes experienced a genome duplication. This topology, referred to as the 'duplication topology', also assumes that no genes have been lost in the taxa surveyed. Supplements to the term homology are described in the figure: 'orthology' (Fitch 1970) describes the relationship between homologous genes (i.e. genes descended from a common ancestral gene) that occur in different species; 'paralogy' (Fitch 1970) describes the relationship between homologous genes that occur within an individual (e.g. genes produced by genome or by tandem duplication). Duplicated zebrafish genes are 'co-orthologues' of their human orthologues (Gates *et al.* 1999).

Human and zebrafish protein sequences were obtained from the non-redundant (NR) protein database at the National Center for Biotechnology Information (NCBI, Bethesda, MD, USA) to determine whether gene numbers and gene phylogenies support the fish-specific duplication hypothesis. We also collected sequences from *Mus musculus*, *Gallus gallus* and *Xenopus laevis* so that we could reconstruct the reliable phylogenies necessary to identify orthologues among the sequences retrieved in our basic local alignment search tool (BLAST) searches. Map data are available for most of the zebrafish genes in our survey and we used these data to determine whether anciently duplicated genes are distributed throughout the zebrafish genome.

2. METHODS

(a) *Database searches*

Protein sequences of zebrafish (*Danio rerio*), human (*Homo sapiens*), mouse (*Mus musculus*), chicken (*Gallus gallus*) and the African clawed frog (*Xenopus laevis*) were obtained by BLASTp (Altschul *et al.* 1990). For all searches we selected the NR search option (see http:/www.ncbi.nlm.nih.gov/blast/html/blastcgihelp.html#nucleotide_databases). With a few exceptions, human 'reference sequences' (Maglott *et al.* 2000) were used as BLASTp query sequences. Most genes surveyed were those used in a gene number comparison between *Drosophila* and humans (Spring 1997), but the mammalian genes that Gates *et al.* (1999) describe as having two zebrafish orthologues were also included. Species were surveyed one at a time to improve the identification of a drop in sequence similarity, which was used as a 'cut-off'. Sequences above the cut-off value were pasted to NCBI clipboards and then downloaded in FASTA format, a format that includes the sequence definition line and sequence characters.

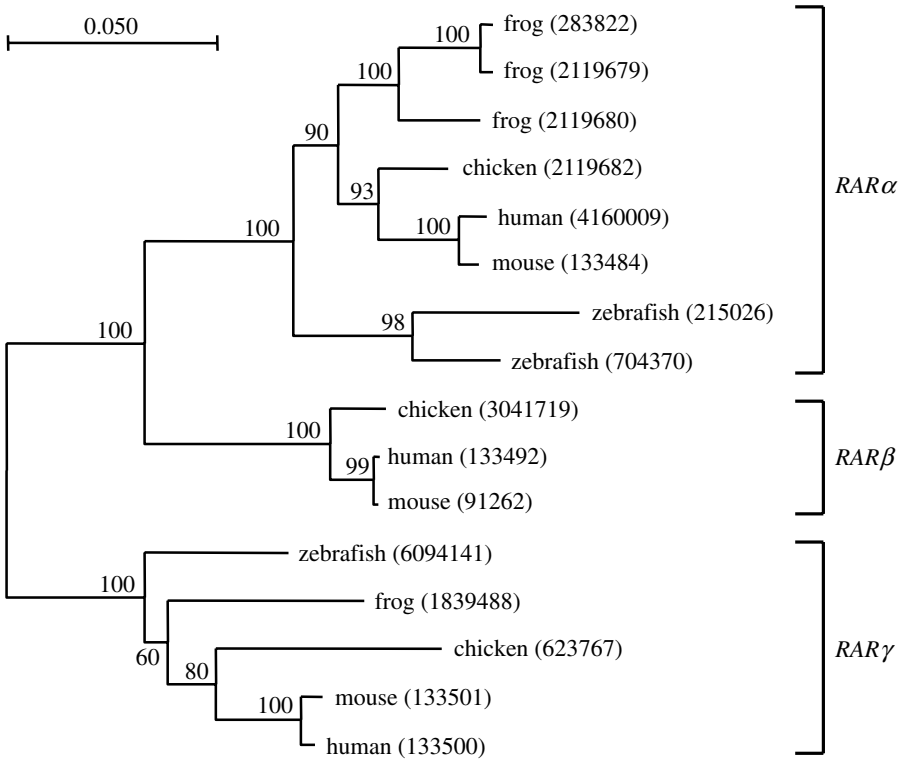

Figure 2. Neighbour-joining tree of the retinoic acid receptor genes retrieved using BLASTp (gene identification numbers shown). Sequences that varied only in length or by very few amino-acid substitutions were removed prior to analysis (see § 2). The tree shows paralogous clades of *RARα*, *RARβ*, and *RARγ* genes. Bootstrap values (Felsenstein 1985) are shown (500 bootstrap reiterations).

(b) *Sequence alignment and phylogeny reconstruction*

When BLASTp identified one or more putative zebrafish orthologues, protein sequences from all species were aligned using CLUSTALX (Thompson *et al*. 1997). For each alignment, a preliminary tree was drawn from the CLUSTAL dendrogram file using TREEVIEW v. 1.6.0 (Page 1996). This tree facilitated the identification of identical sequences, sequences that varied only in length, and sequences within species that differed by few amino acids, all of which were removed from the alignment. Very similar sequences could be alleles at one locus or evidence of recent tandem duplications. In either case they were not likely to be important for our study of genome duplication in the teleost ancestor.

Phylogenies were reconstructed from the remaining sequences using Poisson-corrected genetic distances and the neighbour-joining (NJ) algorithm (Saitou & Nei 1987) in TREECON (Van de Peer & De Wachter 1994). These first NJ phylogenies included many clades of orthologous and paralogous genes (e.g. figure 2). From these large trees we identified sets of orthologous genes (i.e. genes which occurred in monophyletic groups that matched the expected organismal topology). Sequences of orthologous genes were realigned and edited using BIOEDIT (http://www.mbio.ncsu.edu/RNaseP/info/programs/BIOEDIT/bioedit.html). Regions where the alignment was unambiguous were retained and reanalysed using NJ and maximum likelihood (ML) methods. For these last phylogenetic analyses the most closely related human paralogues (identified from the first NJ analyses) were used as outgroups. Support for nodes was evaluated by 500 bootstrap reiterations (Felsenstein 1985). TREE-PUZZLE v. 5.0 (Strimmer & Von Haeseler 1996) was used to reconstruct ML trees (substitution models were selected for each analysis automatically by the program).

(c) *Dating duplication events*

In order to estimate the age of zebrafish paralogues, the number of nucleotide substitutions at third codon positions was plotted against divergence dates for different taxa (Nei & Kumar 2000). Since most third-codon position substitutions do not result in amino-acid replacements, the rate of fixation of these substitutions is expected to be relatively constant in different protein-coding genes (e.g. Nei *et al*. 2000) and to reflect the overall mutation rate (Hughes 1999*b*). Alternatively, one can use the number of synonymous substitutions per synonymous sites to estimate divergence times (Nei & Kumar 2000; Nei *et al*. 2000). However, for the genes surveyed here, there is an approximately linear relationship between the number of third-position substitutions and the number of synonymous substitutions and therefore both approaches are expected to give similar results. Estimation of the number of substitutions at third-codon positions, corrected for multiple events per site according to Tajima & Nei (1984), was done for 26 pairs of genes (no DNA sequence was available for the two zebrafish *GDF6* genes). All computations were done with the software package MEGA2 (Nei & Kumar 2000).

Divergence dates between different taxa were taken from literature and were as follows: genome duplication in *Xenopus*, 30 Myr ago (Hughes & Hughes 1993); divergence between human and mouse, 100 Myr ago (Li *et al*. 1990; Kumar & Hedges 1998); divergence between reptiles (represented by the bird *Gallus gallus*) and mammals, 310 Myr ago (Kumar & Hedges 1998); divergence between amphibians and amniotes, 360 Myr ago (Kumar &

Hedges 1998); and divergence between ray-finned fish and Sarcopterygii, 450 Myr ago (Kumar & Hedges 1998).

3. RESULTS

(a) *Gene numbers and phylogenetic analyses*

BLASTp searches uncovered a large number of sequences for each species, many of which differed only in length or by very few amino-acid replacements. Neighbour-joining analyses of the longest sequences often identified many (up to 15) different monophyletic groups of orthologous genes (e.g. figure 2). Groups of orthologous and paralogous genes analysed together are listed together in different blocks in table 1. Groups of orthologous genes within these clades are presented on separate rows within blocks in table 1.

Variation in the length of sequences in different species meant that for some genes a large proportion of the available data could not be used for phylogenetic analyses. Furthermore, sequence variation among taxa meant that large portions of some sequences could not be unambiguously aligned.

