J Mol Evol (1999) 48:565–576

JOURNAL OF MOLECULAR EVOLUTION

Phylogenies of Developmentally Important Proteins Do Not Support the Hypothesis of Two Rounds of Genome Duplication Early in Vertebrate History

Austin L. Hughes

Department of Biology and Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, University Park, PA 16802, USA

Received: 22 December 1997 / Accepted: 5 October 1998

Abstract. It has been proposed that two rounds of duplication of the entire genome (polyploidization) occurred early in vertebrate history (the 2R hypothesis); and the observation that certain gene families important in regulating development have four members in vertebrates, as opposed to one in *Drosophila*, has been adduced as evidence in support of this hypothesis. However, such a pattern of relationship can be taken as support of the 2R hypothesis only if (1) the four vertebrate genes can be shown to have diverged after the origin of vertebrates, and (2) the phylogeny of the four vertebrate genes (A–D) exhibits a topology of the form (AB) (CD), rather than (A) (BCD). In order to test the 2R hypothesis, I constructed phylogenies for nine protein families important in development. Only one showed a topology of the form (AB) (CD), and that received weak statistical support. In contrast, four phylogenies showed topologies of the form (A) (BCD) with statistically significant support. Furthermore, in two cases there was significant support for duplication of the vertebrate genes prior to the divergence of deuterostomes and protostomes: in one case there was significant support for duplication of the vertebrate genes at least prior to the divergence of vertebrates and urochordates, and in one case there was weak support for duplication of the vertebrate genes prior to the divergence of deuterostomes and protostomes. Taken together with other recently published phylogenies of developmentally important genes, these results provide strong evidence against the 2R hypothesis.

Key words: Gene duplication — Genome duplication — Vertebrates — Evolution of development — Protein phylogeny

Introduction

Several authors have proposed that there were two rounds of duplication of the entire genome, presumably resulting from polyploidization, early in the history of the vertebrates (Ohno 1970; Holland et al. 1994; Sidow 1996; Kasahara et al. 1996). As evidence in favor of this hypothesis, Sidow (1996, p 715) mentions the following: "When comparing *Drosophila* with vertebrates, one finds an uncanny consistency in the multiple by which vertebrate developmental regulator genes outnumber their *Drosophila* homologues: it is often the number four (e.g. *Hox* clusters, *Cdx, MyoD, 60A, Notch, elav, btd/SP* . . .) and sometimes two (e.g. *Wnt-5, decapentaplegic, Eve* . . .) or three (e.g. *Msx, Hedgehog* . . .)." As further evidence for this hypothesis, Sidow (1996) states that vertebrates are estimated to have approximately four times as many genes as does *Drosophila*, an estimate which he attributes to Miklos and Rubin (1996). In fact, the estimates presented by Miklos and Rubin (1996) place the number of genes in the bony fish *Fugu rubripes*, in the mouse, and in the human at about 5.8 times the number in *Drosophila*. Sidow (1996, p. 715) further

Correspondence to: A.L. Hughes; *e-mail:* austin@hugaus3.bio.psu.edu

hypothesizes that gene families "with only two vertebrate paralogs which lost one copy after the first genome duplication; those vertebrate gene families with three lost one paralog after the second genome duplication."

In spite of widespread citation of the hypothesis of two rounds of genome duplication early in vertebrate history (the 2R hypothesis), no study has attempted to subject it to rigorous testing by phylogenetic analysis of gene families. The purpose of the present paper is to conduct such tests. Although, as pointed out by Sidow (1996) in the passage cited above, the occurrence of families having two or three paralogues in vertebrates can be reconciled with the 2R hypothesis if we assume that deletions of duplicate genes have occurred, these families cannot really be used to test the 2R hypothesis because their occurrence is also consistent with several alternative explanations. Even the occurrence of four paralogues in vertebrates cannot in itself be taken as supporting the 2R hypothesis. For example, if the four vertebrate paralogues are shown by phylogenetic analysis to have duplicated prior to the origin of vertebrates, then clearly their duplication could not have occurred as part of the hypothetical genome duplications early in vertebrate history. An example of a phylogeny of this sort is shown in Fig. 1C. In this example, the duplication occurred prior to the divergence of protostomes (including insects) from deuterostomes (including vertebrates).

Furthermore, even when four vertebrate paralogues can be shown to have diverged after the origin of vertebrates, their phylogenetic relationship must exhibit a specific topology in order to be counted as supporting the 2R hypothesis. This topology is illustrated in Fig. 1A. It can be referred to as a topology of the form (AB) (CD), because in it the four genes (A–D) form two clusters, with A being a sister group to B and C a sister group to D. An alternative topology is one in which one of the four paralogues diverged prior to the others (Fig. 1B). This topology can be symbolized as (A) (BCD). Note that the topology of the relationships among B, C, and D is not relevant to the question of support for the 2R hypothesis. Clearly a topology of the (A) (BCD) type does not support the 2R hypothesis. Of course, it is possible to invent ad hoc scenarios to reconcile such a topology with the 2R hypothesis; for example, one can hypothesize a series of events of deletion and of tandem gene duplication occurring independently of the hypothesized genome duplications. Nonetheless, the widespread occurrence of topologies of the (A) (BCD) type in gene families having four paralogues in vertebrates would be evidence against the 2R hypothesis.

In order to test the 2R hypothesis, I reconstructed nine phylogenies of proteins which play important roles in regulating development, which have at least one known homologue in *Drosophila*, and which have four paralogues in vertebrates. To provide additional tests of the hypothesis, the results were compared with those of recently published studies of additional gene families having similar properties.

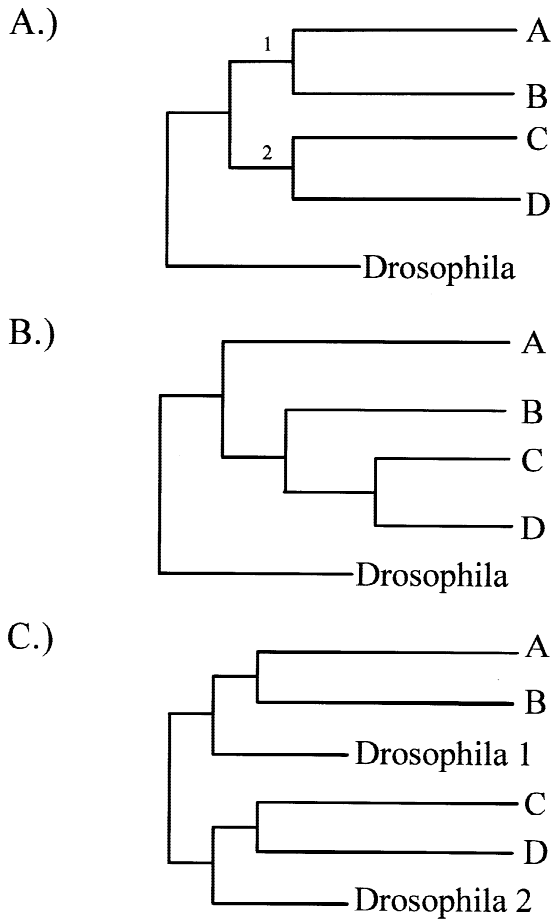

Fig. 1. Examples of possible phylogenies for a gene family having four members (A–D) in vertebrates: **A** a topology of the form (AB) (CD), which supports the hypothesis of two rounds of gene duplication early in vertebrate history; **B** an example of a phylogeny with topology of the form (A) (BCD); **C** a case in which the gene duplication separating the ancestor of A and B from that of C and D occurred prior to the divergence of deuterostomes and protostomes.

The expectation was that, if the 2R hypothesis is true, in a majority of the families there would be strong support for an (AB) (CD) topology. In contrast, a high proportion of gene families with paralogues that diverged prior to the origin of vertebrates (Fig. 1C) or with topologies of the (A) (BCD) type would argue against the 2R hypothesis. From a methodological point of view, it is important to realize that the null hypothesis in such an analysis must be the hypothesis of no effect, that is, in this case, the hypothesis that two rounds of genome duplication did not occur early in vertebrate history. Only if the data provide compelling reason to reject the null hypothesis—i.e., a large proportion of gene families showing the (AB) (CD) topology—can we reject the null hypothesis and accept the 2R hypothesis.

