J Mol Evol (2000) 51:88–96
DOI: 10.1007/s002390010069

JOURNAL OF MOLECULAR EVOLUTION

Evolutionary Patterns of Gene Families Generated in the Early Stage of Vertebrates

Yufeng Wang, Xun Gu

Department of Zoology and Genetics, Center for Bioinformatics and Biological Statistics, Program of Bioinformatics and Computational Biology (BCB), 332 Science II, Iowa State University, Ames, IA 50011, USA

Received: 13 December 1999 / Accepted: 7 April 2000

Abstract. In this paper we have analyzed 49 vertebrate gene families that were generated in the early stage of vertebrates and/or shortly before the origin of vertebrates, each of which consists of three or four member genes. We have dated the first (T_1) and second (T_2) gene duplications of 26 gene families with 3 member genes. The means of T_1 (594 mya) and T_2 (488 mya) are largely consistent to a well-cited version of two-round (2R) genome duplication theory. Moreover, in most cases, the time interval between two successive gene duplications is large enough that the fate of duplicate genes generated by the first gene duplication was likely to be determined before the second one took place. However, the phylogenetic pattern of 23 gene families with 4 members is complicated; only 5 of them are predicted by 2R model, but 11 families require an additional gene (or genome) duplication. For the rest (7 families), at least one gene duplication event had occurred before the divergence between vertebrate and *Drosophila*, indicating a possible misleading of the 4:1 rule (member gene ratio between vertebrates and invertebrates). Our results show that Ohno's 2R conjecture is valid as a working hypothesis for providing a most parsimonious explanation. Although for some gene families, additional gene duplication is needed, the credibility of the third genome duplication (3R) remains to be investigated.

Key words: Gene (genome) duplication — Vertebrate evolution — Gene family — Duplication time dating

Introduction

Gene duplication is believed to play an important role during evolution by providing opportunities to evolve new gene functions that can lead to novel morphologies and physiologies (Ohno 1970; Li 1983). Sequence accumulation by genome projects has shown a rapid increase of gene families in the early stage of vertebrates, which implies the existence of large-scale gene (genome) duplication(s) during this evolutionary period (Holland et al. 1994; Kasahara et al. 1996; Endo et al. 1997; Nadeau and Sankoff 1997; Ruddle 1997; Pebusques et al. 1998; Suga et al. 1999). Because the ratio of member genes of human to fruitfly in these gene families is typically 2:1, 3:1, or 4:1, a two-round genome duplication theory in the early lineage of vertebrates has become popular (e.g., Lundin 1993; Holland et al. 1994; Fryxell 1996; Sidow 1996; Spring 1997).

In spite of its simplicity and elegance (e.g., in the case of four mammalian Hox gene clusters), recent studies have raised several questions about the classical two-round genome duplication (2R) theory (Skrabanek and Wolfe 1998; Hughes 1999; Smith et al. 1999). Their argument can be briefly described as Fig. 1B, where the 2R hypothesis predicts a phylogenetic pattern ((A,B),(C,D)). This prediction does not always hold in those gene families that were generated in the early stage of vertebrates (for a recent review, see Skrabanek and