For 27 genes, NJ analyses produced a well-supported clade with two zebrafish genes, one human, mouse and/or chicken gene and one or two *Xenopus* genes. Eighteen of these 27 trees had the 'duplication topology' (figure 3*a*). In one tree (*EN2*) zebrafish genes are sister sequences but, unexpectedly, they cluster with the two *Xenopus* genes (figure 3*a*). For eight trees (figure 3*b*) one of the two zebrafish genes was the sister sequence to a monophyletic group that included the second zebrafish gene and orthologous genes from *Xenopus*, chicken, mouse and human. Phylogenies of the eight genes shown in figure 3*b* have the 'outgroup topology'. Eighteen of the 19 genes with zebrafish orthologues as sister sequences using NJ methods also had this sister sequence relationship when ML methods were used (for *ISL2*, ML analyses produce the 'outgroup topology'). Among the eight genes in figure 3*b*, ML analysis produced the 'duplication topology' for *FKD5*, *HOXC6* and *SOX11*. Maximum likelihood analyses of *SNAP25* data supported the hypothesis that the two zebrafish genes (*snap25,1* and *snap25,2*) were sister sequences, but the zebrafish, mouse and human *SNAP25* sequence did not form a monophyletic group when ML methods were used. Both phylogenetic methods produced the 'outgroup topology' for four genes (*DLX2*, *JAK2*, *NTN1* and *OTX1*).

Bootstrap support for the duplication topology or the outgroup topology was low for some trees in figure 3, even when the same topology was produced by both phylogenetic methods. To test whether the tree topologies shown in figure 3 were significantly better than the alternative topology, we performed a Kishino–Hasegawa test (Kishino & Hasegawa 1989) as implemented in TREE-PUZZLE (Strimmer & Von Haeseler 1996). As already might have been expected on the basis of the bootstrap analysis, user-defined trees where the two zebrafish genes are sister sequences were not found to be significantly worse than the *DLX2*, *JAK2*, *NTN1* and *SOX11* trees shown in figure 3*b*. However, our application of the Kishino–Hasegawa test also produced unexpected results. The Kishino–Hasegawa test failed to reject the 'outgroup topology' in many cases even when NJ and ML analyses produced the 'duplication topology' with high bootstrap support. For these genes the likelihood of a sister sequence relationship between zebrafish paralogues (i.e. the 'duplication topology') was always the highest, but the 'outgroup topology' was not significantly worse. The Kishino–Hasagawa test appears to have low resolving power for our datasets, which may be too conserved and include too few samples (A. von Haeseler, personal communication).

(b) *The age of the duplicated genes*

To estimate the date of the fish-specific duplication, we plotted known divergence dates between different taxa against the number of nucleotide substitutions at third-codon positions (see § 2). Although we initially included the split between ray-finned fish (Actinopterygii) and Sarcopterygii, this divergence and the corresponding number of substitutions between zebrafish and the other vertebrates were omitted from the final analysis since the nucleotide substitutions at third codon positions were clearly saturated (not shown). This is probably also true for the amphibian–amniote divergences (as shown by the large differences in number of substitutions; figure 4) and to some extent for the divergence between the chicken and mammals (Nei & Kumar 2000). However, based on the plot of figure 4, complete saturation probably does not occur much earlier.

Divergence dates for different vertebrate lineages are controversial and may differ considerably whether based on palaeontological or molecular calibration (Kumar & Hedges 1998; Gu 1998; Lee 1999). Nevertheless, if we consider the dates used as reliable, and using 1.02 (s.d. = 0.24) as the average number of substitutions per site between the 23 pairs of unlinked zebrafish co-orthologues (see below), the fish-specific genome duplication occurred *ca.* 350 Myr ago. Since the third codon positions have probably reached saturation, as indicated by the high number of estimated substitutions per site when both zebrafish genes are compared, this calculation is at the limit of our ability to estimate dates. In conclusion, the fish-specific genome duplication is probably older than 300 million years, if we assume that third-codon positions are not completely saturated at the time of the reptilian–mammalian divergence. Furthermore, assuming that the genome duplication is not older than the divergence of the Actinopterygii and Sarcopterygii, the duplication probably occurred between 300 and 450 Myr ago.

(c) *Map positions*

Zebrafish co-orthologues shown in figure 3 are distributed among 16 of the 25 zebrafish linkage groups (table 2). For *DLL* and *MSX3*, one co-orthologue occurs on linkage group (LG) 1 and the other on LG13, and for *DLX2* and *EN1*, one zebrafish co-orthologue occurs on LG1 and the other on LG9. For *EN2* and *SHH*, one zebrafish co-orthologue occurs on LG2 and the other on LG7. For *BMP2*, *SNAP25* and *SOX11* one co-orthologue occurs on LG17 and the other occurs on LG20. Lastly, for three genes (*HOXB5*, *HOXB6* and *RARα*) one co-orthologue occurs on LG3 and the other on LG12. Thus, portions of LG1 and LG13, LG1 and LG9, LG2 and LG7, LG17 and LG20, and LG3 and LG12 appear to be paralogous (table 2).

Table 1. Surveyed genes.

(Blocks separated by blank lines identify families of genes uncovered in BLAST searches and used for tree reconstruction. Rows (some comprised of more than one line) identify genes that are orthologous to a single human gene according to our phylogenetic analyses. Genes with topologies that support the fish-specific genome duplication hypothesis are shaded. '—', no orthologous genes found in databases.)

human gene name	*Homo sapiens*	*Danio rerio*	*Mus musculus*	*Gallus gallus*	*Xenopus laevis*
ABL1	4885045	—	125137	—	—
ABL2	6382060	—	—	—	7248894
ALDOA	4557305	—	7548322	—	1944025
ALDOB	4557307	—	—	113610	—
ALDOC	113613	—	113614	226855	3928511
APP	4502167	8050809	6680708	6465892	320195
APLP1	4885065	—	6680700	—	—
APLP2	4502147	—	1086521	—	—
ANK1	4502089	—	1168457	1245423	—
ANK2	4502091	—	—	1245425	—
ANK3	4502093	—	710549	1245427	—
BMP2	4557369	2804175 2149148	6680794	2501173	115070
BMP4	4502423	2149144	461633	2501175	399122 477512
BMP5	339560	—	6671642	1881823	—
BMP6	4502425	—	6680798	—	—
BMP7	4502427	6573121	—	6970053	4096790
BMP8	4502429	—	6671644	—	—
BRNI (*POU3*-tf2)	5453936	1730449 2495310	6679425	—	—
POU3-tf3 (outgroup)		5031983			
BTK	4557377	—	2507603	—	—
ITK	7949058	2353318	—	—	—
TEC	4507429	—	420220	—	—
TXK	4507743	—	1174826	—	—
CDH1/3/14	4757960	—	—	115417	13432108
	4502721	—	—	416739	13432110
CDH2	14589889	2133885	—	115422	416743 115425
CDH12[a]	2119627	— —	6680904	3023428 2134302	— —
cad7	—	—	7549750	2134303	2119628
cad11	—	1345125	6753372	3511021	3377485
CALM[b]	5901912	—	6680832	3415119	6137739
CALM2[b]	4502549	—			
CALM3[b]	4885109	—			
CDX1	4502763	—	1170313	1170316	435578
CDX2	4502765	—	1170314	1737445	—
CDX4	4885127	283775	1083362	547650	2134077
COL4A1	7656985	—	115312	7271901	—
COL4A3	177894	—	6680968	—	—
COL4A5	4502955	—	2119170	—	—
CTSH	4758096	—	7106279	—	—
CTSK	4503151	—	6681085	1017831	—
CTSL	4503155	1752664	6753558	2144502	2706547
CTSS	4758098	—	3850787 2961621	—	—
Catlrp-p	—	—	5306071	—	—
Catm	—	—	7715970	—	—

continued

Table 1. *continued*

human gene name	*Homo sapiens*	*Danio rerio*	*Mus musculus*	*Gallus gallus*	*Xenopus laevis*
DLL1	10518497	2809389 1888392	6681197	2134296	807696
DELTA4 (outgroup)	8926615				
DLX1	2829447	2842747	6753644	—	—
DLX2	4758168	2842748 1708243	6753646	—	1079297 1708249
DLX3	4885185	1346299	2495277	5830236	2134092 1708245
DLX4	4503343	—	6681201	—	—
DLX5	4885187	1708248	2495278	1708250	2134167
DLX6	4885189	2842749	6014979	—	1708242
DLX7	—	2842750	—	—	—
DLX8	—	2842751	—	—	—
TCF3/E2a	181906	2118448	—	506759	283796
TCF4/E2b	4507399	—	7305551	—	—
TCF12/E2c	4507391	—	346644	416847	—
E2F2	4758226	—	—	—	—
E2F3	4503433	—	3122045	—	—
EGF	4503491	—	6753732	—	—
TGFA	4507461	—	1351229	—	—
HGL	4758526	—	—	9297019	—
AREG	4502199	—	6753100	—	—
DTR	4503413	—	6754178	4761593	—
TDGF1	4507425	8132035	—	—	—
EGFR	4885199	—	1352359	1070476	—
ERBB2	4758298	—	—	—	—
ERBB3	4503597	—	—	—	—
ERBB4	4885215	—	—	4884676	—
EGR1	4503493	1352361	6681285	—	7673684
EGR2	4557549	462005	2507546	—	1169500
EGR3	4758252	—	9055212	—	—
EGR4	4503495	—	4704780	6707678	—
EMX1	31140	2133842	729412	—	—
EMX2	31142	2133843	729414	—	—
EN1	7710119	4322044 417127	7106305	483162	1708255 399907
EN2	7710121	417128 417129	6753752	483259	1708257 1708256
EPA1	2827756	—	—	—	—
EPA2	4758278	3005903	6753758	—	3861464
EPA3	4885211	—	125338	125337	—
EPA4	4758280	3005933	6679657	2833208	8134439 8134440
EPA5	1706628	—	6679659	1706627	—
EPA7	4758282	1754761	2497573	8134447	—
EPA8	7263928	8134436	6679663	—	—
EPB1	2739208	—	—	8134448	8134450 8134449
EPB2	1706664	—	1706665	2827774	2739062
EPB3	4758288	2198795	1708165	2134386	974710
EPB4	4758290	3005901 3163942	6753760	—	6689570 6689572
EPB6	4758292	—	—	2833209	—
EVX1	4503615	4322046	6679711	—	1708342
EVX2	553284	1617040	6679713	—	—
eve1[c]		630922			