Table 1. Sequences used in analyses

CDX
- Nematoda: *Caenorhabditis elegans* C38D4.8 (Z46241)
- Arthropoda: *Drosophila melanogaster* caudal (M21070); silkmoth (*Bombyx mori*) cdd (D16683)
- Chordata: Vertebrata: zebrafish (*Brachydanio rerio*) CDX4 (x66958); carp (*Cyprinus carpio*) CDX4 (X80668); clawed frog (*Xenopus laevis*) CAD2 (U04032), CAD3 (U02034); chicken (*Gallus gallus*) CDX-C (U080614), CAD (X57760); mouse (*mus musculus*) CDX1 (L08063), CDX2 (U00454), CDX4 (L08061); golden hamster (*Mesocricetus auratus*) CDX2 (X81404); human (*Homo sapiens*) CDX1 (U16360), CDX2 (Y13709)

BMP
- Arthropoda: *Drosophila melanogaster* 60A (M77017), dpp (U63857), screw (U17573); *Drosophila pseudoobscura* dpp (U63857); *Drosophila virilis* 60A (U48595), dpp (U63855); flour beetle (*Tribolium castaneum*) dpp (U63132)
- Echinodermata: purple sea urchin (*Strongylocentrotus purpuratus*) DRV1 (Z48313)
- Chordata: Urochordata: ascidian (*Holocynthia roretzi*) BMPa (D83183)
- Chordata: Vertebrata: zebrafish BMP2/4a (U82232), BMP2/4b (U82233), BMP4a (U82231), BMP4b (U90122); clawed frog BMP2A (X55031), BMP2B (X63425), BMP4 (X64583), BMP7 (X63427); chicken BMP5 (S83278); mouse BMP2 (L25602), BMP6 (X80992), BMP7 (X56906), BMP8A (M97017), BMP8B (U39545); human BMP2 (M22489), BMP4 (M22490), BMP5 (M60314), BMP6 (M60315), BMP7 (X51807), BMP8 (M97016)

Elav
- Nematoda: *C. elegans* F35H8.5 (Z36752)
- Arthropoda: *Drosophila melanogaster* elav (M21152), sex-lethal (M23636), RBP9-2 (L04930); phorid fly *Megaselia scalaris* sex-lethal (X98769)
- Chordata: Vertebrata: zebrafish HuC (U62018), HuD (U17602); clawed frog HuA (U17596), HuB (U17597), HuC (U17598), HuD (U17599); mouse HuA (U65735), HuB (U29088), HuC (U29148); rat HuD (S583320); human HuA (U38175); HuB (U12431); HuC (L26405); HuD (M62843)

Egr/SP
- Nematoda: *C. elegans* C27C12.2 (Z69883), T22C8.5 (Z49071)
- Arthropoda: *Drosophila melanogaster* stripe b (U42402)
- Chordata: Vertebrata: zebrafish EGR1 (U12895), EGR2 (X70322); clawed frog EGR2 (S56884); mouse EGR1 (M20157), EGR2 (M24377), SP4 (U62522); rat EGR3 (U12428), EGR4 (M65008), SP1 (D12768); human EGR1 (X52541), EGR2 (J04076), EGR3 (X63741), EGR4 (X69438), SP1 (J03133), SP2 (M97910), SP3 (X68560), SP4 (X68561)

Brachyury
- Nematoda: *C. elegans* ZK328.6 (U50193), F40H6.4, Tbx9 (Z29443), T07C4.2 (Z29443), F21H11.3 (U11279)
- Arthropoda: *Drosophila melanogaster* trg (S74163), omb (S61744)
- Echinodermata: sea urchin (*Hemicentrotus pulcherrimus*) TbxT (D56332)
- Chordata: Cephalocordata: lancelet (*Branchiostoma floridae*) T (X91903). Urochordata: ascidian (*Halocynthia roretzi*) T (D16441). Vertebrata: zebrafish T (S57147); clawed frog T (M77243), Tbx6 (S83518); chicken T (U25176), Tbx6 (U67088), TbxT (U67087); mouse T (X51683), Tbx2 (U15566), Tbx6 (U57331), T-brain-1 (S78858); human Tbx2 (U28049), Tbx5 (Y09445)

MyoD
- Arthropoda: *Drosophila melanogaster* MyoD (M68897)
- Chordata: Urochordata: ascidian (*Halocynthia roretzi*) AMD1 (D13507); ascidian (*Cionia intestinalis*) CiMDFa (U80079). Vertebrata: rainbow trout (*Oncorhynchus mykiss*) MyoD; zebrafish MyoD; clawed frog MyoD (X56677), MF25 (M31118), MYF5 (X56738), MYF6 (S34392); chicken MyoD (X16189), MYF5 (X75250), MyoG (M95800), MYF6 (D10599); mouse (M18779), MYF5 (X56182), MyoG (M95800), MYF6 (M30499); human MyoD (X56677), MYF5 (X14894), MyoG (X62155), MYF6 (X52011)

Notch
- Arthropoda: *Drosophila melanogaster* NOTCH (M16149-M16153), crumbs (M33753); blowfly (*Lucilia cuprina*) SCL (U58977)
- Chordata: Vertebrata: zebrafish NOTCH1 (X69088), goldfish (*Carassius auratus*) NOTCH3 (U09191); clawed frog NOTCH1 (M33874); mouse NOTCH1 (Z11886), NOTCH3 (X74760), NOTCH4 (U43691); rat NOTCH1 (X57405), jagged (L38483); human NOTCH1 (M73980), NOTCH2 (M99437), NOTCH4 (D63395), jagged (U61276)

Methods

Phylogenetic analyses were applied to seven protein families—Cdx, BMP, Elav, Egr/SP, Brachyury, MyoD, and Notch; sequences analyzed are listed in Table 1. Six of these families are among the seven families with four vertebrate paralogues listed by Sidow (1996) in the passage quoted above. Sidow (1996) also mentions the *Hox* gene clusters. These have been subjected to phylogenetic analysis in two recent studies (Zhang and Nei 1996; Bailey et al. 1997); therefore, they were not included here. Two of the seven families include two subfamilies, each of which contains four vertebrate paralogues: dpp and BMP5-8 in the BMP family and Egr and SP in the Egr/SP family. Therefore, the number of sets of four vertebrate paralogues used to test the 2R hypothesis was nine.

Amino acid sequences were aligned using the CLUSTAL V program (Higgins et al. 1992); the alignments are available from the author upon request. In phylogenetic analyses, any amino acid site at which the alignment postulated a gap in any of the sequences was excluded from all pairwise comparisons; this was done so that a comparable set of data was used in each pairwise comparison. Because the sequences aligned were quite distantly related, in each case the alignment appeared reliable in only a certain conserved portion of the polypeptide; thus, phylogenetic analysis was applied only to this conserved area. For each family, only a subset of the sequences in the database was used in the phylogenetic trees presented here, although preliminary analyses included all available sequences. For ease of presentation in this paper, sequences were chosen to provide representatives of major taxonomic groups and to exclude highly divergent sequences for which the alignment was uncertain. Nonetheless, the results of preliminary analyses using larger data sets were essentially the same as those presented here (data not shown).

