

# Increasing Genomic Complexity by Gene Duplication and the Origin of Vertebrates

Andrew P. Martin\*

Department of Environmental, Population and Organismic  
Biology, CB 334, University of Colorado, Boulder, Colorado  
80309-0334

Submitted October 2, 1998; Accepted March 10, 1999

---

**ABSTRACT:** Prevailing hypotheses concerning the origin of the vertebrate genome postulate successive genome duplications before the origin of fishes or tetrapods. These hypotheses predict episodic expansion of gene families early in vertebrate evolution (mostly before the origin of fishes), tetralogous relationships between gene copies samples from invertebrates and vertebrates, and gene family trees with symmetrical shapes. None of these predictions were evident from a phylogenetic analysis of 35 gene families. Overall, the results do not refute the hypothesis that gene family evolution is governed by independent gene duplications occurring with identical probability across gene lineages. These results suggest that the genome complexity of contemporary vertebrates mostly reflect small-scale (regional) DNA duplications instead of large-scale (genomic) duplications.

**Keywords:** gene duplication, vertebrates, gene families, genome duplication.

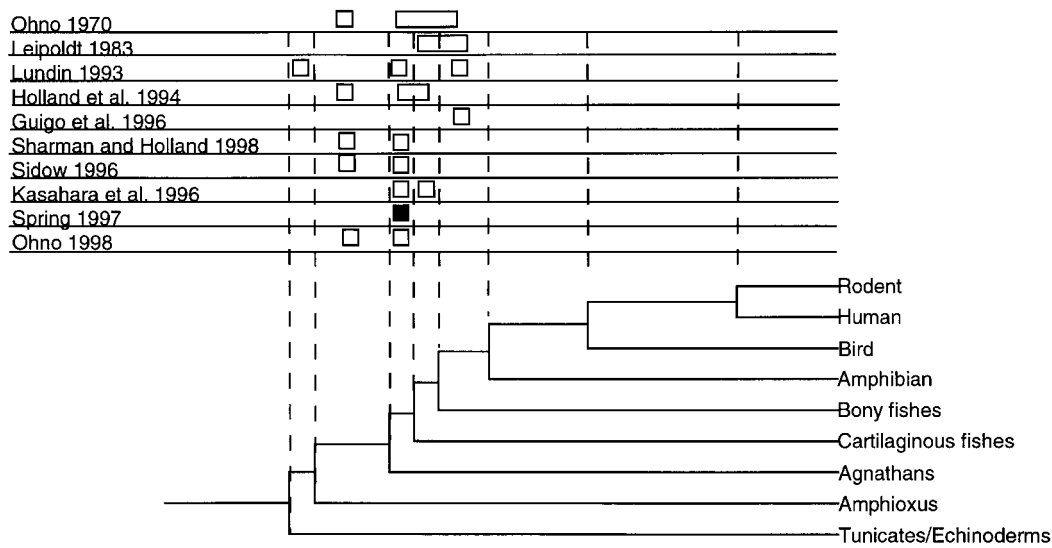
---

The connection between genetic and organismal complexity is of general interest to biologists (Bonner 1988; McShea 1996). One arena that has been the subject of considerable interest is the genome changes underlying the origin of vertebrates. There is a clear indication for genomic elaboration before the origin of fishes. For instance, there are approximately 15,000 genes in nematodes, *Drosophila*, and sea squirts, whereas there are thought to be 70,000 in vertebrates (Miklos and Rubin 1996; Simmen et al. 1998). The correlation of genomic and organismal diversification at the base of vertebrates lends force to the argument that increase in organismal complexity (i.e., the elaboration of morphological, physiological, biochemical,

cellular, and behavioral phenotypes) stems from a corresponding diversification of the underlying genome (Ohno et al. 1967; Ohno 1970; Gordon 1994; Iwabe et al. 1996).