continued

Table 1. *continued*

VIL2	4507893	—	6678571	4514720	—
RDX	4506467	—	6677699	6179570	—
MSN	4505257	—	462608	—	6648536
FGFr1	182532	—	309240	120045	214900
FGFr2	4503709	—	2144423	116098	544293
FGFr3	4503711	8886017	477423	116097	2425188
FGFr4	4503713	773667	6679789	—	2541908 1213275
FKD5	8134472	2982343 2982347	2494502	—	3695057
FXL1 (outgroup)	13638268				
FLOT1	5031699	12751185 12751187	6679811	— —	— —
flotillin1 (outgroup)	3115387 (*Dros.*)				
gdf6[d]	—	914116 1906321	1707885 (bovine)	—	5052013
GDF5	1346125	—	742374	4836456	—
GLI1	4885279	—	6009644	2501700	3915716
GLI2	4885277	6554167	—	2564663	2501705 4704617
GLI3	13518032	—	6680021	7141288	2501704
GPC1	4504081	—	—	1707999	—
GPC3	4758462	—	7710030	—	—
GPC4	4504083	—	6680059	—	—
HH (*DHH*)	6166118	6014963	6681181	—	6014961 6014962
(*IHH*)	1581789	1616585	6166227	6016342	6016351
(*SHH*)	4506939	6174983 6136068	6094284	6094281	6175032 530994
HOXA2	6016292	6016291	6754230	585280	—
HOXB2	4504465	—	90630	—	—
HOXA3	6016293	—	2811092	6016301	385342
HOXB3	4504467	6016297 5679191	1708353	1708352	399999
HOXD3	6325469	6016300	1708360	—	—
HOXA5[e]	123225	4322062	6754232	—	—
HOXB5	4504469	123245 4322074	6680251	—	123297
HOXB6	400001	4233076 123250	123253	—	—
HOXC6	4758554	4322098 4322100	1083364	—	123243
HOXA9[e]	6166219	4322064 4322066	6166220	2495322	—
HOXB9	—	4322080	1708355	—	901848
HOXC9	—	4322102	6680255	—	—
HOXD9	7657170	4322104	7305153	123285	—
HOXA10	2822167	2661785	6680243	—	—
HOXB10	—	4322068	—	—	—
HOXC10	—	4322082	400011	—	—
HOXD10	4504471	1731637	7305151	400019	—

continued

Table 1. *continued*

human gene name	*Homo sapiens*	*Danio rerio*	*Mus musculus*	*Gallus gallus*	*Xenopus laevis*
HOXA11[e]	5031759	4322049 1707451	6754226	399992	2995957
HOXC11[e]	7657166	4322084 4322086	—	—	—
HOXD11	400021	974813	123292	400020	—
HOXA13[e]	4504457	4322051 4322053	6680245	—	—
HOXC13[e]	7689387	4322090 4322092	1708359	—	—
ID1	4504569	2253424	2827752	—	—
ID2	4504571	—	109791	2935461	2134185 2134043 4587148
ID3	2135331	—	6680341	—	—
ID4	4504573	—	729812	—	—
INSR	4557884	—	6754360	4588602	5420052
INSRR	186555	—	6754362	—	—
IGF1R	4557665	—	3025894	2808533	1150692 3037089
ISL1	124927	1708559	4469284	1708560	—
isl2	—	1708564 1708561	1708563 (rat)	1708562	—
ITGA2B	4504745	—	7262859	—	—
ITGA5	4504751	—	6754378	—	3183037
ITGA4	4504749	—	—	—	—
ITGB3/4	124968	—	7949057	631019	2119641
ITGB6	9446402	—	4324977	—	—
ITGB7	4504777	—	—	—	—
ITGB1	4504767	—	124964	124962	124961 124965
ITGB2	4557886	—	—	—	—
ITGB5	4504773	—	3478697	—	—
JAK1	4504803	1938358	1708580	4558482	—
TYK2	4507749	—	5733095	—	—
JAK2	4826776	3687398 3687400	6680508	—	—
JAK3	4557681	—	2499670	—	—
LI (*CAM*)	4557707	1065714 1065716	6651057	104799	—
NRCAM (outgroup)	6651380				
LAMA1	34226	—	6678656	1246110	—
LAMA2	4557709	—	2497588	—	—
LAMA3	4557711	—	1922889	—	—
LAMB1	4504951	—	126367	—	—
LAMB2	4504953	—	6678658	2708707	—
LAMB3	4557713	—	6678660	—	—
LHX1	5031867	2497670 2155289	6678688	1708826	267419
Lhx5 (outgroup)			6678690		

continued

Table 1. *continued*

MEF2A	5031907	1518141	7305265	4914481	913313 913312
MEF2C	4505147	1518143	477011	—	—
MEF2D	5174545	1518145	2500877	—	2500878
MSX1	123310	—	11177822	1708273	234375
MSX2	1082306	—	547660	1170325	547691
Msx3	—	399912 2506531	6754756	—	—
MsxD[f]		399913			
MsxA[f]		2506530			
MYOD1	4505309	3914105	6996932	3915780	127711 127053
MYOG	4505311	—	—	—	—
MYOD5	5031929	—	6678982	—	127629
MYH9	189030	—	—	127759	3660672
MYH10	641958	—	—	212449	422615
MYH11	2104553	—	7441402	3915778	—
NFKB1	189180	—	6679044	222839	—
NFKB2	4505383	—	5081604	2134380	3116208
REL	4506473	—	6677707	136185	1004330
REL A	307300	—	6677709	1729913	548721
REL B	5730007	—	6677711	5305228	1710086
NOS1	987662	—	6724321	—	—
NOS2A/B/C	1228940	—	6754872	2498062	—
NOS3	189212	—	—	—	—
NTN1	4758840	2327065 2394302	4732097	2497605	2655297
NTN2 (outgroup)	5453810				
OTX1	417425	3024322 3024327	417426	—	—
OTX2	417427	3024329	417428	—	644782 3024328
OTX5	—	—	—	—	6624755 6252982
PAX2	4557821	3420031 3024368	417447	6683012	5815455 2765055
PAX5 (outgroup)	417449				
PBX1	4505623	7160792	2432009 7110681	8096555 8096557	— —
PBX2	4505625	7160798	—	—	—
PBX3	5453852	7160796	2432017	—	—
PBX4	—	5679283	—	—	—
PTC1	4506247	4539024	6679519	6225890	—
PTC2	4506245	6225889	6679517	—	—
RAF1	4506401	534977	—	125489	125654
ARAF1	4502193	—	125646	—	—
BRAF	4757868	—	—	464647	—
RAN	131845	2500061	6677677	1172839	6729160
RAN (outgroup)	6857182 (*Dros.*)	—	—	—	—
NRAS	4505451	3334308	7242162	—	3334309
HRAS	4885425	—	6680271	31868	—
KRAS2A	131875	—	417590	—	2072749
KRAS2B	131879	—	131880	—	3599487 464552

continued

Table 1. *continued*

human gene name	*Homo sapiens*	*Danio rerio*	*Mus musculus*	*Gallus gallus*	*Xenopus laevis*
RALA	4885569	—	131836	—	—
RALB	4506405	—	—	—	3955067
RARα	4160009	704370 215026	133484	2119682	2119679 2119680 283822
RARβ	133492	—	91262	3041719	—
RARγ	133500	6094141	133501	623767	1839488
RB1	4506435	—	6677679	459445	—
RBL1	4506443	—	2498835	—	—
RBL2	5032029	—	6685841	—	—
RXRA	4506755	1583309	6755384	—	283824
RXRB	1350911	1046299 1046297	1350912	—	1085220 840922
RXRG	5902068	8478106	1350914	133700	1710810
SRC	4885609	—	6678129	6175046	125705
YES1	4885661	—	6678617	125869	321075
FGR	4885235	—	6753860	—	—
FYN	4503823	—	6679879	479367	125371
LCK	4885449	—	2117800	1170731	—
LYN	4505055	—	—	—	2114076
HCK	4504357	—	6754166	—	—
BLK	4502413	—	6680786	—	—
SDC1	4506859	—	6755438	—	2547264
SDC2	386787	—	6677891	—	2547266
SDC4	4506861	—	6755442	1351051	—
SNA 11	5729674	841424 545350	6755586	—	—
SLUG (outgroup)	2832266	—	—	—	—
SNAP25	134583 —	3703098 3703100	6755588	481202 —	—
SNAP23	6685971	—	6678049	—	—
SOX11	4507161	4099263 7572947	6678065	2982742	2522255
SOX4 (outgroup)	4507163				
STAT1	6274552	3687402	6678153	—	—
STAT2	4885615	—	6561853 6014655 5051642	—	—
STAT3	4507253	3687429	1711553	—	6177821
STAT4	4507255	—	1174461	—	—
STAT5a	4507257	—	6755672	4960028	—
STAT5b	6912688	—	7242209	—	—
TNC	4504549	1065718	7106435	135584	—
TNXB	7671639	—	7441741	1419546	—
TNR	5730098	—	—	86419	—