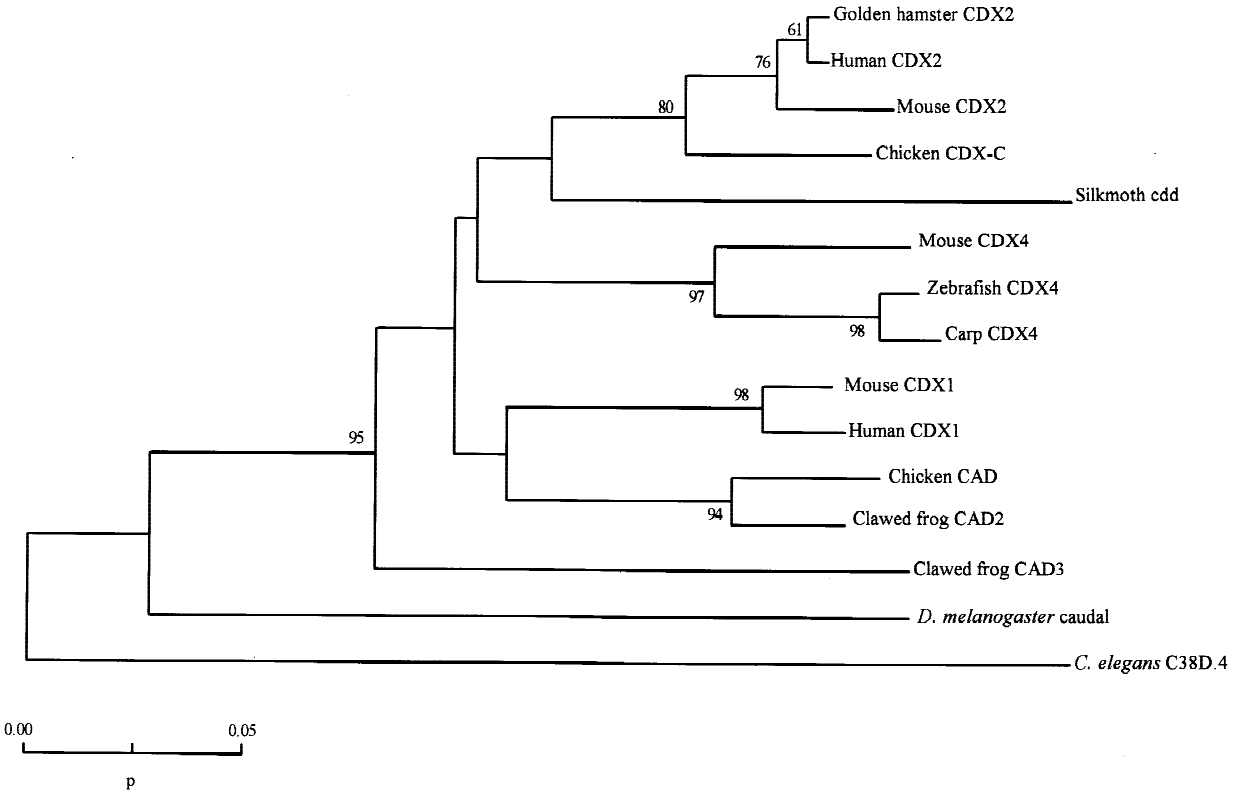

Fig. 2. Phylogenetic tree of the Cdx family. Numbers on branches represent the percentage of bootstrap samples supporting that branch; only values ≥50% are shown.

Here I briefly describe the seven gene families used in the analyses and the portion of the polypeptide analyzed. In most cases, one or more sequences could be used as an outgroup to root the tree, and I describe outgroups used.

Cdx. The Cdx family includes homeobox proteins expressed in early embryogenesis in both vertebrates and invertebrates (Doll and Niessing 1993; Hu et al. 1993; Mlodzik and Gehring 1987). Phylogenetic analysis was based on the conserved homeodomain (49 aligned residues). The tree was rooted by using as an outgroup a Cdx homologue from the nematode worm *Caenorhabditis elegans.*

BMP. The bone morphogenetic proteins (BMP) of vertebrates, members of the transforming growth factor β (TGF-β) superfamily, are involved in regulating the growth of bone and certain other organs (Celeste et al. 1990; Oh et al. 1996). Homologous genes have been found to play a role in development in *Drosophila* also (Arora et al. 1994; Padgett et al. 1987; Wharton et al. 1996). Phylogenetic analysis was based on the conserved C-terminal region of the protein (177 aligned residues), which contains the conserved TGF-β homology region (Wharton et al. 1991). No outgroup was used to root the tree, but the root was placed in the midpoint of the longest internal branch. However, because the vertebrate genes consisted of two subfamilies (designated BMP5-8 and dpp), each subfamily could be used to root the other subfamily.

Elav. The *Drosophila elav* gene is required for the development of neurons (Robinow et al. 1988), while the distantly related *sex-lethal* controls sex determination and dosage compensation (Penalva et al. 1996). Along with certain vertebrate genes, these genes belong to a family encoding proteins believed to regulate developmental processes posttranscriptionally through a role in RNA metabolism (Ma et al. 1996; Perron et al. 1995). Phylogenetic analysis was based on the conserved C-terminal portion of the protein (218 aligned residues), which includes the putative RNA-binding sites RNP1 and RNP2 (Robinow et al. 1988). The tree was rooted with insect sex-lethal proteins.

Egr/SP. This family includes zinc-finger proteins that act as transcription factors for a wide variety of genes. In *Drosophila*, the stripe b gene is involved in head segmentation (Wimmer et al. 1993), while vertebrate members of this family are involved in differentiation of a variety of cell types including those of the nervous and immune systems (Kingsley and Winoto 1992; Milbrandt 1987; Supp et al. 1996). The btd protein of *Drosophila*, which is related to these Egr and SP, was not included in phylogenetic analysis because in preliminary analysis it showed only very low sequence similarity to the vertebrate Egr and SP, to *Drosophila* stripe b, and to related proteins of *C. elegans* (data not shown). The analysis was based on the conserved zinc-finger region (Supp et al. 1996) (88 aligned residues). Although the tree was unrooted, there were two subfamilies (Egr and SP), each of which served to root the other.

Brachyury. The vertebrate *Brachyury* or *T* gene, which is essential for notochord formation, encodes a DNA-binding protein (Kispert and Herrmann 1993). Insect homologues are expressed throughout embryogenesis, particularly in the hindgut (Kispert et al. 1993). Phylogenetic analysis was based on the conserved DNA-binding domain (Kispert et al. 1994) (164 aligned residues). The tree was rooted with a number of homologues from *C. elegans.*

MyoD. The MyoD family includes DNA-binding proteins involved in development of muscle and certain other tissues in both vertebrates and invertebrates (Hopwood et al. 1989; Krause et al. 1990; Miner and Wold 1990; Paterson et al. 1991). Phylogenetic analysis was based on