Correspondence to: X. Gu; *e-mail:* xgu@iastate.edu

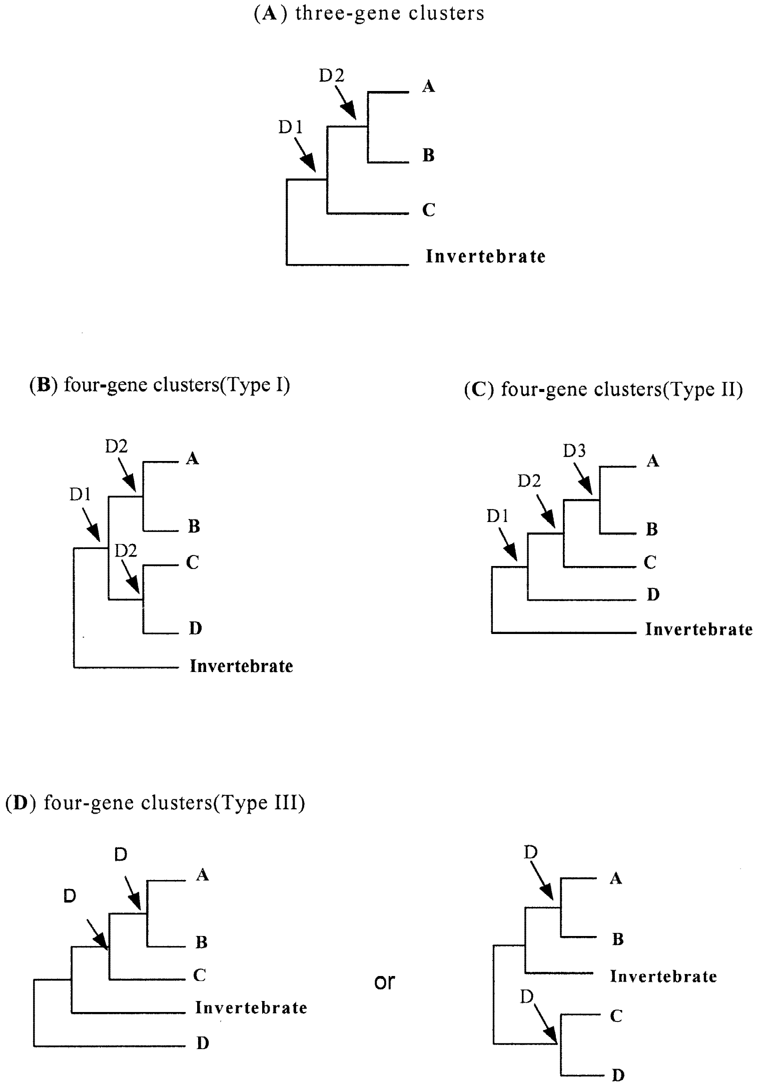

Fig. 1. Phylogenies of gene families that have undergone gene duplications in the early vertebrate stage. *Drosophila melanogaster* or/and *C. elegans* were used as an invertebrate outgroup. **A** A topology of gene families having three gene clusters, which were likely generated by two gene (genome) duplications (D1 and D2, respectively). Panels **B, C, D** show three possible phylogenetic patterns of gene families with four member genes. **B** Type I topology that has been possibly arisen from two rounds of gene duplications. **C** Type II topology that has been possibly arisen from three-round genome duplications (D1, D2, and D3). **D** Type III topology in which at least one gene duplication had occurred before the divergence of invertebrates and vertebrates.

Wolfe 1998). For example, if the phylogenetic pattern is (((A,B), C), D), an additional gene duplication is needed so that the hypothesis of three-round genome duplication (3R) is more favorable (Fig. 1C). However, testing 2R or 3R theory is complicated by the fact that in many cases, two phylogenetic patterns, ((A,B),(C,D)) and (((A,B), C), D), are not statistically distinguishable or the topology is sensitive to the tree-making method.

Some authors (e.g., Spring 1997) attempted to identify vertebrate gene families due to 2R genome duplications using only the synteny information and the member gene ratio between vertebrates and invertebrates. Phylogenetic analysis of vertebrate gene families has shown that such an approach can be misleading (Hughes 1999). However, the view of Hughes (1999) that the role of genome duplication(s) in vertebrate evolution has been disapproved seems to be extreme. Although the classical 2R model may be oversimplified, the concept of genome duplication (Ohno 1970) in early vertebrates is still valid as a working hypothesis. Otherwise, we are faced with the difficulty to explain the big-bang fashion of tissue-specific isoforms in the early stage of vertebrates. Indeed, in the human genome, a significant portion of gene families with two or three members is generated in the early vertebrate lineage, which are consistent with Ohno's 2R hypothesis: one or more member genes may have been lost (Nadeau and Sankoff 1997). In this paper, we interpret the genome duplication hypothesis as only a few large-scale chromosome- or genome-wide duplication(s) can explain the generation of many duplicated genes during a relatively short time period in the early stage of vertebrate evolution. Although at this moment, it may be difficult to provide conclusive evidence for 2R, 3R, or more round genome duplications (Skrabanek and Wolfe 1998), our understanding for vertebrate genome evolution can be significantly improved by exploring the evolutionary pattern of these gene families in a large scale.

Materials and Methods

Amino acid sequences of vertebrate gene families were obtained from several databases including Genbank, EMBL, and Hovergen (http://pbil.univ-lyon1.fr/). *Drosophila* and *C. elegans* homologues were obtained by BLAST-psi search. Two subsets of gene families were selected for our study.

Subset I: Gene Families with Three-Gene Clusters. From more than 1,100 gene families, we have obtained 112 gene families with three gene clusters. Among them, 26 gene families meet the following criteria: (1) Each gene family has three member genes. (2) These member genes are likely to be generated after the origin of vertebrates, which can be confirmed by using *Drosophila* and/or *C. elegans* homologues as an outgroup. (3) The *Drosophila* homologue can be reasonably considered as an orthologue of vertebrate gene clusters, so that the evolutionary distance between *Drosophila* and vertebrates can be used for calibration. (4) Amino acid sequences of each gene family should at least include species from mammals, birds or reptiles, and teleosts; and the phylogenetic tree should show a clear-cut evidence (i.e., bootstrapping value >70%) that these duplicate genes were generated before the divergence between teleosts and tetrapods. And (5) Gene families with less than 100 amino acids (e.g., HOX) are not used because of statistical uncertainty. They are listed as follows (the number in the parenthesis is the number of sequences): Adrenergic receptor $\alpha1$, ADRA1 (16); adrenergic receptor $\alpha2$, ADRA2 (29); adrenergic receptor β, ADRB (24); aldolase, ALDO (18); Alzheimer β-amyloid precursor protein, APLP (13); calponin (19); carboxypeptidase (14); caudal, CDX (14); cyclin D (15); glycogen phosphorylase, PYG (13); early growth response protein, EGR (12); fibroblast growth factor 8/17/18, FGF 8/17/18 (12); Fos (18); GLI (10); hedgehog, HH (20); insulin receptor, INSR (9); ITGA5/8/VNTR (10); inositol 1,4,5-triphosphate receptor, ITPR (10); JUN (16); MINOR (15); MYB (14); nitric oxidase synthase, NOS (13); REL (14); ROR (14); stimulating hormone receptor, SHR (26); zonula occludens, ZO (8). Accession numbers of these sequences are available on request.

Subset II: Gene Families with Four Gene Clusters. In total, 58 gene families with 4 clusters (member genes) have been found. To elucidate the phylogenetic pattern, we selected 23 4-cluster gene families with invertebrate homologues available. They are: activin β (20); cathepsin (20); cyclic nucleotide-gated cation channel, CNCG (18); discs large protein, DLG (12); epidermal growth factor receptor, EGFR (13); ELAV (20); ezrin (16); glycine receptor, GLR (12); ID (15); janus tyrosine kinase, JAK (16); myocyte enhancer factor 2, MEF2 (14); MAP kinase phosphatase, MKP (13); myogenic determining factor, MyoD (30); NOTCH (14); neuropeptide receptor, NPYR (20); octamer-binding protein, OCT (23); phosphodiesterase-4, PDE4 (10); 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase, PFKFB (15); peroxidase, PO (12); RAS (23); Src-related (13); syndecan (13); TRK (14).