continued

Table 1. *continued*

WNT1	4885655	139740	139744	—	139748
WNT2a	4507927	2501661	139751	—	—
WNT2b	13518017	—	6678591	5901876	3123031
WNT3b	6136371	263558	6678593	5821261	401416
WNT3a	6136340	—	7106447	—	—
WNT11[g]	4759320	7579033 3169687	6678589	1351423	1722841
WNT10b	5803223	263561	6756003	—	—
WNT10a	—	1175018	6678587	6141561	—
WNT6	—	—	227508	—	401424
WNT16	5732946	—	6249635	—	—
	7706773	—	—	—	—
WNT7a	5509901	—	6678603	—	401418
WNT7b	6136361	263560	6678605	1245763	401419
WNT7c	—	—	—	—	401420
WNT5a	4507929	—	6678597	4512218	731158
WNT5b	—	2501662	6678599	—	465484
WNT4[g]	—	1351427 4894948	6678595	1351428	477511

[a] A well supported monophyletic group including human *CDH12*, *Cad6* from *M. musculus*, and two divergent *G. gallus* sequences (*cad10* and *cad6b*) did not show the expected organismal topology (*CDH12* was the 'basal' sequence) and, therefore, may not be true orthologs.
[b] *CALM* genes in the databases for human, mouse, chicken, and frog were identical. Thus, the placement of the mouse, chicken, and frog genes on the same row as *CALM1* is arbitrary.
[c] BLASTp turned up two zebrafish EVX genes. One was the sister sequence of the EVX1+EVX2 clade when Drosophila *even-skipped* (gi 123364) was used to root the tree.
[d] GenBank included a short mouse sequence labelled *Gdf6*. The phylogenetic relationship between this gene and the *GDF6* sequences included in table 1 was not resolved.
[e] For many *Hox* genes, only short conserved sequences that could not be placed within expected clades of orthologs were available (see § 4). Thus, in some cases, *Hox* genes are assigned to rows according to their names.
[f] All *MSX* genes shown formed a well-supported monophyletic group. However, the relationship between zebrafish *msxD* and *msxA* genes and the other *MSX* genes was not resolved.
[g] *WNT4* and *WNT11* genes each form monophyletic groups with two zebrafish genes, but the tree topologies differ significantly from the expected organismal tree and may include two sets of orthologous genes as is the case for *WNT2*, *WNT3*, *WNT5*, *WNT7* and *WNT10* genes.

For *ISL2*, *L1(CAM)* and *PAX2*, zebrafish co-orthologues occur next to one another on the same chromosome (table 2). This observation suggests that duplicated *ISL2*, *L1(CAM)* and *PAX2* genes in zebrafish were formed by tandem duplications. For this reason these three genes were not included in the estimate of the age of the fish-specific genome duplication reported above.

4. DISCUSSION

A genome duplication in the ancestor of teleost fishes is the most parsimonious explanation for the following observations: (i) many genes that occur once in chicken, mouse and man, and twice in *Xenopus*, a tetraploid frog, also occur twice in zebrafish; (ii) the phylogenetic analyses that were necessary to identify the two zebrafish co-orthologues show, in most cases, that zebrafish genes are sister sequences as predicted by the genome duplication hypothesis; (iii) zebrafish co-orthologues are approximately the same age; and (iv) zebrafish co-orthologues are distributed throughout the zebrafish genome.

(a) *Gene number comparisons and gene tree topologies*

The genome duplication hypothesis predicts that zebrafish will have more genes than humans. However, we found 140 cases among the 240 human genes included in our survey in which the database contained no zebrafish orthologues. In a few cases (e.g. *Hox* genes) the shortage of zebrafish orthologues may be an artefact of our inability to assign some genes to specific clades. However, the shortage of fish genes is primarily due to the incomplete nature of the database: NCBI contains 1591 protein entries for zebrafish and 96 009 protein entries for humans (23 November 2000).

Phylogenetic analyses identified 27 genes where orthologues that occur once in man, mouse and chicken, and often twice in *Xenopus*, also occur twice in zebrafish. For all of these genes, monophyly of the two zebrafish genes, plus orthologues from *Xenopus*, chicken, mouse and man, was well supported. For three of these genes, zebrafish co-orthologues are closely linked. Therefore, despite our estimation that they are approximately the same age as the other duplicates, they are unlikely to have been produced by genome duplication. Although not all of the remaining 24 genes had the topology predicted by the fish-specific genome duplication hypothesis, most examples of the 'outgroup topology' are poorly supported by bootstrap reiterations and/or are not present when ML methods are used. A genome duplication event (or many gene duplications) prior to the Sarcopterygii–Actinopterygii divergence might explain the 'outgroup topologies' in figure 3*b*. However, if this is the case, then true orthologues of each of the 'basal' zebrafish genes must have been lost in Sarcopterygii. We believe it is more likely that some or all of the outgroup topologies shown in figure 3*b* are tree

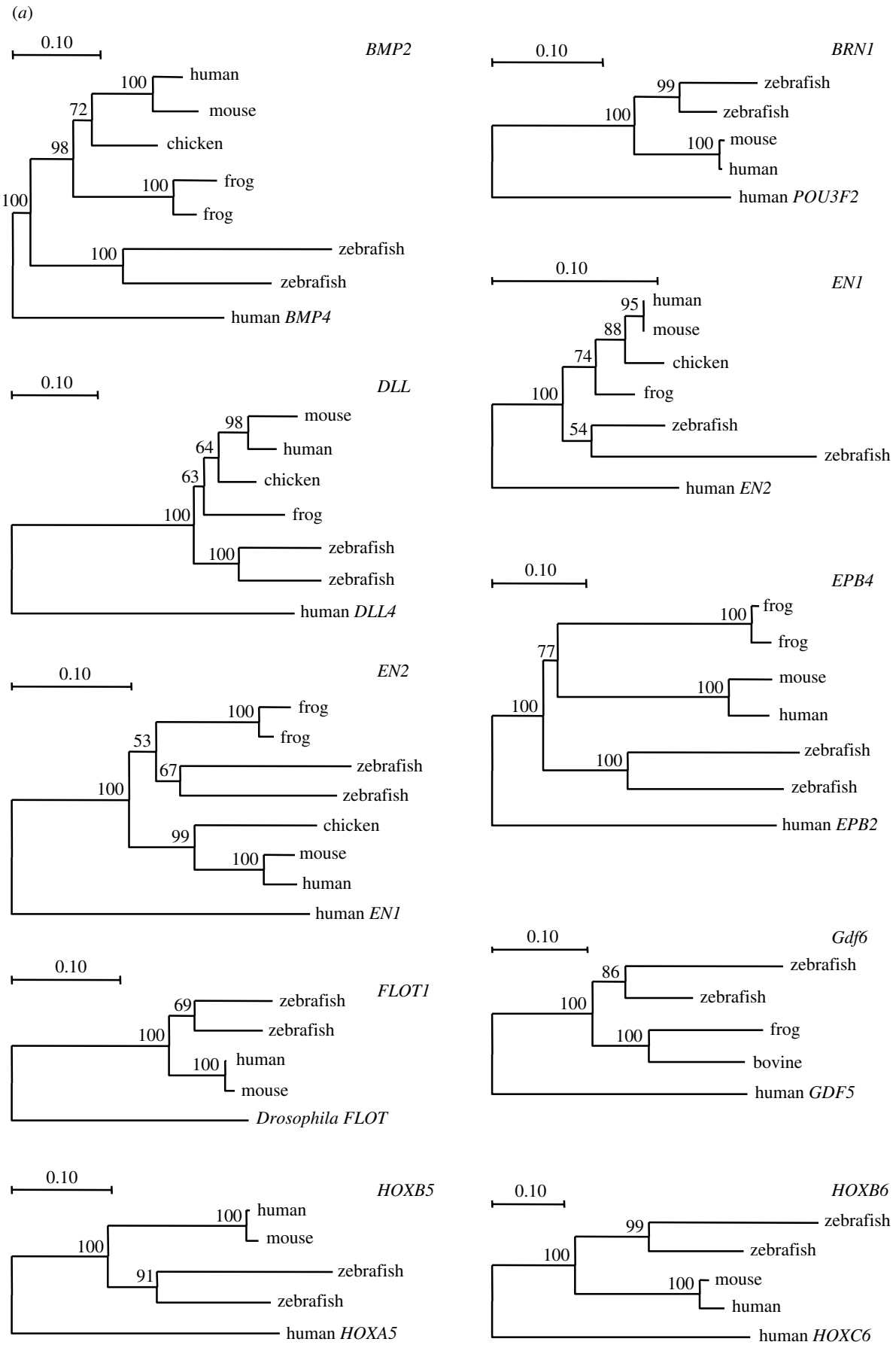

Figure 3. (See caption opposite.)