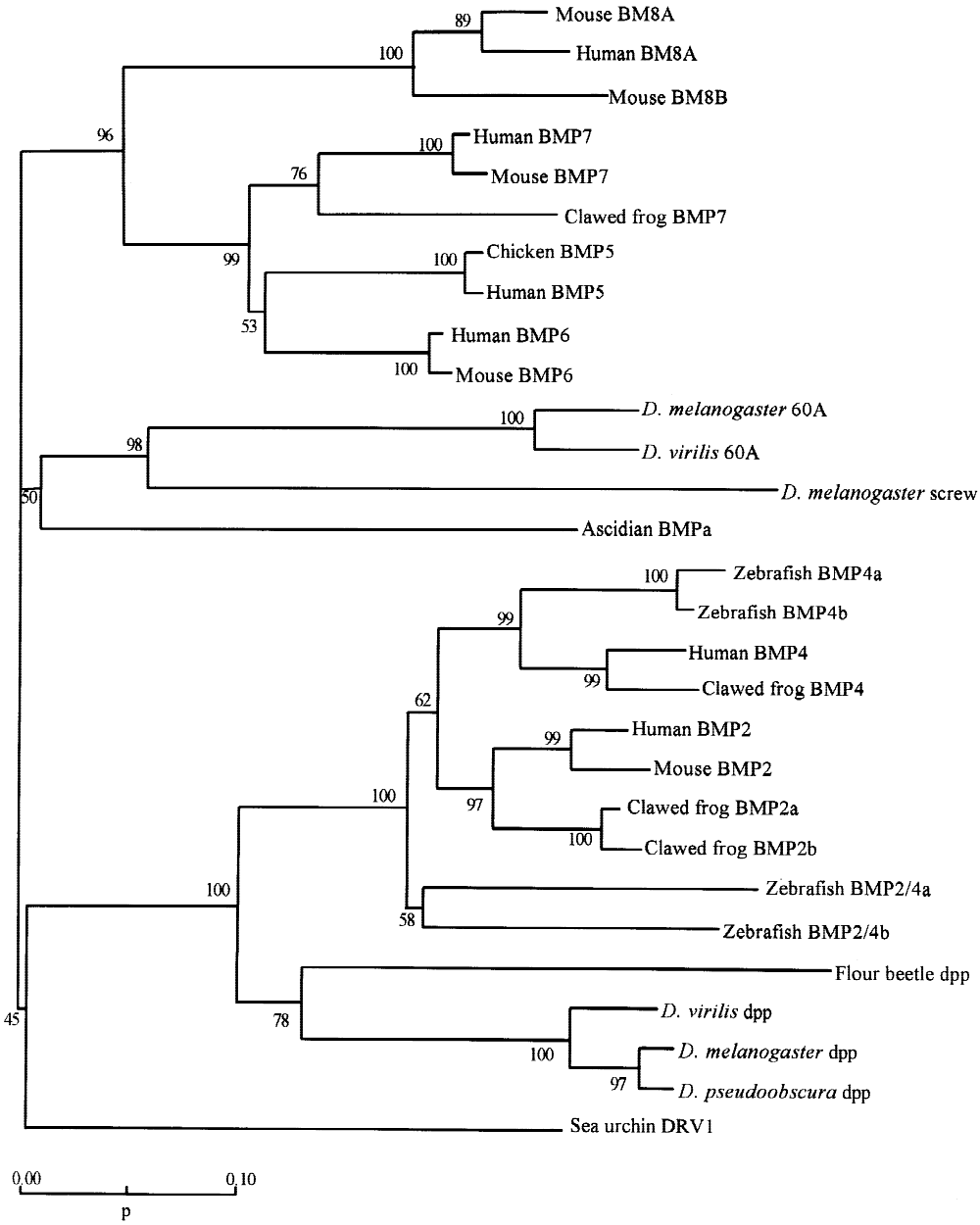

Fig. 3. Phylogenetic tree of the BMP family. Numbers on branches are as in Fig. 2.

the central region of the protein, including the conserved and functionally important basic and Myc-like regions (Hopwood et al. 1989) (112 aligned residues). In the absence of an outgroup, the tree was rooted in the midpoint of the longest internal branch.

Notch. *Drosophila* Notch and its vertebrate homologues are involved in the development of many tissues, playing a role in cell–cell interaction (Larsson et al. 1994; Wharton et al. 1985). The phylogenetic tree was based on the conserved central region of the polypeptide (172 aligned residues). The tree was rooted with *Drosophila crumbs* and vertebrate homologues.

Phylogenetic trees were constructed by the maximum-parsimony (MP) method (Swofford 1990) and by the neighbor-joining (NJ) method based on the following three distances: the uncorrected proportion of amino acid difference (p), the Poisson-corrected estimate of the number of amino acid replacements per site (Nei 1987), and the gamma-corrected estimate of the number of amino acid replacements per site (Ota and Nei 1994). All of the methods yielded essentially the same results; therefore, only NJ trees based on p are presented here. Trees based on p are preferable when the sequences involved are very distantly related, as is true in this case, because its variance is lower than that of other distances (Kumar et al. 1993). The reliability of branches in the phylogenetic trees was assessed by bootstrapping (Felsenstein 1985), which involves repeated sampling from the data with replacement and construction of a tree based on each sample; 1000 bootstrap samples were used.

Results

Cdx. Presumably because the number of sites available for analysis of this family was quite limited, the phylogenetic analysis of Cdx and related proteins showed poor

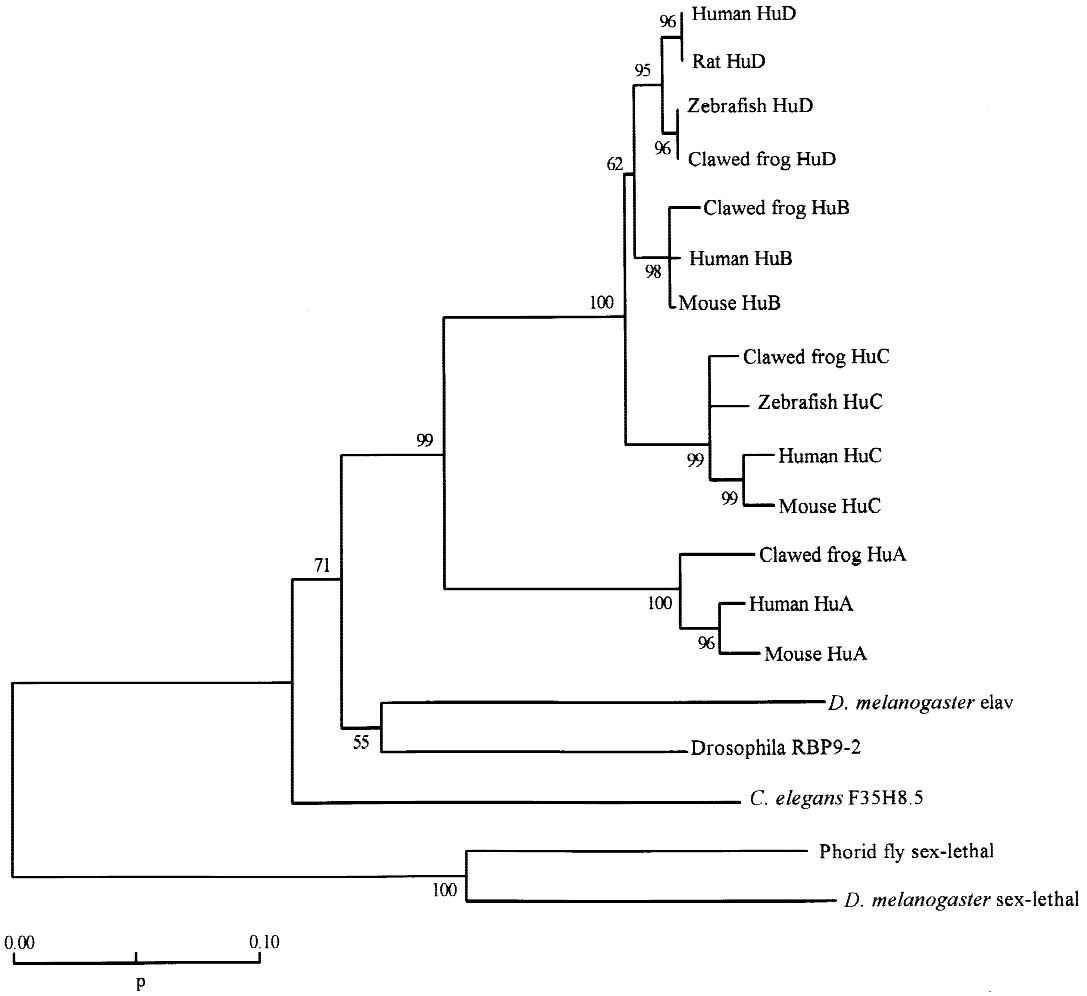

Fig. 4. Phylogenetic tree of the Elav family. Numbers on branches are as in Fig. 2.

resolution. The silkworm cdd clustered with the vertebrate CDX2 and CDX4 sequences (Fig. 2). This topology suggests that the gene duplication giving rise to the ancestor of CDX2 and CDX4 may have occurred prior to the divergence of deuterostomes and protostomes. However, bootstrap support for clustering of silkworm cdd with vertebrate CDX2 and CDX4 was quite low (44%). Even if the clustering of silkworm cdd among vertebrate genes could be attributed to stochastic error, the tree still would not support the 2R hypothesis. Rather, the tree's topology would be of the (A) (BCD) type, because clawed frog CAD3 clustered outside the other family members from vertebrates (Fig. 2). But again, bootstrap support for this topology was low (47%).