It should be noted that the classification of phylogenetic patterns shown in Fig. 1 could be complicated by the possibility that some member genes have not been sequenced yet. However, we think that it may not be the serious problem for our data. Indeed, new member gene search has been extensively conducted in many gene families because they can be easily found through the human/mouse EST data. Keeping this problem in mind, here we simply assume the effect of incompleteness is trivial. This assumption can be tested in the near future when the complete human genome sequence is available.

For each gene family, amino acid sequences were aligned by CLUSTALW (Thompson et al. 1994). The phylogenetic tree was reconstructed by the neighbor-joining method (Saitou and Nei 1987). The root was determined by using the *Drosophila* and/or *C. elegans* homologous gene as an outgroup.

For each gene family with three-gene clusters, these two gene duplications were dated as follows. A linearized neighbor-joining tree (Takezaki et al. 1995) is used to convert the (average) distance to the geological time scale when several speciation events are used for calibrations. We have fully recognized that it is highly controversial about the geological dates of vertebrates between paleontologists and molecular evolutionists (e.g., Kumar and Hedges 1998; Gu 1998). Therefore, it is necessary to test whether our results are sensitive to the choice of calibration points. Fortunately, our preliminary results have shown that *the uncertainty of geological dates of vertebrates, i.e., primate-rodent (t_1 = 65–115 mya), and tetrapod-teleost (t_4 = 410–450 mya), has little effect on our time estimation of duplicate genes* (data not shown). For simplicity, middle values (i.e., $t_1 = 80$ mya and $t_4 = 430$ mya) were used in our study. However, our estimation is indeed sensitive to the vertebrate-*Drosophila* divergence time (t_5). Thus, two calibration systems were used to address this problem. The vertebrate calibrations are the same: primate-rodent ($t_1 = 80$ mya), mammal-bird ($t_2 = 310$ mya), mammal-amphibian ($t_3 = 350$ mya), tetrapod-teleost ($t_4 = 430$ mya), while the first calibration system uses the vertebrate-*Drosophila* ($t_5 = 830$ mya) as an upper bound, according to the molecular data (Gu 1998). The second calibration time uses $t_5 = 700$ mya as a lower bound, which seems reasonable because recent fossil records found in South China (Shu et al. 1999) imply that the first agnathans may have evolved in the earliest Cambrian (~555 mya). Our calibration systems minimize the effect of rate variation among vertebrate lineages since duplication events have occurred within the range of the calibration times, i.e., after the divergence time of vertebrate-*Drosophila* and definitely before that of primate-rodent (Gu 1998).

Results

Dating Gene Duplications in Early Vertebrates

Figure 2 shows the original and linearized phylogenetic tree for aldolase (ALDO) gene family, which was inferred by the neighbor-joining method with Poisson distance. Bootstrap values more than 50% were presented. 1st and 2nd represent the time points (T_1 and T_2, respectively) of the first and the second gene duplications, respectively. The linearized neighbor-joining tree was used to compute the average distance (*d*). We found that the linearity between *d* and the calibration time (*t*) is very good for both calibration systems I and II; the correlation coefficients for two calibration systems are 0.95 and 0.94, respectively. Thus, as shown in Fig. 2C and Fig. 2D, we can estimate times of the first and second gene duplications.

The estimated times of first and second gene duplications in 26 gene families are presented in Table 1. At first we discuss the result based on the first calibration system (i.e., the split of vertebrate-*Drosophila* is about 830 mya). For most of cases (21 of 26), T_1 is larger than 500 mya, and for all except one case, T_1 is less than 700 mya. The mean of T_1 over all families is 594 mya, with standard error 17 mya. Similarly, the mean of T_2 over all families is 488 mya, with standard error 15 mya. It is noteworthy that the biological interpretation of these means of T_1 and T_2 is meaningful only under the 2R hypothesis, i.e., all these gene families were generated by