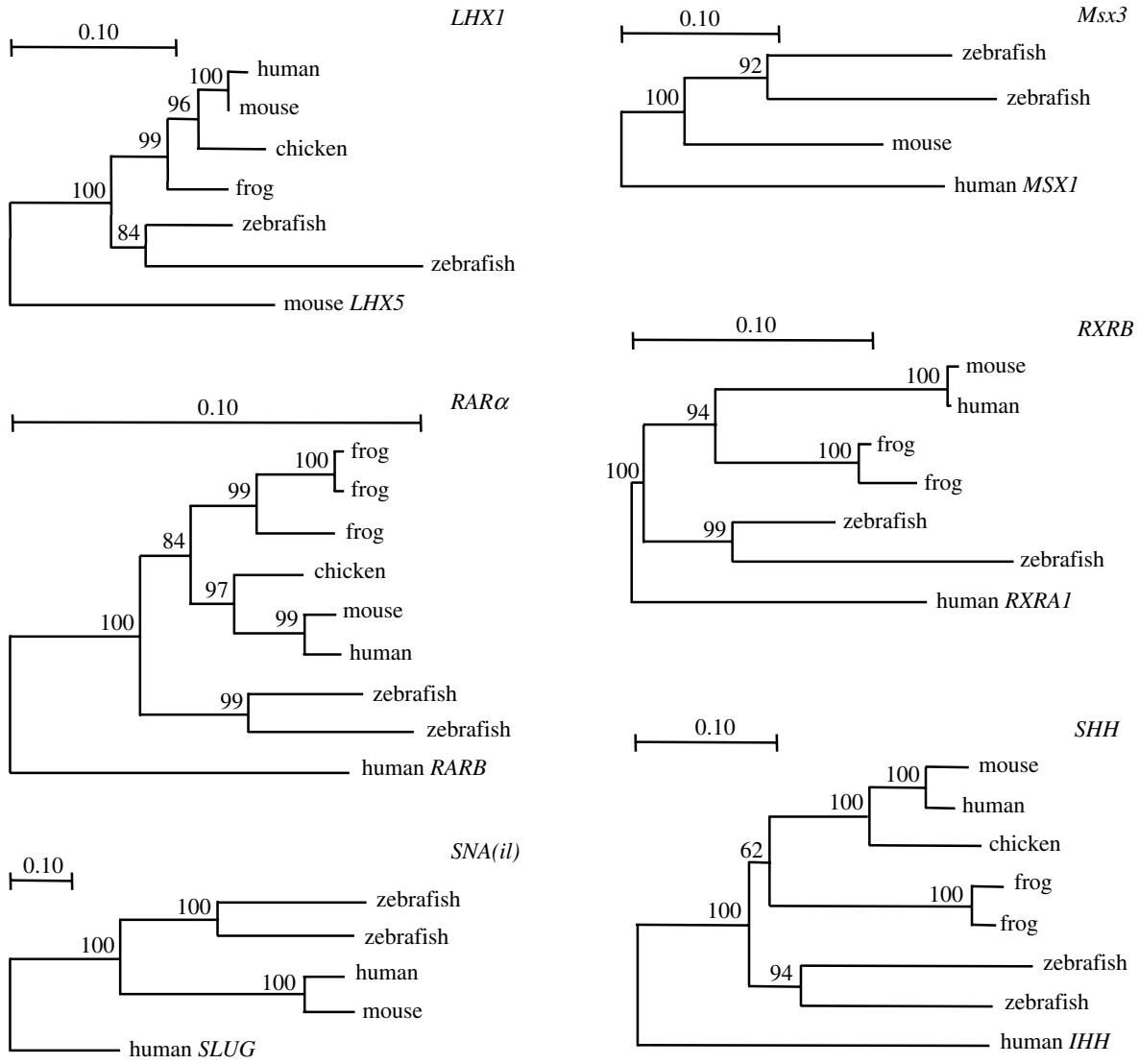

Figure 3. (*Continued*.) Phylogenies of duplicated fish genes. Trees were reconstructed using Poisson-corrected genetic distances and the neighbour-joining algorithm of Saitou & Nei (1987) as implemented in TREECON (Van de Peer & De Wachter 1994). Bootstrap values shown for nodes supported by more than 50% of 500 bootstrap reiterations (Felsenstein 1985). In all cases monophyly of the ingroup is well supported in an analysis that included other paralogues (see figure 2). The most closely related human paralogue was used to root the tree. (*a*) Phylogenies showing a sister sequence relationship for the zebrafish paralogues. Phylogenies of *ISL2*, *L1*(*CAM*) and *PAX2* genes had the same topologies as the genes shown here but the map positions of the zebrafish co-orthologues (table 2) suggest that they were not produced during genome duplication. (*b*) Phylogenies that include two zebrafish co-orthologues but not the expected sister sequence relationships. Maximum likelihood analyses (not shown) produce the duplication topology for *FKD5*, *HOXC6* and *SOX11*.

reconstruction artefacts, perhaps caused by unequal rates of evolution in one of the zebrafish co-orthologues.

Synteny data indicate that zebrafish have two co-orthologues for 10 human *Hox* genes: *B1*, *B5*, *B6*, *C6*, *B8*, *A9*, *A11*, *C11*, *A13*, *C13* (Amores *et al*. 1998). If these additional *Hox* genes in zebrafish were produced by genome duplication, then we should have been able to reconstruct the 'duplication topology' for each of them. Instead, we found the topology predicted by the genome duplication hypothesis for only *HoxB5* and *HoxB6* genes (and for *HoxC6* genes when ML methods were used). For *HoxB1*, *HoxA11*, *HoxC11*, *HoxA13* and *HoxC13*, one or both of the zebrafish sequences in the database was 73 amino acids long or less and was comprised almost entirely of the highly conserved homeodomain, which is 60–63 amino acids long (Bürglin 1994). The lack of variation in these short sequences precluded reliable tree reconstruction. For *HoxB8*, only one zebrafish sequence (*hoxB8b*) occurred in the database. For *HoxA9* the two zebrafish genes, *hoxA9a* and *hoxA9b*, occurred within a well-supported *Hox9* clade and were sister sequences, but were not assigned to any of the four *Hox9* clades.

Gates *et al*. (1999) and Barbazuk *et al*. (2000) included *Hes5* among their list of genes with two zebrafish co-orthologues. Both studies report that zebrafish genes *her2* and *her4* are orthologous to mouse *Hes5*. However, our BLASTp searches turned up three additional zebrafish genes (*her1*, *her3* and *her7*) that cluster with mouse *Hes5* and the topology of the expanded tree (whether based upon NJ or ML methods) does not support the hypothesis that any pair of zebrafish genes are co-orthologues of mouse *Hes5*.