BMP. The phylogenetic tree of BMP-related molecules (Fig. 3) contained three major clusters: (1) a cluster including vertebrate BMP5, BMP6, BMP7, and BMP8, referred to here as the "BMP5-8 subfamily"; (2) a cluster including insect dpp as well as vertebrate BMP2 and BMP4, referred to here as the "dpp subfamily"; and (3) a cluster containing *Drosophila* 60A and screw proteins. The position of ascidian BMPa and sea urchin DRV1 relative to these major clusters was not well resolved (Fig. 3). Within the BMP5-8 subfamily, the vertebrate genes showed a topology of the form (A) (BCD) (Fig. 3). Vertebrate BMP8 fell outside the other vertebrate members of the subfamily, and the branch supporting this pattern received highly significant (99%) bootstrap support (Fig. 3). In the dpp subfamily, the vertebrate members showed a topology of the form (AB) (CD); zebrafish BMP2/4a clustered with BMP2/4b, while BMP2 of various vertebrates clustered with BMP4 (Fig. 3). However, support for this pattern was quite weak, bootstrap percentages for the two relevant branches being 58 and 62% (Fig. 3).

Elav. In the Elav family, a topology of the form (A) (BCD) received strong bootstrap support (1005), with HuA falling outside the cluster of HuB, HuC, and HuD (Fig. 4).

Egr/SP. Vertebrate members of this family formed two major clusters: (1) the Egr subfamily, including

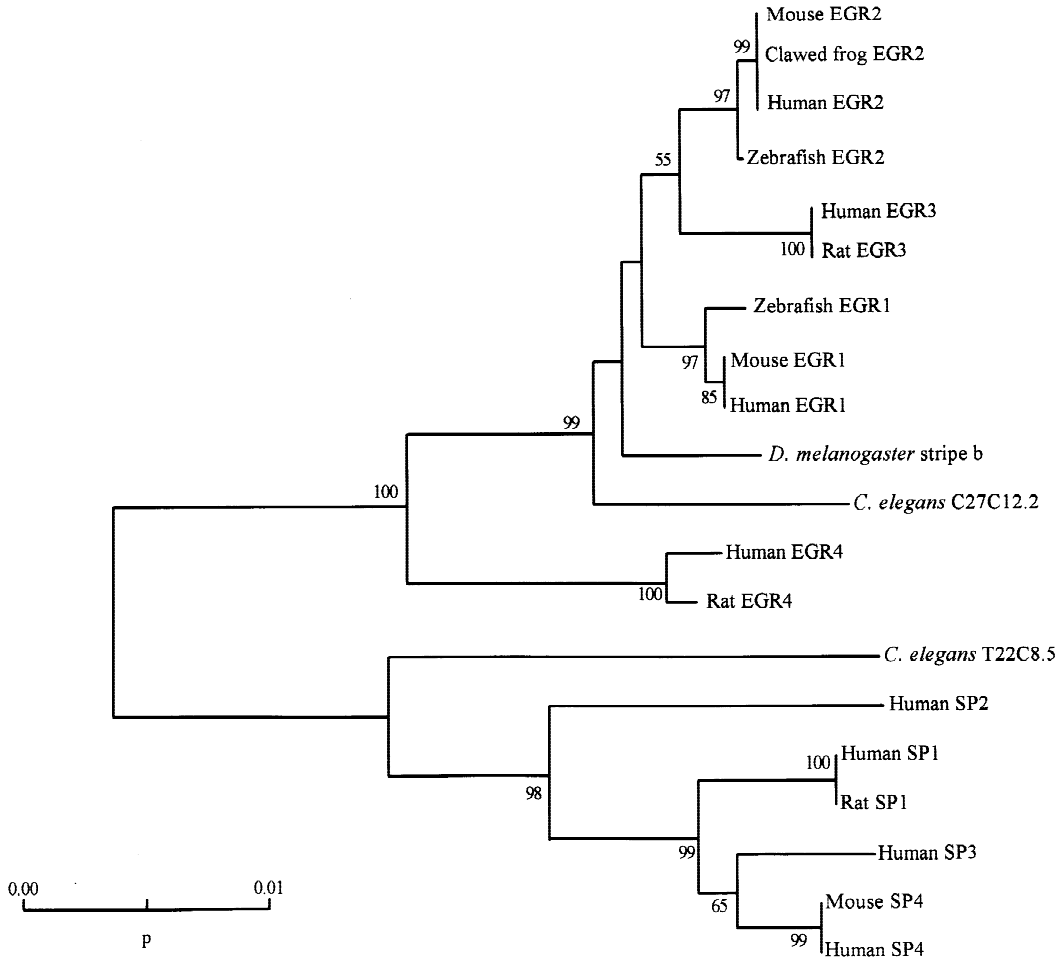

Fig. 5. Phylogenetic tree of the Egr/SP family. Numbers on branches are as in Fig. 2.

EGR1-4; and (2) the SP subfamily, including SP1-4 (Fig. 5). Vertebrate members of the Egr subfamily showed strong evidence of duplication prior to the divergence of deuterostomes and protostomes; EGR4 grouped outside the cluster of EGR1-3 and homologues from *Drosophila* and *C. elegans*, and the branch supporting this topology received highly significant bootstrap support (99%) (Fig. 5). In the SP subfamily, the topology was of the form (A) (BCD). SP2 clustered outside the other three vertebrate molecules, and this pattern received significant bootstrap support (98%) (Fig. 5).

Brachyury. In the Brachyury family, the vertebrate genes fell into two major groups, and the topology of the phylogenetic tree (Fig. 6) indicated that the gene duplication giving rise to these groups occurred prior to the divergence of deuterostomes and protostomes. Vertebrate and ascidian T sequences clustered with Drosophila tbx, and this pattern received highly significant bootstrap support (100%) (Fig. 6). Similarly, *Drosophila* omb and a sequence from *C. elegans* clustered with vertebrate Tbx2, again a pattern that received strong bootstrap support (99%) (Fig. 6).

MyoD. Although an outgroup was lacking to root the MyoD tree, the tree supported the hypothesis that two major clusters of MyoD genes arose before the origin of vertebrates. A significant internal branch (96% bootstrap support) separated the following: (1) a cluster including vertebrate MyoD, vertebrate MYF5, and *Drosophila* MyoD and (2) a cluster including vertebrate MyoG, vertebrate MYF6, and ascidian homologues (Fig. 7). However the tree is rooted; this implies that the divergence of these two groups occurred at least before the separation of ascidians (Urochordata) from vertebrates.

Notch. The phylogenetic tree placed vertebrate NOTCH4 outside the cluster of vertebrate NOTCH1-3 and insect NOTCH, with a highly significant branch (100% bootstrap support) (Fig. 8). Thus, the tree supported the hypothesis that NOTCH4 diverged from other vertebrate NOTCH family members prior to the divergence of protostomes and deuterostomes.

Discussion

Since only phylogenies having a topology of the form (AB) (CD) provide explicit support for the 2R hypoth-