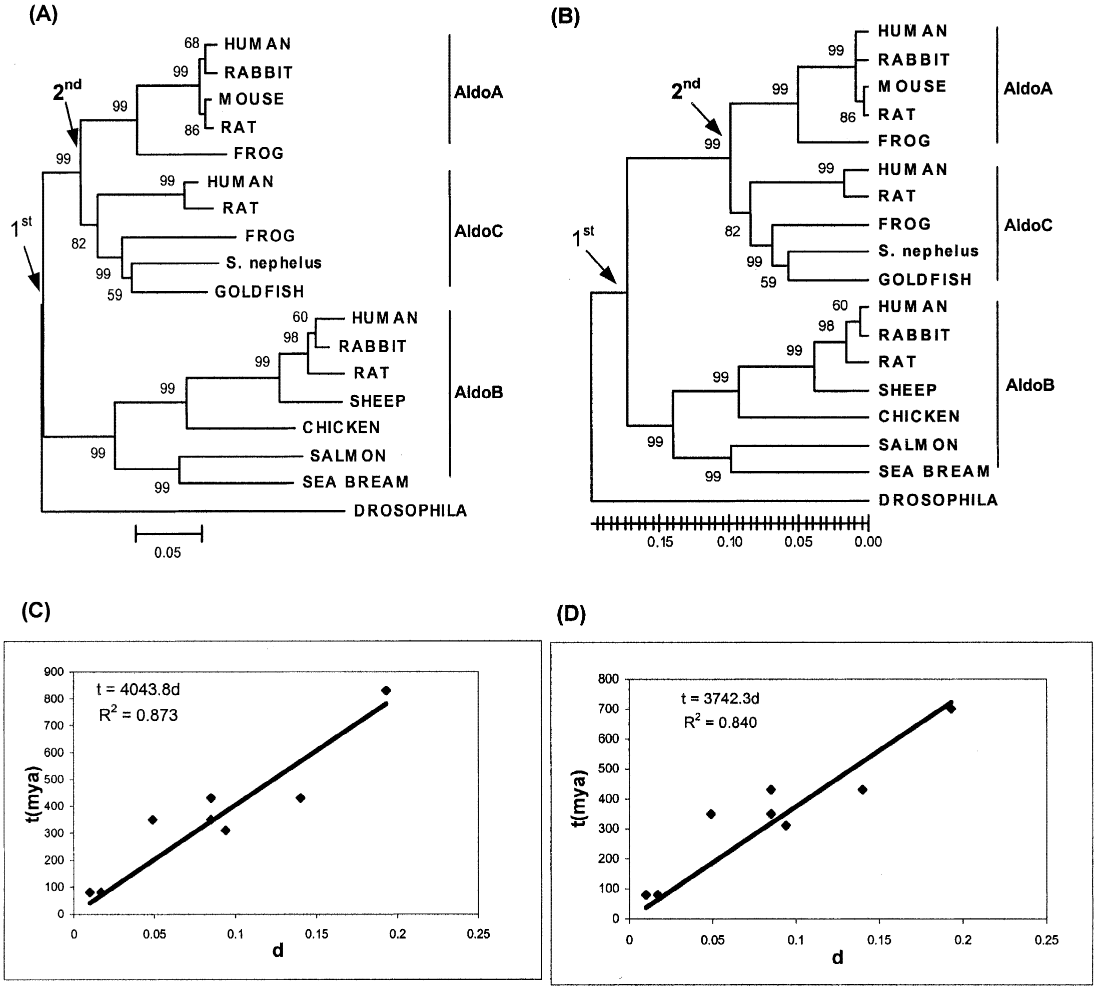

Fig. 2. Phylogenetic analysis of aldolase gene family. **A** The original phylogenetic tree. Neighbor-joining method with Poisson distance was used to infer the tree topology. Bootstrapping values more than 50% were presented. 1st and 2nd represent the time points of the first and the second gene duplications, respectively. **B** The linearized neighbor-joining tree used to compute the average distance. Bootstrapping values more than 50% were presented. **C** The linear relationship of evolutionary distance (*d*) against time (*t*) (correlation coefficient r = 0.95) based on Calibration System I. **D** The linear relationship of evolutionary distance (*d*) against time (*t*) (correlation coefficient r = 0.94) based on Calibration System II.

two-round genome duplications in the early stage of vertebrates. At least for some of them, this assumption may not hold. This is because we cannot rule out the possibility that due to the incompleteness of data or losses of member genes, a gene duplication event can actually be the result of a third genome duplication (3R model) or simply is irrelevant to any genome duplication. At any rate, this dating is largely consistent to the most widely cited version of the 2R hypothesis. (1) The first round of genome duplication is assumed to have occurred prior to the divergence of *Agnatha,* while the mean of T_1 is 594 mya. And (2) the second round of genome duplication is assumed to have occurred before the divergence of *Chondryichthyes*, while the mean of T_2 is 488 mya.

To explore the pattern of functional divergence after successive gene (genome) duplications, it is important to know the separation in geological time between them, i.e., $\delta = T_1 - T_2$ (Nadeau and Sankoff 1997). Two successive gene duplications may act as a single event in the time scale of long-term evolution if $\delta < 30$ million years (myr), which can be explained as follows. From the population genetics viewpoint, if two gene duplication events occurred in a very close time period, these (as many as four) duplicate genes are polymorphic that have not been fixed (or lost) in the population. As a result, the long-term effect may not be distinguished from one single event (Li 1983). As extensively discussed by Nadeau and Sankoff (1997), δ may provide biological insights for understanding the relative rate of functional divergence and loss after gene (genome) duplications. As shown in Table 1, the distribution of the separation of geological time (δ) between T_1 and T_2 (the first calibration system) can be summarized as follows: 4 cases with $\delta < 50$ myr, 9 cases for $50 < \delta < 100$ myr, and 13 cases with $\delta \geq 100$ myr. The mean of δ is 106 myr with standard error 12 myr. That is, in most cases δ is significantly longer than the average survival time of a duplicate gene in the genome without functional divergence

Table 1. Time estimation of the first and the second gene duplications of 26 gene families with 3-gene clusters

		Calibration System I			Calibration System II		
Gene family	Aligned sites	T_1 (mya)	T_2 (mya)	$T_1 - T_2$ (myr)	T_1 (mya)	T_2 (mya)	$T_1 - T_2$ (myr)
EGR	308	724	624	100	647	558	89
Aldo	360	699	404	295	629	364	265
Myb	506	691	528	163	626	478	148
Nos	1,100	690	563	127	602	491	111
ZO	821	685	458	227	585	491	94
Carboxypeptidase	390	682	598	84	627	549	78
APLP	501	657	473	184	572	411	161
ROR	431	643	444	199	544	376	168
ITPR	2,561	632	614	18	543	527	16
Rel	380	629	461	168	575	422	153
ITGA5/8/VNTR	981	628	619	9	562	553	9
Insulin receptor	1,199	623	589	34	538	509	29
Hedgehog	277	618	552	66	580	517	63
Caudal	145	616	495	121	566	455	111
Jun	239	598	471	127	528	417	111
Pyg	836	595	452	143	505	384	121
ADRA1	383	592	480	112	519	421	98
MINOR	276	579	496	83	500	428	72
SHR	473	568	515	53	498	452	46
Cyclin D	279	541	468	73	480	415	65
GLI	392	520	413	107	466	371	95
FGF 8/17/18	189	490	423	67	435	375	60
Fos	242	472	410	62	415	361	54
Calponin	174	440	377	63	382	327	55
ADRA2	287	424	400	24	369	348	21
ADRB	310	419	362	57	365	316	49
Average ± SE		594 ± 17	488 ± 15	106 ± 12	525 ± 15	435 ± 13	90 ± 11