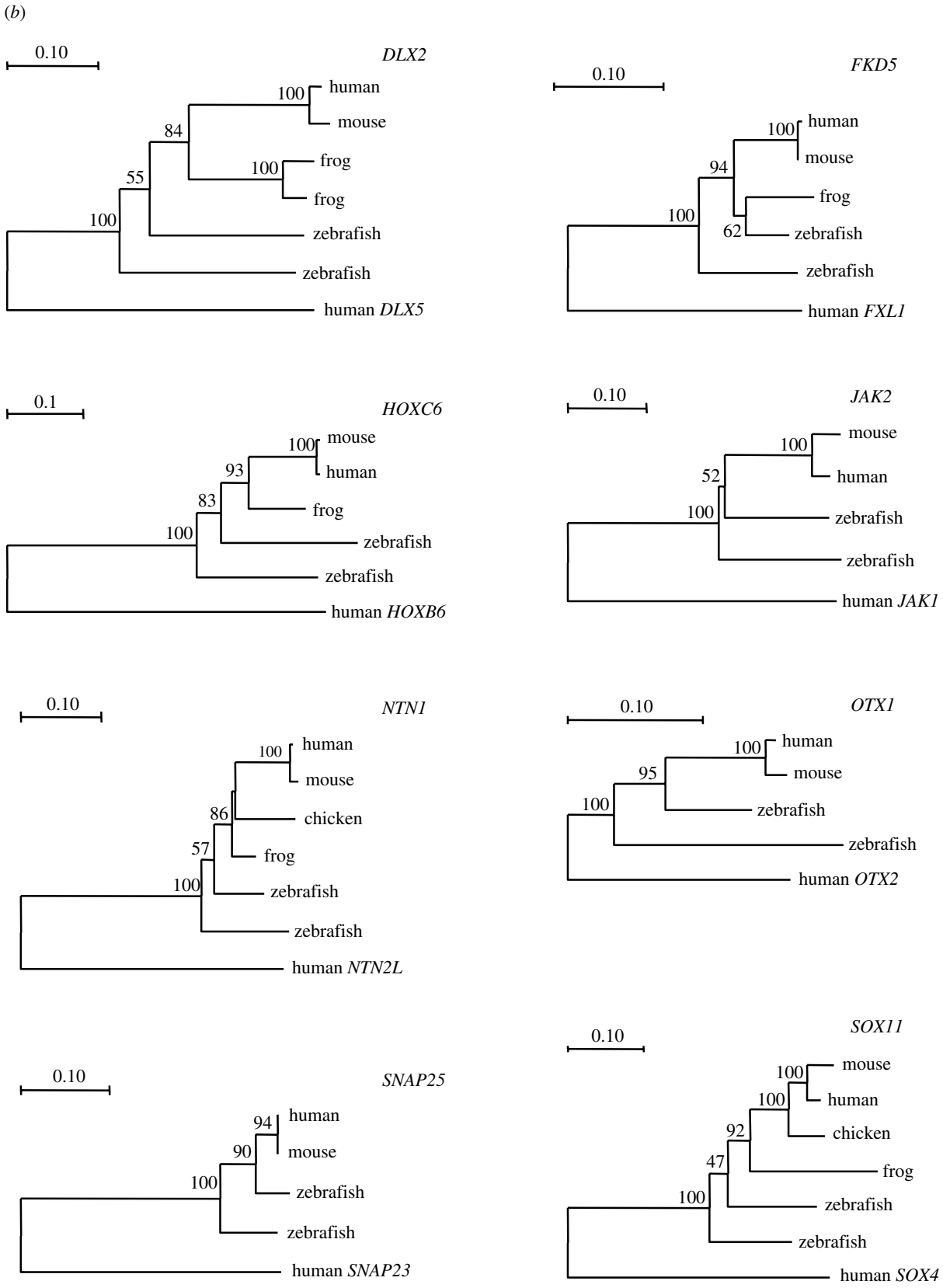

Figure 3. (*Continued*.)

(b) *Age of co-orthologues*

Since additional *Hox* clusters are present in both zebrafish and *Takifugu* (see § 1), the fish-specific genome duplication is believed to have happened before the divergence of Cypriniformes (zebrafish) and Tetraodontiformes (*Takifugu*), at least 150 Myr ago (Nelson 1994; Cantatore *et al*. 1994). On the other hand, the duplication most probably took place after the divergence of ray-finned and

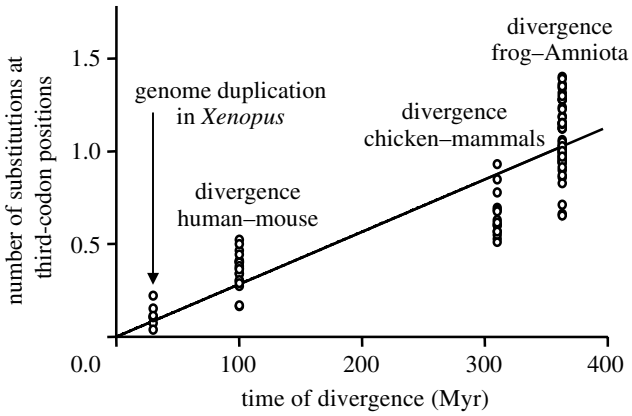

Figure 4. Substitutions at third-codon positions plotted against divergence dates (see § 2) for taxa included in this study. The divergence of Actinopterygii and Sarcopterygii (*ca*. 450 Myr ago) was excluded because third positions are saturated and the inclusion of these data would erroneously influence the regression. The average number of third-codon position substitutions between pairs of zebrafish co-orthologues is 1.02 (s.d. = 0.24).

lobe-finned fishes, *ca*. 450 Myr ago (Kumar & Hedges 1998; Lee 1999), since all sarcopterygian species studied so far have four or fewer *Hox* gene clusters. This is consistent with our observations that for many phylogenetic trees, zebrafish paralogues appear to have been formed during the time interval between the divergence of amphibians and amniotes, and the divergence between reptiles (i.e. birds) and mammals (figure 3*a*).

A comparison of synonymous and non-synonymous substitutions in duplicated genes of varying ages and from a diversity of species suggests that genes experience a period of accelerated evolution shortly after gene duplication (Lynch & Conery 2000). Acceleration in the rate of evolution of both zebrafish genes compared with frog, chicken, mouse and human genes might mean that the genome duplication is younger than it appears to be on our phylogenies (though an increase in non-synonymous mutations following a duplication event should not affect our genetic distance estimates based upon third-codon positions). Allotetraploidy might have also confounded our ability to date the fish genome duplication. Gene duplication (i.e. tetraploidy) occurs when cytokinesis fails during the first mitotic division of a fertilized egg (Sheppard *et al*. 1982). In autotetraploidy, 'duplicate' genes come from two individuals of the same species and are identical or are alleles at a given locus. With allotetraploidy the two genomes involved come from different species and may have diverged extensively at the faster-evolving loci before the tetraploidy, i.e. duplication event (Spring 1997). Thus, for genome duplication via allotetraploidy, divergence between co-orthologues begins before the tetraploidy event (i.e. genome duplication).

Despite these possible sources of error in the estimation of the fish genome duplication, our estimate that the duplicated zebrafish genes are between 300 and 450 million years old indicates that genome duplication preceded the teleost radiation. Study of 'basal' actinopterygians (e.g. bichir, sturgeon, bowfin, gar) will help to determine more accurately the date of the fish genome duplication.

(c) *Gene location*

Comparative genomics has provided many new insights into the evolution of chromosomes. Radiation hybrid maps have shown that there are orthologous chromosome regions in human and mouse (Nadeau & Sankoff 1998), in human and cat (Murphy *et al*. 2000), human and cattle (Band *et al*. 2000), and in human and zebrafish (Barbazuk *et al*. 2000). Genome duplication means that many species also possess paralogous chromosome regions (e.g. Morizot *et al*. 1991; Lundin 1993; Amores *et al*. 1998; Pébusque *et al*. 1998). Indeed, the term 'co-orthology' can be applied to regions of chromosomes as well as genes.

The duplicated zebrafish genes uncovered in this study occur on a large proportion of the 25 zebrafish linkage groups, but they do not appear to be randomly distributed in the zebrafish genome. Our phylogenetic data indicate that regions of zebrafish LG1 and LG9, LG2 and LG7, LG3 and LG12, LG11 and LG23, LG17 and LG20 are paralogous (table 2).

(d) *The retention and loss of duplicated genes*

Several models have been proposed to explain the evolutionary persistence of duplicated genes in zebrafish. Gibson & Spring (1998) argue that selection can prevent the loss of redundant genes (i.e. duplicates) if those genes code for components of multidomain proteins because mutant alleles disrupt multidomain proteins (i.e. are dominant negative mutations). Force *et al*. (1999) argue that when a gene with multiple functions is duplicated, the duplicates are redundant only for as long as each retains the ability to perform all ancestral roles. When one duplicate experiences a mutation that prevents it from carrying out one of its ancestral roles, the other duplicate is no-longer redundant. This is consistent with Sidow's (1996) proposition that a single unique function in an ocean of redundancy is enough to keep the gene afloat and prevent degenerative substitutions. According to Force *et al*.'s (1999) 'duplication degeneration–complementation' model, degenerative mutations preserve rather than destroy duplicated genes. Force *et al*. (1999) present *EN1* as an example of their model. Zebrafish *eng1a* and *eng1b* appear to have divided the roles of their orthologues (e.g. human *EN1*). It will be interesting to find out if the other co-orthologues reported here have divided the roles of their sarcopterygian orthologues or are components of multidomain proteins. De Pinna (1996) provided a list of teleost synapomorphies. One convincing way to show that extra genes originating from genome duplication were responsible for the radiation of Teleostei would be to demonstrate that duplicated genes code for teleost-specific traits.

An alternative evolutionary link between the teleost radiation and genome duplication involves 'divergent resolution' (Lynch & Conery 2000; Taylor *et al*. 2001). Lynch and Conery proposed that the loss of different duplicates in geographically isolated populations could reduce the fecundity of hybrids. They considered a young pair of functionally redundant, unlinked, duplicate genes in an ancestral species. One of the two duplicates is likely to be silenced (i.e. become a pseudogene) within the next one

Table 2. Genome location and genetic distance between pairs of co-orthologous genes.