A variety of published hypotheses explain the inferred increased genetic complexity accompanying the origin of vertebrates, and most invoke whole-genome duplications (fig. 1). The most widely accepted hypothesis is that the apparent octoploidy of the vertebrate genome resulted from two rounds of genome duplication before the origin of fishes (Holland et al. 1994; Sidow 1996; Nadeau and Sankoff 1997; Spring 1997; Postlethwait et al. 1998; fig. 1). Three important observations have been marshaled in favor of this hypothesis. First, large sections of chromosomes have similar gene order between fishes and mammals (Brenner et al. 1993; Ruvinsky and Silver 1997; Postlethwait et al. 1998); moreover, there is evidence of paralogy of chromosomes in mammalian genomes (Lundin 1993). Second, there is one homeobox gene cluster in arthropods, echinoderms, and cephalochordates and four clusters in vertebrates (Holland et al. 1996), although teleost fishes are exceptional with the recent discovery of seven Hox gene clusters and evidence for extensive gene loss (Aparicio et al. 1997; Amores et al. 1998). In addition, there are three gene clusters in lampreys (Sharman and Holland 1998). Finally, the number of coding genes in vertebrates is roughly four times the number estimated for invertebrates (Brenner et al. 1993; Fields et al. 1994; Miklos and Rubin 1996). Moreover, for many genes in *Drosophila* (or nematodes and sea urchins), there are typically three or four putative vertebrate orthologs (Holloway et al. 1997; Semenov and Snyder 1997; Spring 1997; Yamaguchi and Brenner 1997). This ratio has been called tetralogy, a term marking dogmatic acceptance of the repeated genome duplication hypothesis (Nadeau and Sankoff 1997; Spring 1997). This paradigm is so entrenched that gene trees have been rearranged post hoc into a branching pattern consistent with repeated genome duplication hypotheses; see, for example, figure 7 in Shain et al. (1997, p. 351). Further, Miklos and Rubin (1996, p. 522) reason, "If the genome projects verify the underlying octoploid nature of the human and mouse genomes, then

\* E-mail: am@stripe.colorado.edu.



**Figure 1:** Alternative hypotheses proposed to explain the elaboration of the genome in vertebrates. The figure is a modified version from Skrabanek and Wolfe (1998). The phylogenetic tree is drawn with branch length proportional to time. Open squares indicate whole-genome duplication events. Solid squares denote two successive genome duplications. The Ohno (1998) citation is from Skrabanek and Wolfe's figure.

the basic vertebrate gene number may be similar to that of the fly and worm, about 12,000 to 14,000 genes. The duplicated pathways in mammals are, however, likely to have adopted specialized expression patterns and biological functions." The debate is, therefore, not about whether genome duplications occurred but about when and how many genome duplication events there were, and whether these events were important in the evolution of increased organismal complexity.

Recognition that regional, independent duplications are common (Brown et al. 1990; Currie and Sullivan 1994; Irwin 1995; Tomarev et al. 1995; Brown et al. 1996; Jobling et al. 1996; Rowen et al. 1996; Saxena et al. 1996; Gilbert et al. 1997; SanMiguel et al. 1998) and that they may also involve large regions of chromosomes (Kenck et al. 1997; Potier et al. 1998; Ritchie et al. 1998) suggests an alternative hypothesis for the elaboration of genomic complexity—namely, that the contemporary vertebrate genome organization is mostly the result of piecemeal assembly from fixation of independent gene duplications. Duplication events by unequal crossing over, replicative transposition, and replicative translocation are known to occur at relatively high rates. For example, rates of unequal crossing over are high in most organisms,  $10^{-3}$ – $10^{-5}$  per locus per generation (Fryxell 1996 and references therein), and vary tremendously across loci. Ribosomal genes arise and are deleted at the rate of  $10^{-2}$ – $10^{-3}$  per loci per generation, whereas rates for protein-coding genes range from  $10^{-4}$  to  $10^{-7}$  per generation (Shapira and Finnerty 1986;

Fryxell 1996). Similarly, the power of replicative transposition for gene propagation is evident from the GADPH gene family in rodents (Garcia-Meunier et al. 1993) and the estimated doubling of the maize genome, from 1,200 Mb to 2,400 Mb, over only the last few million years (SanMiguel et al. 1998). Demonstration that a 10-kb portion of a trypsinogen homology unit, complete with the whole gene and genetic elements for tissue-specific and developmentally specific expression, was replicated and translocated from chromosome 7 to chromosome 9 in humans (Rowen et al. 1996) suggests that pathways of illegitimate recombination may have been important for genome evolution. Rowen et al. (1996) speculate that this type of translocation may be a mechanism for the origin of multigene families.