Note: All estimates are based on the calibration time of primate-rodent (80 mya), mammal-bird (310 mya), mammal-amphibian (350 mya), tetrapod-teleost (430 mya), plus different calibration times for the vertebrate-*Drosophila* split. Calibration system I is based on the molecular dating of vertebrate-*Drosophila* split (830 mya) (Gu 1998), whereas calibration system II assumes the divergence time of vertebrate-*Drosophila* as 700 mya. Note that the controversy about the time of mammal-mammal split (65–115 mya) (Kumar and Hedges 1998) has almost no effect on our estimation (mya: million years ago; myr: million years)

(Li 1983). It implies that the fate of most duplicate genes from the first gene (genome) duplication might have been determined (either acquiring some new functions or becoming a pseudogene) when the second gene (genome) duplication took place. This result holds regardless of the debate whether these gene families were generated by two- or three-round genome duplications, but has an indication that these gene (genome) duplications may have distinct impacts on the evolution of vertebrate genomes.

As expected, our estimates based on the second calibration system, on average, are lower than those based on the first calibration system, though they are similar (Table 1). The mean of T_1 is 525 mya and that of T_2 is 435 mya. Since phylogenetic analysis of each gene family clearly show that these gene duplications had occurred before the divergence between tetrapods and teleosts (i.e., >430 mya, see Materials and Methods, results not shown), it seems unreasonable if the estimate of T_2 is less than 400 mya. This can be caused by the nonlinear property of the calibration system, which is apparently nontrivial in the second calibration system; indeed there are 9 such cases (Table 1). We find that the split time between vertebrate and *Drosophila* ($t_5 = 700$ mya) is not consistent with the vertebrate calibrations because it is not the case for the first calibration system. Therefore, it seems that our results based on the first calibration system are more reasonable. Interestingly, the distribution of the separation of geological time between T_1 and T_2 is fairly invariant, with a mean of 90 myr.

Instead of a Poisson distance, we have used various distances, including PAM and gamma distance, to examine the robustness of our results; they usually give similar results with ~5% margin (data not shown). To examine the extent to which our dating is affected by the rate variation among lineages, we have calculated the relative time ratio of each gene duplication event to a single speciation event that has been used for calibration. For instance, the means of relative dates of the first gene duplication are 0.67, 1.78, and 2.71, respectively, to the divergence time of vertebrate-*Drosophila*, tetrapod-teleost, and mammal-bird; those of the second gene duplication are 0.48, 1.31, and 1.93, respectively. Therefore, if we only use vertebrate calibrations, the dates of these genome duplications can be inflated up to $T_1 \approx 800$ mya and $T_2 \approx 600$ mya, about 25% higher than our

Table 2. Phylogenetic patterns of four-gene cluster families that have been used in this study

Tree topology		Gene family	Aligned sites	Bootstrap value(s) (%)
Type I	(((A,B),(C,D)), invertebrate)	Activin β	293	85[a]; 99[b]
		EGFR	634	97; 99
		ID	97	96; 57
		Jak	663	99; 98
		PFKFB	388	97; 83
Type II	((((A,B),C),D), invertebrate)	CNCG	537	99[c]; 69[d]
		DLG	679	35; 99
		ELAV	243	98; 95
		GLR	328	99; 99
		MKP	254	96; 76
		PDE4	435	99; 90
		PO	687	99; 99
		Ras	179	76; 76
		Src-related	452	99; 99
		Syndecan	156	88; 92
		Trk	711	99; 99
Type III	((((A,B),C), invertebrate),D)	Cathepsin	312	98[e]; 99[f]
		Ezrin	552	99; 99
		Mef2	291	99; 99
		Notch	1,052	96; 99
		NPYR	249	92; 99
	(((A,B), invertebrate),(C,D))	MyoD	185	99[g]; 99[h]
		OCT	219	99; 99

Note: See Fig. 1 for the topologies of three types of phylogenetic pattern. Invertebrate (*Drosophila melanogaster* or/and *C. elegans*) were used as an outgroup

[a,b] Bootstrap values of branching points D2 in Fig. 1B for Type I gene family

[c,d] Bootstrap values of branching points D2 and D3 in Fig. 1C for Type II gene family, respectively. Only for one case (DLG gene family), the bootstrap value = 35%, which should be considered ambiguous

[e–h] Bootstrap values of branching points Ds in Fig. 1D for Type III gene family

estimates even if the upper bound (t_5 = 830 mya) is used. We believe that it may be caused by a fast evolution after gene duplication, which may be as a result of functional redundancy (Li 1983). If that is the case, $T_1 - T_2$ is also inflated. This effect has been significantly alleviated if *Drosophila* is used as an outgroup.