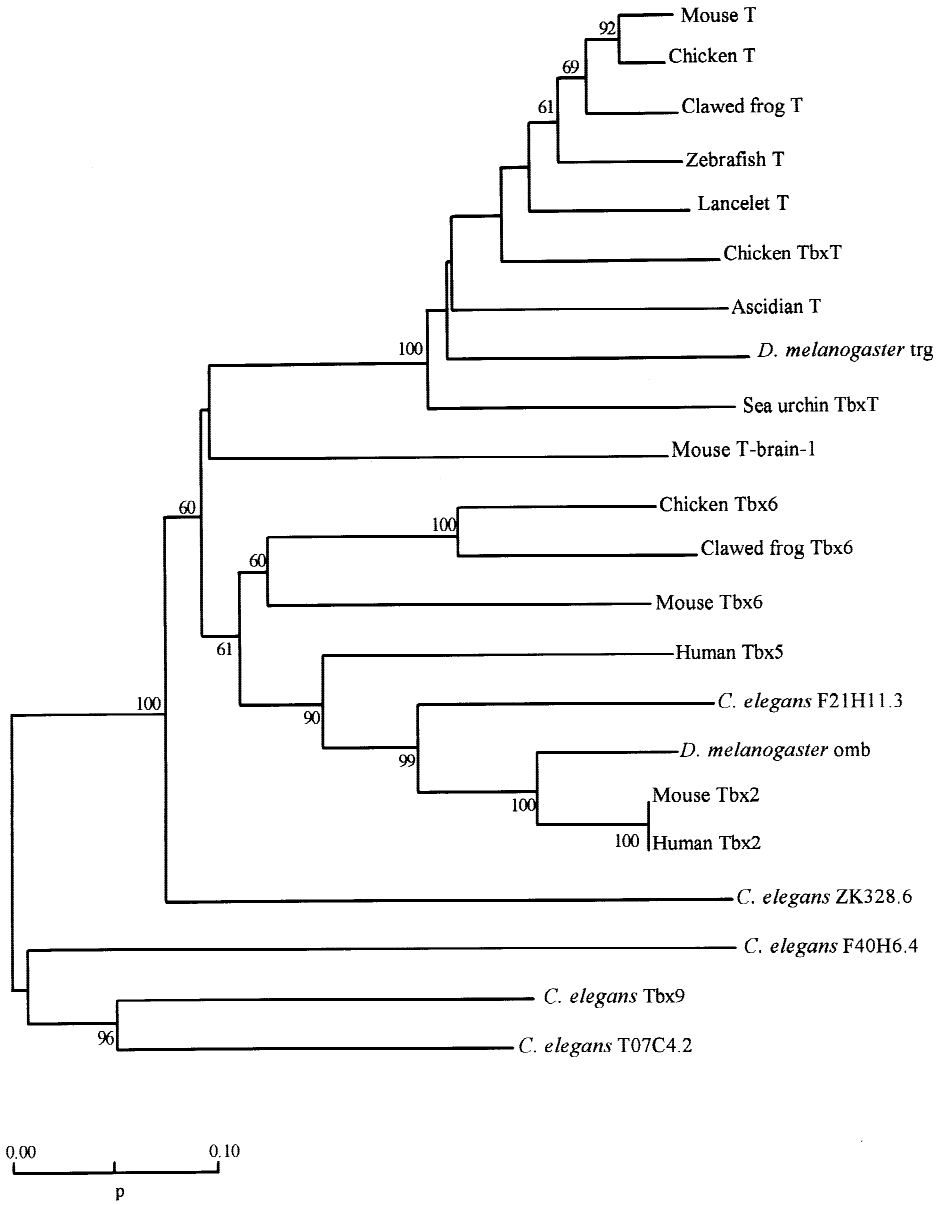

Fig. 6. Phylogenetic tree of the Brachyury family. Numbers on branches are as in Fig. 2.

esis, the present analyses provided essentially no support for this hypothesis. This topology occurred only once (in the dpp subfamily of the BMP family; Fig. 3), and it received very weak bootstrap support. Interestingly, in other recently reported analyses of other developmentally important families, topologies of the (AB) (CD) form were not found. Table 2 summarizes the results of 13 independent phylogenies of developmentally important gene families from the present study and others. The other families considered are the fibroblast growth factor receptor (FGFR) genes, antennapedia-class homeobox genes (antp), hox-linked collagen (COL) genes, and Pax genes (Table 2). Of the 13 phylogenies, only dpp had a topology of the form (AB) (CD).

In contrast, seven phylogenies were of the form (A) (BCD) (Table 2). Five of these received significant bootstrap support (>95%), while one received 93% bootstrap support. Furthermore, the five remaining phylogenies supported the hypothesis that the vertebrate family members initially diverged prior to vertebrate origins (Table 2). In four cases, the duplication clearly took place prior to the divergence of deuterostomes and protostomes, while in the other case (MyoD) it may have taken place early in the chordate lineage before the divergence of Urochordata and Vertebrata. Also, in four of the five cases, the branch supporting a duplication before the origin of vertebrates received significant bootstrap support (Table 2).

Therefore, available data from protein phylogenies do not support the 2R hypothesis. Other data that might be relevant to this hypothesis include estimates of gene number in different organisms and estimates of DNA

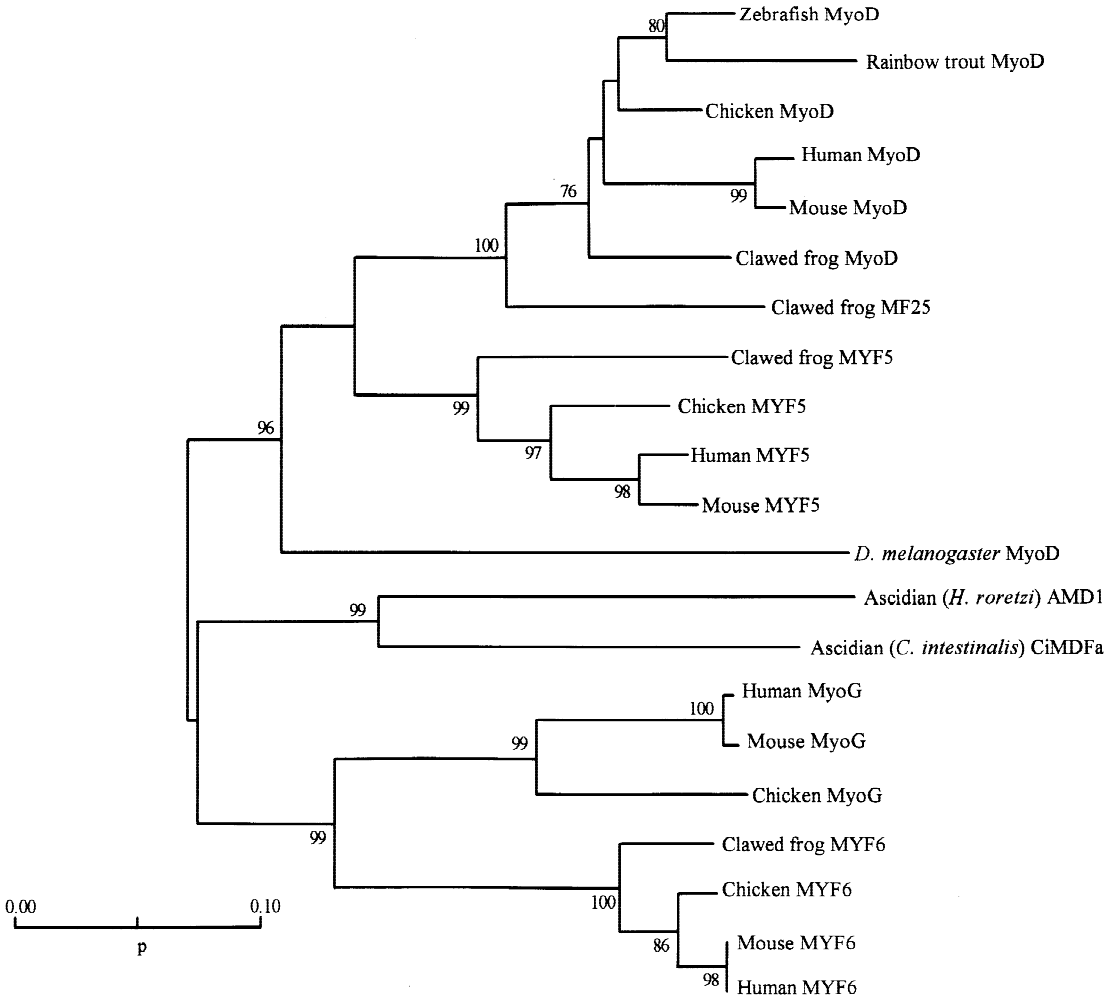

Fig. 7. Phylogenetic tree of the MyoD family. Numbers on branches are as in Fig. 2.

content per diploid nucleus (measured either as the number of base pairs in the genome or as DNA mass). However, both of these types of data are problematic as potential sources of information about past events of genome duplication. After genome duplication, population genetics theory predicts that duplicate loci that do not evolve new functions, and thus are redundant, will be silenced as a result of mutation, a prediction supported by studies of electrophoretic variation in animal species that have undergone recent polyploidization (Ferris and Whitt 1977; Li 1980). Likewise, noncoding DNA may be lost over time as a result of deletions. Thus, after a long evolutionary time, neither gene number nor genome size may reliably reflect past genome duplication events.