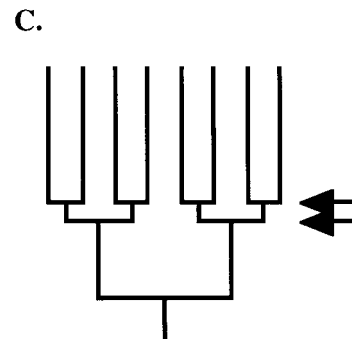
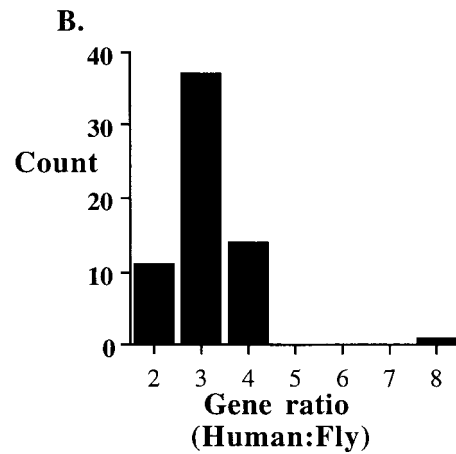
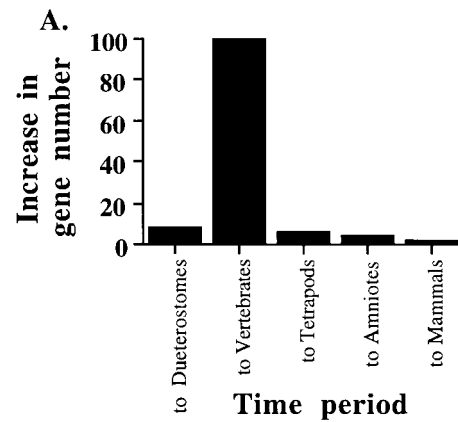
Of course, as Fryxell (1996, p. 364) notes, "spontaneous gene duplications occur much too frequently to be the rate limiting step in the evolutionary process of gene duplication and divergence." Models of the fate of duplicate genes in populations indicate that genes can increase in frequency if they are advantageous (Clark 1994; Walsh 1995) or when there is recurrent duplication of an essential gene that is facing inactivation by deleterious mutation (Clark 1994). In addition, duplications that are neutral with respect to fitness will also occasionally increase to fixation by drift, even if the duplicate loci accumulate deleterious mutations (Clark 1994; Walsh 1995). Conversely, either genes are lost through the elimination of the segregating mutation in populations by drift or selection or

they can be lost after fixation by the steady accumulation of mutations that render the gene inactive (a phenomenon known as gene silencing). The important point is that gene family expansion is a product of mutation and fixation. Given that the supply of duplication mutations may not be limited over the time period during which vertebrates originated, apparent episodic increase in genome complexity may reflect the effects of processes governing the fixation more than signal an explosion of new genes.

Alternative hypotheses of gene family evolution can be tested in a phylogenetic framework. Hypotheses invoking multiple, ancient genome duplications predict episodic increase in gene duplication early in the evolution of vertebrates, many single-copy genes in *Drosophila* (and other invertebrates) will have multiple vertebrate orthologs (tetralogs) and symmetry of gene family trees (fig. 2). Proponents of hypotheses invoking multiple genome duplications argue that there is abundant evidence for the retention of genes arising from ancient genome duplications. Nevertheless, a number of difficulties are associated with testing alternative hypotheses of the number and timing of polyploidization events based on examination of the differences among paralogous genes. Individual gene duplications and loss and gene conversion will erase the signal of ancient genome duplication (Sidow 1996; Wolfe and Shields 1997; Skrabanek and Wolfe 1998). Long periods of tetrasomic inheritance may prohibit the differentiation of duplicated loci (Allendorf and Danzmann 1997; Gaut and Doebley 1997) such that the timing of genome duplication appears to have happened more recently. Conversely, the formation of an allotetraploid through hybridization of divergent genomes (Gaut and Doebley 1997; Spring 1997) will push the inferred duplication event back in time. In addition, chromosomal translocations will erode syntenic relationships between paralogous chromosomes. In general, these processes of genome evolution will tend to blur the distinction among the different hypotheses (fig. 1), reducing the power of testing alternatives. Thus, the goal of this study is not to test the alternative hypotheses outlined in figure 1 but to test whether there is evidence for multiple genome duplications early in the evolutionary history of vertebrates. The alternative hypothesis is that gene family complexity has increased over time by regional duplications occurring independently and with identical probability across gene lineages.

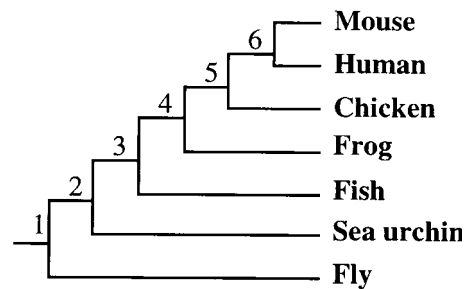
### Methods

Gene families selected for this analysis satisfied the criterion of having a reasonably even and broad taxonomic representation of sequences with some evidence of past duplication. All vertebrates for which there were sequence



**Figure 2:** Graphical representation of the predictions of genome duplication hypotheses. *A*, peak in the relative numbers of gene duplication events in the period of time preceding the origin of fishes. *B*, peak in the distribution of the ratio of orthologous genes in vertebrates and invertebrates. The distribution shown is from Spring (1997). *C*, Symmetry in the shape of gene trees. Arrows mark putative genome duplications.