An important but difficult problem is how to deal with the gene clusters with different evolutionary rates. We have indeed observed that in several cases the evolutionary rates among clusters vary so much that they could potentially affect the accuracy of time estimation. Fortunately, for most cases under study, the evolutionary rates among clusters do not vary significantly so that we could use the average rate as a compromise. We may take two approaches. The first one is our current approach, i.e., first compute the average evolutionary rate and then compute the duplication time. The other approach is to compute the duplication time from each member gene, and then take an average. Though these two approaches are feasible in practice, we found that statistically the first one is more stable (data not shown).

Complicated Patterns of Four Gene Clusters

Phylogenetic analysis for 23 gene families with 4 member genes shows complicated patterns, as denoted by type I, II, and III, respectively (Fig. 1). As summarized in Table 2, only five gene families show type I phylogenetic pattern that is predicted by the 2R model. For example, the phylogenetic trees of activin β and PFKF gene families (Fig. 3) show that four isoforms were likely to be generated after the divergence between vertebrates and *Drosophila*. However, 11 gene families show type II phylogenetic pattern (Fig. 1C). For instance, see Fig. 4 for ELAV (Hu) and RAS gene families. The branching points of duplication events are highly supported by the bootstrap test, indicating the statistical reliability of this tree topology (Table 2). Thus, the 2R model is not sufficient, and an additional gene (or genome) duplication event is needed. Moreover, the remaining seven gene families indicate that at least one gene duplication event is prior to the divergence between vertebrates and *Drosophila*. As an example, see Fig. 5 for the Mef2 and MyoD gene families. Therefore, these four gene clusters should not be considered as the product of two-round genome duplication (Hughes 1999). Indeed, the Mef2 gene family has been used as a “good” example according to the 4:1 rule (Spring 1997), which turns out to be incorrect. However, denying the generation of isoform Mef2B as the result of any possible genome duplication does not necessarily rule out the possibility that the origin of other three iso-

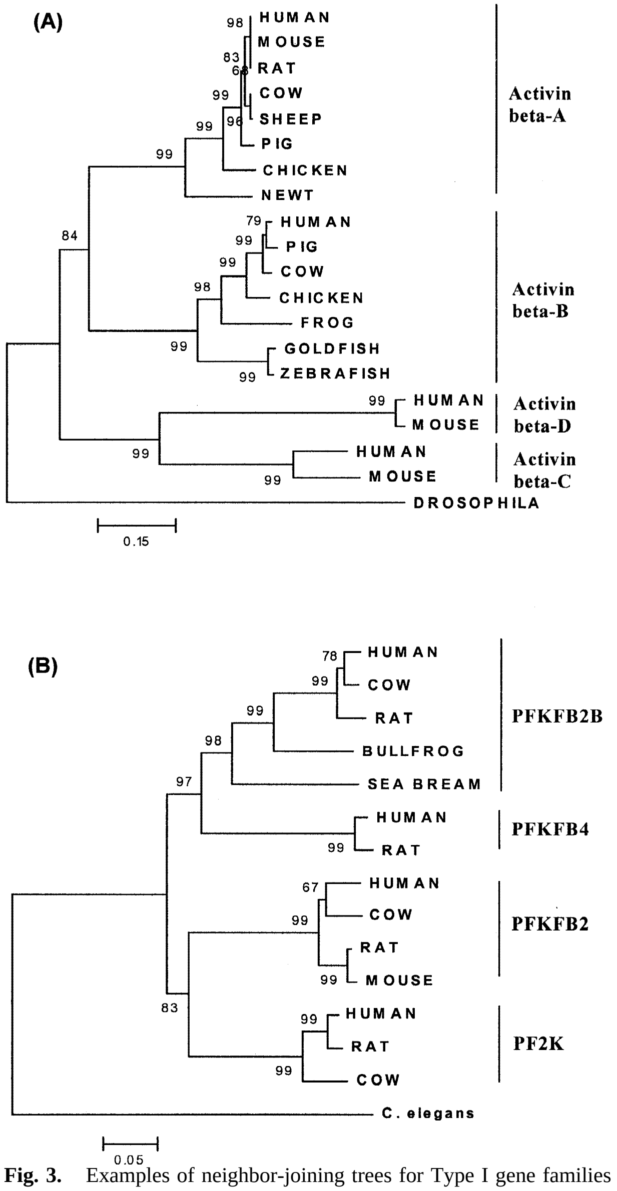

Fig. 3. Examples of neighbor-joining trees for Type I gene families with four member genes. **A** Activin β gene family. **B** PFKF gene family.

forms (Mef2A, D, and C) are related to genome duplication(s).

Discussion

Is Genome Duplication Hypothesis Wrong?

Since the hypothesis of genome duplication, in particular the 2R model, has been widely cited in the field of vertebrate development and evolution, it deserves a rigorous test. Our exhaustive search over more than 1,100 gene families have shown that 112 gene families with 3 members, 58 gene families with 4 members, and 130 gene families with 2 member genes; multiple families that are apparently generated recently, such as Ig, TCR, or olfactory receptors, have been excluded. Sequence diversity among member genes is usually more than four times