Furthermore, it is clear that the relationship between gene number and genome size is not a simple one. For example, the bony fish *Fugu rubripes* is estimated to have the same number of genes (70,000) as mammals such as human and mouse, yet its genome is only about 12% as large (400 vs. 3300 megabases) as that of human or mouse (Miklos and Rubin 1996). However, there is a good linear relationship between the logarithm of gene number and that of genome size for a wide range of organisms, including bacteria, invertebrates, and vertebrates (Fig. 9). Deviations from this overall trend are so far poorly understood. The issue is further complicated by the fact that, at least in certain cases, there may be adaptive aspects to genome size (Szarski 1970; Olmo et al. 1989; Hughes and Hughes 1995).

As mentioned previously, typical vertebrates are estimated to have about 5.8 times as many genes as does *Drosophila*. Known diploid genome sizes of insects cover an extraordinarily large range, from 0.3 pg/.nucleus in dixid flies to over 25.0 pg/nucleus in acridid grasshoppers (Finston et al. 1995). With a genome size of 0.36 pg/nucleus (Rasch et al. 1971), *Drosophila melanogaster* has one of the smallest genomes known from insects, about 20 times smaller than a typical vertebrate genome. Thus it is possible that *Drosophila* has an unusually small number of genes as well. Evidence that this may be so comes from the fact that *Drosophila* is estimated to have fewer genes even than the morphologically much simpler *C. elegans* (Miklos and Rubin 1996). Thus *Drosophila* seems a poor comparison for reconstructing possible events of genome duplication early in vertebrate history.

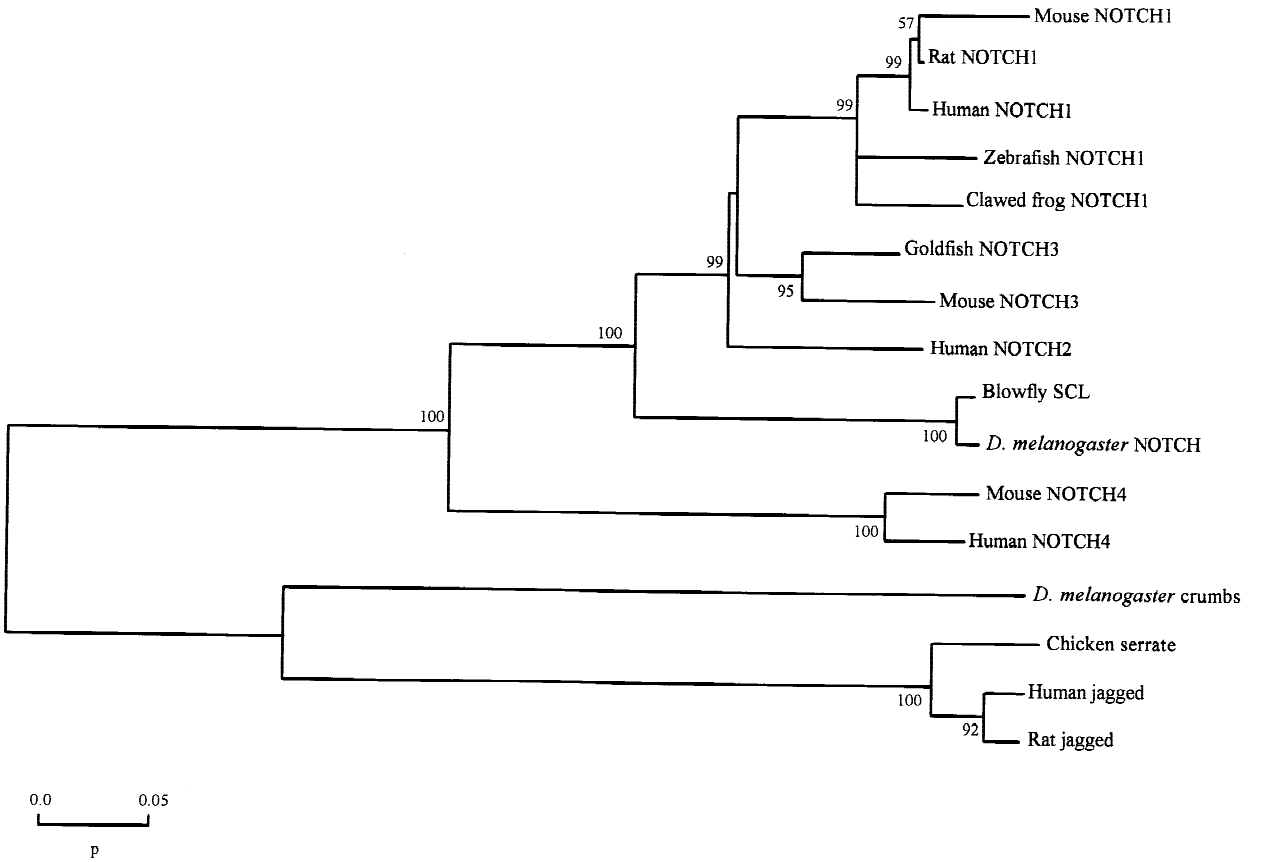

Fig. 8. Phylogenetic tree of the Notch family. Numbers on branches are as in Fig. 2.

Table 2. Summary of phylogenetic analyses of gene families having four members in vertebrates

Gene	Duplication before vertebrate origin[a]	Topology[b]	Source
CDX	+ (47)		This study
dpp	−	(AB) (CD) (58, 62)	" "
BMP5-8	−	(A) (BCD) (99)	" "
Elav	−	(A) (BCD) (100)	" "
Egr	−	(A) (BCD) (97)	" "
SP	−	(A) (BCD) (96)	" "
Brachyury	+ (100)		" "
MyoD	+ (96)		" "
NOTCH	+ (100)		" "
FGFR	−	(A) (BCD) (98)	Coulier et al. (1997)
antp	−	(A) (BCD) (<50)	Zhang and Nei (1996)
Hox-linked COL	−	(A) (BCD) (93)	Bailey et al. (1997)
Pax	+ (99)		Balczarek et al. (1997)

[a] Percentage bootstrap support in parentheses. For all genes but MyoD, the duplication could be shown to have occurred before the divergence of protostomes and deuterostomes.

[b] Topologies are shown in Fig. 1. Figure 1A represents the (AB) (CD) topology, while Fig. 1B represents an example of (A) (BCD) topology. Percentage bootstrap support in parentheses; for (AB) (CD) topology bootstrap percentages for branches 1 and 2 (as in Fig. 1A) are given.

Although the available evidence does not support the 2R hypothesis, the phylogenies presented here are all consistent with the hypothesis that a single genome duplication occurred either at some point in deuterostome history before the origin of vertebrates or within the vertebrate lineage shortly after its origin. However, the available data can be easily explained without hypothesizing any genomewide duplication event; known gene phylogenies can be explained by independent duplication of individual genes or chromosomal segments, processes well known to occur in eukaryotic genomes. Indeed it will prove difficult to test the hypothesis of a single

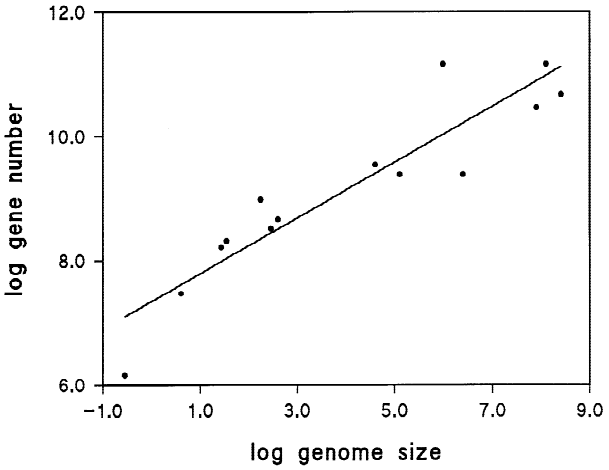

Fig. 9. Relationship between the natural logarithm of estimated gene number and the natural logarithm of genome size (Mbp). The line shown is the linear regression line $Y = 7.343 + 0.448X$ ($R^2 = 0.869$). The data are from Miklos and Rubin (1996).

genome duplication in vertebrate ancestry because it is difficult to devise ways to discriminate between this hypotheses and alternatives. Probably the best evidence will come from reliable estimates of gene number in nonchordate deuterostome phyla, in chordate subphyla other than vertebrates, and in early-branching vertebrate taxa such as Agnatha and Chondryichthyes. So far, no such data are available. Therefore, the hypothesis that genome duplication has played a key role in vertebrate evolution remains entirely speculative.