(Map data were obtained from the Zebrafish Information Network: http://zfish.uoregon.edu/ZFIN/, Gates *et al*. (1999) and Barbazuk *et al*. (2000). Symbols denote possible paralogous chromosomes. 'Confidential' means that the gene has been mapped but data are not available. Genetic distances were computed using only third codon positions and corrected for multiple events per site according to Tajima & Nei (1984). Estimated number of mutations per site are shown for *ISL2*, *L1(CAM)* and *PAX2* but these data are not included in the calculation of the mean because these zebrafish co-orthologues were probably produced by independent tandem duplications. Woods *et al*. (2000) recently reported that the two zebrafish *ISL2* genes and the two zebrafish *Pax2* genes do not occur on the same linkage groups (contrary to Barbazuk *et al*. 2000). Our phylogenies of *ISL2* and *Pax2* genes were consistent with the fish-specific genome duplication hypothesis (i.e. 'duplication topology' with high bootstrap support for all nodes) and the Tajima–Nei distance estimates for the *ISL2* and *Pax2* duplicates (table 2) are approximately the same as those for the other unlinked duplicates.)

	symbol	symbol (zebrafish)	location (zebrafish)	Tajima–Nei distance
1	*BMP2*	*bmp2a*	LG 17●	1.207
		bmp2b	LG 20●	
2	*BRN1*	*brn1.1*	LG 9	1.119
		brn1.2	LG 6	
3	*DLL1*	*dla*	LG 1	1.233
		dld	LG 13*	
4	*DLX2*	*dlx2*	LG 9	1.364
		dlx5	LG 1†	
5	*EN1*	*eng1a*	LG 9	0.931
		eng1b	LG 1†	
6	*EN2*	*eng2*	LG 7	1.199
		eng3	LG 2Ψ	
7	*EPB4*	*rtk4*	unmapped	0.975
		epa4	unmapped	
8	*FKD5*	*fkd3*	LG 25	1.027
		fkd5	unmapped	
9	*FLOT1*	*re2a*	unmapped	0.720
		re2b	unmapped	
10	*Hedgehog*	*shh*	LG 7	1.389
		twhh	LG 2Ψ	
11	*HOXB5*	*hoxb5a*	LG 3	0.749
		hoxb5b	LG 12Φ	
12	*HOXB6*	*hoxb6a*	LG 3	0.876
		hoxb6b	LG 12Φ	
13	*HOXC6*	*hoxC6a*	LG 23	1.009
		hoxC6b	LG 11Θ	
14	*JAK2*	*jak2a*	confidential	1.054
		jak2b	confidential	
15	*LHX1*	*lhx1*	LG 15	1.089
		lim6		
16	*msx3* (mouse)	*msxb*	LG 1	1.590
		msxc	LG 13*	
17	*NTN1*	*ntn1*	LG 3	0.863
		ntn1a	LG 6	
18	*OTX1*	*otx1*	LG 17	1.047
		otx3	LG 1	
19	*RARA*	*rara2a*	LG 12	0.964
		rara2b	LG 3Φ	
20	RXRB	*rxre*	LG 19	0.931
		rxrd	unmapped	
21	*SNA(il)*	*snail1*	LG 11	0.809
		snail2	LG 23Θ	
22	*SNAP25*	*snap25,1*	LG 20●	0.594
		snap25,2	LG 17●	
23	*SOX11*	*sox11a*	LG 17●	0.749
		sox11b	LG 20●	
Mean (s.d.)				1.02 (0.23)
	gdf6	*dynamo*	LG 19	NA
	(*bovine*)	*radar*	confidential	
	ISL2	*isl2*	LG 25	1.128
		isl3	LG 25	
	L1(CAM)	*l1.1*	LG 23	1.187
		l1.2	LG 23	
	PAX2	*pax2*	LG 13	0.873

to two million years. If the ancestral species is divided into geographically isolated populations, then a different copy of the duplicated gene could become fixed in the two populations. If the two populations hybridize, the F_1 progeny would be heterozygous in two respects. With respect to homologous chromosomes, one homologue would have a functional allele and the other a pseudogene. With respect to the entire genome, an F_1 individual would have two functional alleles of the locus but those alleles would occur on different chromosomes. In the F_2 generation, there is a 6.25% chance that an individual will receive only pseudogenes of a given duplicated and differentially resolved gene. If the gene in question is an essential gene, then 6.25% of the F_2 generation would not survive. Furthermore, 25% of F_2 individuals may also suffer reduced fitness because they would be haploid at this locus. Lynch & Conery (2000) stated that with tens to hundreds of young unresolved gene duplicates present in most eukaryotic genomes, such genes could provide a common substrate for the passive origin of isolating barriers. However, genome duplication (e.g. in the ancestor of teleost fishes) provides many more than tens to hundreds of unlinked, duplicated genes. Divergent resolution of thousands of genes might be a very powerful isolating mechanism. One prediction of this model in which genome duplication leads to speciation is that tetraploid taxa should have more species than their diploid sister groups.

(e) *Terminology*

In this paper we have adopted the term 'co-orthologue' (Gates *et al*. 1999). In our opinion, this term is useful because it conveys information about genome duplications that is not obvious from the term 'orthologue'. Supplements to orthology and paralogy have also been introduced by Holland (1999) and Sharman (1999): 'pro-orthologue' describes the relationship of a gene to one of the post-duplication descendants of its orthologue. Human *RARA* is, for example, a pro-orthologue of the zebrafish genes *rara2a* and the *rara2b* (figure 2).'Semi-orthologue' describes the relationship of one of a set of duplicated genes to a gene directly descended from the ancestor of the whole set (e.g. *rara2a* is semi-orthologous to *RARA*). Because semi-orthologue implies 'half orthologue' it might be a more appropriate term than co-orthologue for comparisons between diploid fish genes and their human pro-orthologues. Such a naming approach could be extended to include other genic relationships. For example, genes in most actinopterygians might be considered 'octalogues' of their respective orthologous genes in invertebrates. However, attempts to describe such gene relationships numerically can become awkward. For example, how would the relationship between genes in tetraploid fish such as the goldfish (*Carassius auratus*) and genes in *Drosophila* be described? In this case a 1:16 gene ratio is expected, based upon the four genome duplications that probably separate these species. Even for a species between which a 1:2 or a 1:4 gene ratio is expected based upon genome duplication data, tandem duplications can disrupt the actual orthologue ratio. Therefore, we prefer the terms pro-orthologue and co-orthologue to describe relationships between genes in taxa separated by any number of tandem or genome duplications.

(f) *Problems with gene nomenclature*

Our conclusion that there was a genome duplication event in fish means that all genes in actinopterygian fish have co-orthologous relationships with their sarcoterygian (e.g. human) orthologues. Currently the names of many zebrafish genes reflect their co-orthologous relationship to orthologues or 'pro-orthologues' in sarcopterygians (e.g. *bmp2a* and *bmp2b*; *eng1a* and *eng1b*). However, in many cases the fact that a given zebrafish gene is one of two orthologues is not clear from its name. For example, the following pairs of genes were shown to be co-orthologues in our study: *dla* and *dld*, *dlx2* and *dlx5*, *eng2* and *eng3*, *isl2* and *isl3*, *rxrE* and *rxrD*, *shh* and *twhh*, *otx1* and *otx3*, *fkd3* and *fkd5*, and *dynamo* and *radar*.

We propose all genes in diploid fish be given the same name as pro-orthologues in humans but that these names be appended with an '*a*' or '*b*' designation to reflect their co-orthologous relationships with human (and other sarcopterygian) genes. In cases where only one co-orthologue appears to have been retained, the '*a*' designation serves as a reminder of the genes' duplication history.

Tiggy-winkle hedgehog (Ekker *et al*. 1995) highlights the potential confusion generated when the name of a gene lacks phylogenetic information. *Tiggy-winkle hedgehog* (*twhh*) and *sonic hedgehog* (*shh*) in zebrafish are equally orthologous (i.e. co-orthologous) to *sonic hedgehog* (*SHH*) in humans (present study; Zardoya *et al*. 1996). A PubMed search suggests that this fact is not widely appreciated: 29 references include the terms; *shh* + zebrafish and only five include *twhh* + zebrafish. Furthermore, a gene named '*twhh*' has been sequenced in goldfish. However, goldfish *twhh* cannot be orthologous to zebrafish *twhh*, as might be expected from its name, because goldfish are tetraploid (Zhang *et al*. 1999). That is, the goldfish *twhh* that has been sequenced can only be co-orthologous to zebrafish *twhh* (i.e. one of two *twhh* co-orthologues).

Our phylogenetic study also turned up naming 'errors' in genes for which only one co-orthologue is currently known. Zebrafish *rxra* clusters with strong bootstrap support within the *RXRc* clade. Conversely, zebrafish *rxrc* clusters with strong support within the *RXRa* clade. As this list of confusing and erroneous names grows a complete review of fish gene nomenclature will become increasingly important just as it was for *Hox* genes in 1992 (De Robertis 1994).

Woods *et al*. (2000) recently reported that the two zebrafish *Isl2* genes and the two zebrafish *Pax2* genes do not occur on the same linkage groups (contrary to Barbazuk *et al*. 2000). Our phylogenies of *Isl2* and *Pax2* genes were consistent with the fish-specific genome duplication hypothesis (i.e.,'duplication topology' with high bootstrap support for all nodes), and the Tajima–Nei distance estimates for the *Isl2* and *Pax2* duplicates (table 2) are approximately the same as those for the other unlinked duplicates.

We thank Jürg Spring, Angel Amores, Tomaso Patarnello and Henner Brinkmann for helpful discussions. Alexander Schmid and Tancred Frickey provided laboratory assistance. J.S.T. is supported by a postdoctoral fellowship from the Natural Sciences and Engineering Research Council of Canada. Y.V.dP. is a Research Fellow of the Fund for Scientific Research, Flanders (Belgium). We thank the Deutsche Forschungsgemeinschaft for grants to Y.V.dP. (842/2-1) and to A.M. (1725/2-1, 1725/3-1, 1725/4-1, 1725/5-1) and the Verband der Chemischen Industrie.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990 Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Amores, A. (and 12 others) 1998 Zebrafish *hox* clusters and vertebrate genome evolution. *Science* **282**, 1711–1714.