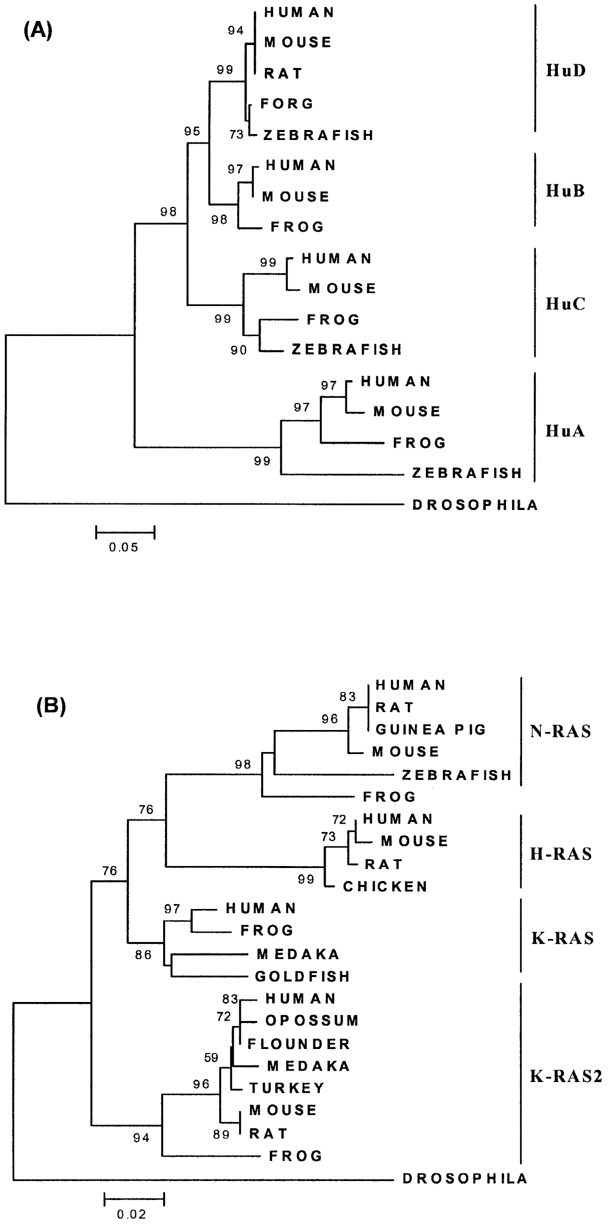

Fig. 4. Examples of neighbor-joining trees for Type II gene families with four member genes. **A** Elav (Hu) gene family. **B** Ras gene family.

larger than that between orthologous human/mouse genes, indicating that they have maintained in the genome since the early stage of vertebrates. The appearance of such large numbers of multiple-cluster families (~27% in the vertebrate genome) implies the possibility of genome-wide duplication(s).

Furthermore we have analyzed 49 vertebrate gene families with at least 3 member genes to investigate the evolutionary pattern of gene (genome) duplications in the early stage of vertebrates or shortly before the origin of vertebrates. It has been clear that the two-round genome duplication theory (2R) can well explain the mean duplication times over 26 gene families with 3 gene clusters and the phylogenetic pattern (type I) of 5 gene families with 4 gene clusters. However, to explain type II

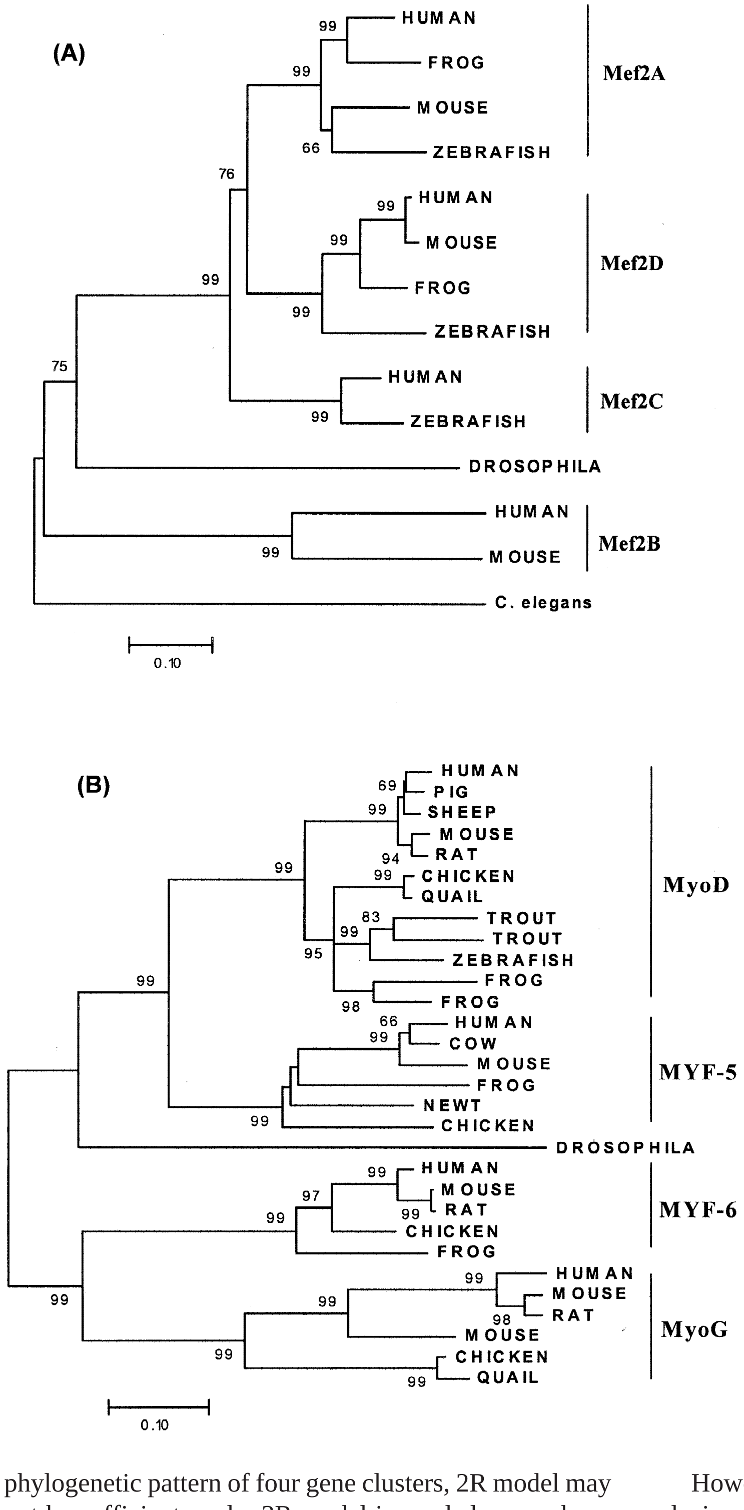

Fig. 5. Examples of neighbor-joining trees for Type III gene families with four member genes. **A** Mef-2 gene family. **B** Ras gene family.