Acknowledgments. This research was supported by Grants R01-GM34940 and K04-GM00614 from the National Institutes of Health.

References

Arora K, Levine MS, O'Connor MB (1994) The *screw* gene encodes a ubiquitously expressed member of the TGF-β family required for specification of dorsal cell fates in the *Drosophila* embryo. Genes Dev 8:2588–2601

Bailey WJ, Kim J, Wagner GP, Ruddle FH (1997) Phylogenetic reconstruction of the vertebrate Hox cluster duplication. Mol Biol Evol 14:843–853

Balczarek KA, Lai Z-C, Kumar S (1997) Evolution and functional diversification of the Paired Box (*Pax*) DNA-binding domains. Mol Biol Evol 14:829–842

Celeste AJ, Iannazi JA, Taylor RC, et al. (1990) Identification of transforming growth factor β family members present in bone-inductive protein purified from bovine bone. Proc Natl Acad Sci USA 87: 9843–9847

Coulier F, Pontarotti P, Roubin R, Hartung H, Goldfarb M, Birnbaum D (1997) Of worms and men: an evolutionary perspective on the fibroblast growth factor (FGF) and FGF receptor families. J Mol Evol 44:43–56

Doll U, Niessing J (1993) Continued expression of the chicken *caudal* homologue in endodermally derived organs. Dev Biol 156:155–163

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:95–105

Ferris SD, Whitt GS (1977) Loss of duplicate gene expression after polyploidisation. Nature 265:258–260

Finston TL, Hebert PDN, Foottit RB (1995) Genome size variation in aphids. Insect Biochem Mol Biol 25:189–196

Higgins DG, Bleasby AJ, Fuchs R (1992) Clustal V: improved software for multiple sequence alignment. Comput Appl Biosci 8:189–191

Holland PWH, Garcia-Fernandez J, Williams NA, Sidow A (1994) Gene duplications and the origins of vertebrate development. Development 1994 Suppl:125–133

Hopwood ND, Pluck A, Gurdon JB (1989) MyoD expression in the forming somites is an early response to mesoderm induction in *Xenopus* embryos. EMBO J 8:3409–3417

Hu Y, Kazenwadel J, James R (1993) Isolation and characterization of the murine homeobox gene *Cdx-1*. J Biol Chem 268:27214–27225

Hughes AL, Hughes MK (1995) Small genomes for better flyers. Nature 377:391

Kasahara M, Hayashi M, Tanaka K, et al. (1996) Chromosomal localization of the proteasome Z subunit gene reveals an ancient chromosomal duplication involving the major histocompatibility complex. Proc Natl Acad Sci USA 93:9096–9101

Kingsley C, Winoto A (1992) Cloning of GT box-binding proteins: a novel Sp1 multigene family regulating T-cell receptor gene expression. Mol Cell Biol 12:4251–4261

Kispert A, Herrmann BG (1993) The *Brachyury* gene encodes a novel DNA binding protein. EMBO J 12:3211–3220

Kispert A, Herrmann BG, Leptin M, Reuter R (1994) Homologs of the mouse *Brachyury* gene are involved in the specification of posterior terminal structures in *Drosophila*, *Tribolium*, and *Locusta*. Genes Dev 8:2137–2150

Krause M, Fire A, Harrison SW, Priess J, Weintraub H (1990) CeMyoD accumulation defines the body wall muscle cell fate during C. elegans embryogenesis. Cell 63:907–918

Kumar S, Tamura K, Nei M (1993) MEGA: molecular evolutionary genetic analysis, Version 1.0. Pennsylvania State University, University Park

Larsson KA, Johansen KM, Xu T, Artavanis-Tsakonas (1985) The human *NOTCH1, 2,* and *3* genes are located at chromosome positions 9q34, 1p13-p 11 and 19p13.2-p13.1 in regions of neoplasia-associated translocation. Genomics 24:253–258

Li W-H (1980) Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fishes. Genetics 95:237–258

Ma W-J, Cheng S, Campbell C, Wright A, Furneaux H (1996) Cloning and characterization of HuR, a ubiquitously expressed elav-like protein. J Biol Chem 271:6144–6151

Miklos GLG, Rubin GM (1996) The role of the genome project in determining gene function: insight from model organisms. Cell 86:521–529

Milbrandt J (1987) A nerve growth factor-induced gene encodes a possible transcriptional regulatory factor. Science 238:797–799

Miner JH, Wold B (1990) Herculin, a fourth member of the *MyoD* family of myogenic regulatory genes. Proc Natl Acad Sci USA 87:1089–1093

Mlodik M, Gehring WJ (1987) Expression of the *caudal* gene in the germ line of Drosophila: formation of an RNA and protein gradient during early embryogenesis. Cell 48:465–478

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Oh S-H, Johnson R, Wu DK (1996) Differential expression of bone morphogenetic proteins in the developing vestibular and auditory sensory organs. J Neurosci 16:6463–6475

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

Olmo E, Capriglione T, Odierna G (1989) Genome size evolution in vertebrates: trends and contrasts. Comp Biochem Physiol 92B:447–453

Ota T, Nei M (1994) Estimation of the number of amino acid substi-

tutions per site when the substitution rate varies among sites. J Mol Evol 38:642–643

Padgett RW, St. Johnston RD, Gelbart WM (1987) A transcript from a *Drosophila* pattern gene predicts a protein homologous to the transforming growth factor-β family. Nature 325:81–84

Patterson BM, Walldorf U, Eldridge J, Dubendorfer A, Frasch M, Gehring WJ (1991) The *Drosophila* homologue of vertebrate myogenic-determination genes encodes a transiently expressed nuclear protein marking primary myogenic cells. Proc Natl Acad Sci USA 88:3782–3786

Penalva LOF, Sakamoto H, Navarro-Sabate A, et al. (1996) Regulation of the gene *sex-lethal:* a comparative analysis of *Drosophila melanogaster* and *Drosophila subobscura.* Genetics 144:1653–1664

Perron M, Theodore L, Wegnez M (1995) Isolation and embryonic expression of *Xel-1,* a nervous system-specific *Xenopus* gene related to the elav family. Mech Dev 51:235–249

Rasch EM, Barr HJ, Rasch RW (1971) The DNA content of sperm of *Drosophila melanogaster*. Chromosoma 33:1–18

Robinow S, Campos AR, Yao K-M, White K (1988) The elav gene product of Drosophila, required in neurons, has three RNP consensus motifs. Science 242:1570–1572

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

Supp DM, Witte DP, Branford WW, Smith EP, Potter SS (1996) Sp4, a member of the Sp1-family of zinc finger transcription factors, is required for normal murine growth, viability, and male fertility. Dev Biol 176:284–299

Swofford DL (1990) PAUP: phylogenetic analysis using parsimony. Illinois Natural History Survey, Champaign

Szarski H (1970) Changes in the amount of DNA in cell nuclei during vertebrate evolution. Nature 226:651–652

Wharton KA, Johansen KM, Xu T, Artavanis-Tsakonas S (1985) Nucleotide sequence from the neurogenic locus notch implies a gene product that shares homology with proteins containing EGF-like repeats. Cell 43:567–581

Wharton KA, Thomsen GH, Gelbart WM (1991) *Drosophila* 60A gene, another transforming growth factor β family member, is closely related to human bone morphogenetic proteins. Proc Natl Acad Sci USA 88:9214–9218

Wimmer EA, Jackie H, Pfeifle C, Cohen SM (1993) A Drosophila homologue of human Sp1 is a head-specific segmentation gene. Nature 366:690–694

Zhang J, Nei M (1996) Evolution of antennapedia-class homeobox genes. Genetics 142:295–303