Aparicio, S. 2000 Vertebrate evolution: recent perspectives from fish. *Trends Genet.* **16**, 54–56.

Band, M. R. (and 10 others) 2000 An ordered comparative map of the cattle and human genomes. *Genome Res.* **10**, 1359–1368.

Barbazuk, W. B., Korf, I., Kadavi, C., Heyen, J., Tate, S., Wun, E., Bedell, J. A., McPherson, J. D. & Johnson, S. L. 2000 The syntenic relationship of the zebrafish and human genomes. *Genome Res.* **10**, 1351–1358.

Burglin, T. R. 1994 A comprehensive classification of homeobox genes. In *Guidebook to the homeobox genes* (ed. D. Duboule), pp. 27–71. Oxford University Press.

Cantatore, P., Roberti, M., Pesole, G., Ludovico, A., Milella, F., Gadaleta, M. N. & Saccone, C. 1994 Evolutionary analysis of cytochrome *b* sequences in some Perciformes: evidence for a slower rate of evolution than in mammals. *J. Mol. Evol.* **39**, 589–597.

Carroll, R. L. 1997 *Patterns and processes of vertebrate evolution*. Cambridge University Press.

Cheng, C.-H. C. & Chen, L. 1999 Evolution of an antifreeze glycoprotein. *Nature* **401**, 443–444.

de Pinna, C. C. M. 1996 Teleostean monophyly. In *Interrelationships of fishes* (ed. M. L. J. Stiassny, L. R. Parenti & G. D. Johnson), pp. 147–162. Academic Press.

De Robertis, E. M. 1994 The homeobox in cell differentiation and evolution. In *Guidebook to the homeobox genes* (ed. D. Duboule), pp. 13–23. Oxford University Press.

Dulai, K. S., von Dornum, M., Mollon, J. D. & Hunt, D. M. 1999 The evolution of trichromatic colour vision by opsin gene duplication in New World and Old World primates. *Genome Res.* **9**, 629–638.

Ekker, S. C., Ungar, A. R., Greenstein, P., von Kessler, D. P., Porter, J. A., Moon, R. T. & Beachy, P. A. 1995 Patterning activities of vertebrate hedgehog proteins in the developing eye and brain. *Curr. Biol.* **5**, 944–955.

Elgar, G. (and 11 others) 1999 Generation and analysis of 25 Mb of genomic DNA from the pufferfish *Fugu rubripes* by sequence scanning. *Genome Res.* **9**, 960–971.

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

Fitch, W. 1970 Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**, 99–113.

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

Gates, M. A., Kim, L., Egan, E. S., Cardozo, T., Sirotkin, H. I., Dougan, S. T., Lashkari, D., Abagyan, R., Schier, A. F. & Talbot, W. S. 1999 A genetic linkage map for zebrafish: comparative analysis and localization of genes and expressed sequences. *Genome Res.* **9**, 334–347.

Gibson, T. J. & Spring, J. 1998 Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. *Trends Genet.* **14**, 46–49.

Gu, X. 1998 Early metazoan divergence was about 830 million years ago. *J. Mol. Evol.* **47**, 369–371.

Hinegardner, R. & Rosen, D. E. 1972 Cellular DNA content and the evolution of teleostean fishes. *Am. Nat.* **106**, 621–644.

Holland, P. W. H. 1992 Homeobox genes in vertebrate evolution. *BioEssays* **14**, 267–273.

Holland, P. W. H. 1999 The effect of gene duplication on homology. In *Homology* (ed. G. R. Bock & G. Cardew), pp. 226–242. Wiley: Chichester.

Hughes, A. L. 1999*a* Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *J. Mol. Evol.* **48**, 565–576.

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

Hughes, M. K. & Hughes, A. L. 1993 Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. *Mol. Biol. Evol.* **10**, 1360–1369.

Kishino, H. & Hasegawa, M. 1989 Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* **29**, 170–179.

Kumar, S. & Hedges, S. B. 1998 A molecular timescale for vertebrate evolution. *Nature* **392**, 917–920.

Lee, M. S. 1999 Molecular clock calibrations and metazoan divergence dates. *J. Mol. Evol.* **49**, 385–391.

Li, W. H., Gouy, M., Sharp, P. M., O'hUigin, C. & Yang, Y. W. 1990 Molecular phylogeny of Rodentia, Lagomorpha, Primates, Artiodactyla, and Carnivora and molecular clocks. *Proc. Natl Acad. Sci. USA* **87**, 6703–6707.

Lundin, L.-G. 1993 Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics* **16**, 1–19.

Lundin, L.-G. 1999 Gene duplications in early metazoan evolution. *Cell Dev. Biol.* **10**, 523–530.

Lydeard, C. & Roe, K. J. 1997 The phylogenetic utility of the mitochondrial cytochrome *b* gene for inferring relationships among actinopterygian fishes. In *Molecular systematics of fishes* (ed. T. C. Kocher & C. A. Stepien), pp. 285–303. San Diego, CA: Academic Press.

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

Maglott, D. R., Katz, K. S., Sicotte, H. & Pruitt, K. D. 2000 NCBI's LocusLink and RefSeq. *Nucleic Acids Res.* **28**, 126–128.

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.

Morizot, D. C., Slaugenhaupt, S. A., Kallman, K. D. & Chakravarti, A. 1991 Genetic linkage map of fishes of the genus *Xiphophorous* (Teleostei: Poeciliidae). *Genetics* **127**, 399–410.

Murphy, W. J., Sun, S., Chen, Z.-Q., Yuhki, N., Hirschmann, D., Menotti-Raymon, M. & O'Brien, S. J. 2000 A radiation hybrid map of the cat genome: implications for comparative mapping. *Genome Res.* **10**, 691–702.

Nadeau, J. H. & Sankoff, D. 1998 The lengths of undiscovered conserved segments in comparative maps. *Mamm. Genome* **9**, 491–495.

Naruse, K. (and 19 others) 2000 A detailed linkage map of Medaka, *Oryzias latipes*: Comparative genomics and genome evolution. *Genetics* **154**, 1773–1784.

Nei, M. & Kumar, S. 2000 *Molecular evolution and phylogenetics*. New York: Oxford University Press.

Nei, M., Rogozin, I. B. & Piontkivska, H. 2000 Purifying selection and birth-and-death evolution in the ubiquitin gene family. *Proc. Natl Acad. Sci. USA* **97**, 10 866–10 871.

Nelson, J. S. 1994 *Fishes of the world*, 3rd edn. New York: Wiley.

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

Ohno, S. 1999 The one-to-four rule and paralogues of sex-determining genes. *Cell. Mol. Life Sci.* **55**, 824–830.

Page, R. D. M. 1996 TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* **12**, 357–358.

Patel, N. H. & Prince, V. E. 2000 Beyond the *Hox* complex. *Genome Biol.* **1**, 1027.1–1027.4.

Pébusque, M.-J., Coulier, F., Birnbaum, D. & Pontarotti, P. 1998 Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *Mol. Biol. Evol.* **15**, 1145–1159.

Saitou, N. & Nei, M. 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.

Sharman, A. C. 1999 Some new terms for duplicated genes. *Cell Dev. Biol.* **10**, 561–563.

Sheppard, D. M., Fisher, A., Lawler, S. D. & Povey, S. 1982 Tetraploid conceptus with three paternal contributions. *Hum. Genet.* **62**, 371–374.

Sidow, A. 1996 Gen(om)e duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev.* **6**, 715–722.

Spring, J. 1997 Vertebrate evolution by interspecific hybridization—are we polyploid? *FEBS Lett.* **400**, 2–8.

Stellwag, E. J. 1999 *Hox* gene duplication in fish. *Cell Dev. Biol.* **10**, 531–540.

Strimmer, K. & Von Haeseler, A. 1996 Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13**, 964–969.

Tajima, F. & Nei, M. 1984 Estimation of evolutionary distance between nucleotide sequences. *Mol. Biol. Evol.* **1**, 269–285.

Taylor, J. S., Van de Peer, Y. & Meyer, A. 2001 Genome duplication, divergent resolution and speciation. *Trends Genet.* **17**, 299–301.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. 1997 The CLUSTAL X Windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**, 4876–4882

Van de Peer, Y. & De Wachter, Y. 1994 TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. *Comput. Appl. Biosci.* **10**, 569–570.

Wang, Y. & Gu, X. 2000 Evolution patterns of gene families generated in the early stage of vertebrates. *J. Mol. Evol.* **51**, 88–96.

Woods, I. G., Kelly, P. D., Chu, F., Ngo-Hazelett, P., Yan, Y.-L., Huang, H., Postlethwait, J. H. & Talbot, W. S. 2000 A comparative map of the zebrafish genome. *Genome Res.* **10**, 1903–1914.

Zardoya, R., Abeiheif, E. & Meyer, A. 1996. Evolution and orthology of *hedgehog* genes. *Trends Genet.* **12**, 496–497.

Zhang, L. J., Zhu, Y. P., Xiao, W. H. & Huang, S. Y. 1999 Genetic diversity in crucian carp (*Carassius auratus*). *Biochem. Genet.* **10**, 267–279.