phylogenetic pattern of four gene clusters, 2R model may not be sufficient, and a 3R model is needed, or we have to assume one of gene duplications is irrelevant to any genome-wide duplication. At this moment, we cannot rule out any possibility. Seven cases in type III phylogenetic pattern shows that it can be very misleading to assign any four (or more) tissue-specific genes in the human genome as "potential" products of the 2R genome duplication, without any detailed phylogenetic analysis. For example, we have already found several such mistakes in the Tetrabase developed by Spring (1997).

However, we do not agree Hughes's (1999) final conclusion that the role of genome duplication in vertebrate evolution is entirely speculative. What we have shown in this paper, as well as many authors (Skrabanek and Wolfe 1998; Hughes 1999; Smith et al. 1999), is that the generation of gene families in the early vertebrate is much more complicated than we previously thought. In principle, there is a simple correlation between genome duplication(s), chromosome synteny, phylogeny, and functional divergence (e.g., tissue specificity). It has been shown that current vertebrate genomes contain substantial such information (e.g., Lundin 1993; Holland et al. 1994; Kasahara et al. 1996; Endo et al. 1997; Nadeau and Sankoff 1997; Ruddle 1997; Pebusques et al. 1998; Suga et al. 1999). Although this type of information is invaluable for understanding the origin of human genome complexity, the difficulty for a direct testing by using these data is that in the past several hundred million years, the vertebrate genome has experienced dramatically chromosome rearrangement, functional divergence, and gene loss. Without a stochastic model that takes these mechanisms into account, it is not surprising that our attempt to prove genome duplication(s) is always unsuccessful (Skrabanek and Wolfe 1998). Nevertheless, the 2R model clearly provides a most parsimonious explanation for most observations; an additional gene duplication for some gene families under study is necessary, but it remains unclear whether its scale is large enough to be considered as a third genome duplication (the 3R) model. If the complete sequence of human genome finally shows a significant portion of duplicate genes is indeed generated during a short time period of the early vertebrate lineage, it is difficult to imagine how can we offer an explanation without the concept of genome duplication. At any rate, adding an additional (or even more) genome duplication to the conventional 2R model is fundamentally different from denying the concept of genome duplication. Currently we are working on these challenging but exciting problems.

Acknowledgments. This study was supported by an ISU startup fund and an NSF grant to X.G. We thank Jianying Gu for assistance. We are grateful to an anonymous referee for helpful comments.

References

Endo T, Imanishi T, Gojobori T, Inoko H (1997) Evolutionary significance of intra-genome duplications on human chromosomes. Gene 205:19–27

Fryxell KJ (1996) The coevolution of gene family trees. Trends Genet 12:364–369

Gu X (1998) Early Metazoan divergence was about 830 million years ago. J Mol Evol 47:369–371

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

Hughes AL (1999) 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

Kasahara M, Hayashi M, Tanaka K, Inoko H, Sugaya K, Ikemura T, Ishibashi T (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

Kumar S, Hedges SB (1998) A molecular timescale for vertebrate evolution. Nature 392:917–920

Li W-H (1983) Evolution of duplicate genes and pseudogenes. In: Nei M, Keohn RK (eds) Evolution of genes and proteins. Sinauer Associates, Sunderland, MA, pp 14–37

Lundin LG (1993) Evolution of the vertebrate genomes as reflected in paralogous chromosomal regions in man and the house mouse. Genomics 16:1–19

Nadeau JH, Sankoff D (1997) Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. Genetics 147:1259–1266

Ohno S (1970) Evolution by gene duplication. Springer-Verlag, Berlin

Pebusques M-J, Courliur F, Birnbaum D, Pontarotti P (1998) Ancient large scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome duplications. Mol Biol Evol 15: 1145–1159

Ruddle FH (1997) Vertebrate genome evolution—the decade ahead. Genomics 46:171–173

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

Shu D-G, Luo HL, Conway Morris S, Zhang X-L, Hu S-X, Chen L, Han J, Zhu M, Li Y, Chen Z (1999) Lower Cambrian vertebrates from South China. Nature 402:42–46

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

Skrabanek L, Wolfe KH (1998) Eukaryote genome duplication—where's the evidence? Curr Opin Genet Dev 8:694–700

Smith NG, Knight R, Hurst LD (1999) Vertebrate genome evolution: a slow shuffle or a big bang? Bioessays 21:697–703

Spring J (1997) Vertebrate evolution by interspecific hybridization—are we polyploid? FEBS Lett 400:2–8

Suga H, Hoshiyama D, Kuraku S, Katoh K, Kubokawa K, Miyata T (1999) Protein tyrosine kinase cDNAs from amphioxus, hagfish, and lamprey: isoform duplications around the divergence of cyclostomes and gnathostomes. J Mol Evol 49:601–608

Takezaki N, Rzhetsky A, Nei M (1995) Phylogenetic tests of the molecular clock and linearized trees. Mol Biol Evol 12:823–833

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680