data were included, but the analysis focuses on inferring gene duplication events for the internodes labeled 1–6 in figure 3. Duplications occurring on more recent branches (e.g., within mammals) were ignored. In addition, a diverse sample of proteins was surveyed (i.e., enzymes, transcription factors, secreted proteins, structural proteins, etc.; table 1). Alignments of proteins for gene families were obtained from the Pfam database that constructs protein sequence alignments of gene families using hidden Markov models (HMM; Eddy 1996; Sonnhammer et al. 1997). These HMM alignments were used because they have been shown to recover character homology of amino acid residues among highly divergent protein sequences efficiently (McClure et al. 1994; Sonnhammer et al. 1997). Each gene family alignment was edited to remove sequences having lengths that were approximately 50%–60% less than the length of the overall alignment. The resultant matrix of aligned protein residues was subjected to maximum likelihood (ML) and maximum parsimony (MP) tree-building algorithms or transformed into a distance matrix for cluster analysis with the neighbor-joining (NJ) and least-squares (FITCH) algorithms. Four different tree-building algorithms were employed because each method makes different assumptions and performs differently depending on the particular pattern of character covariation across taxa (Huelsenbeck and Hillis 1993). Because different tree-building algorithms can yield very different trees depending on the pattern of character covariation across taxa, we avoid a consistent bias by not relying on a single method. Maximum-likelihood trees were calculated using PUZZLE 3.1 (Strimmer and von Haeseler 1996), which implements a tree search algorithm based on quartet puzzling (Strimmer et al. 1997). The analysis incorporated gamma-distributed rate heterogeneity among sites with five rate classes, one invariant and four variable. A hidden Markov model was used for determining the distribution of the different rates and their corresponding sequence positions (Felsenstein and Churchill 1996). Maximum-likelihood trees are computationally expensive, a cost that is further exacerbated by the estimation of the parameters for an HMM and incorporation of rate heterogeneity into the model of evolution. Some computational cost is saved through the use of an inexact search method (Quartet Puzzling), but this method still searches a huge proportion of the tree space. Construction of ML trees was made possible through use of a Silicon Graphics Origin 2000 parallel processing supercomputer, estimated to perform 1.23 GFLOPS (billion floating point operations per second; Ramey 1997). MP trees were generated with PAUP \* (beta release 4.50 DOS and 4.54 SGI) by permission from D. Swofford (1993), Smithsonian Institution. A heuristic search algorithm employed tree bisection reconnection branch swapping with random sequence addition repli-



**Figure 3:** Organismal phylogeny used to investigate rate and timing of gene duplications. Numbers above nodes are time periods for which the number and timing of duplication events were inferred.

cated 10 times. If <75 trees were found, all were retained. Majority rule, Adams, and strict consensus trees were generated and used in subsequent analyses for searches returning >75 equally parsimonious shortest-length trees. Searches not having begun to converge during the first replicate within 12 h were moved to a Silicon Graphics Origin 2000 supercomputer. Searches that failed to finish in reasonable time or found more than 50,000 shortest-length trees were taken as grounds for exclusion of the data set from this analysis. Distance matrices were constructed during the maximum-likelihood calculations and with the PROTDIST algorithm of PHYLIP (Felsenstein 1993, version 3.37), using the PAM Dayhoff substitution-weighting matrix. The NJ algorithm used the PROTDIST matrix, whereas we used the FITCH algorithm (Felsenstein 1993) on maximum-likelihood distances. All best (or minimum-length) trees were retained as hypotheses of gene family phylogeny. Nine additional trees were drawn from the literature (see table 1). All of these gene trees were generated with methods consistent with the assumptions of at least one of the methods described in this article.

Gene trees inferred from alignments include information about relationships among species (speciation) and about relationships among genes (duplication). In this sense, each species can be considered as a sample of the evolutionary history of the multigene family and will contain genes that were born before the origin of the species and may contain some members of the gene family that are unique to that lineage. A problem with inferring the evolutionary history of multigene families is that most gene trees are complex, and it is difficult to decipher when gene duplications occurred relative to speciation events and whether genes are missing (either because of a failure to sample or because of the loss of genes). Therefore, reconciled trees were constructed using COMPONENT (Page 1993a, 1993b) for all gene trees using the species

**Table 1:** Gene families used to test hypotheses about the evolution of complexity

	Gene family	Function
Transcription factors (turn on and off genes):		
1	dlx <sup>a</sup>	Development
2	PAX	Development
3	Forkhead	Development
4	POU	Generalist
5	MyoD <sup>b</sup>	Muscle development
6	MADS <sup>c</sup>	Development
7	ETS	Transcription regulation of diverse function
Growth factors (stimulate growth and differentiation):		
8	GATF	Transforming factor
9	TGF	Generalist
10	FGF	Fibroblast growth factor
11	NGF	Growth and differentiation of neurons
Receptors (bind specific ligands):		
12	Nicotinic receptors <sup>d</sup>	Binds ligands
13	Steroid receptors <sup>e</sup>	Binds ligands
Enzymes (perform work):		
14	Creatine kinase	Phosphotransferase
15	Trypsinogen	Protease
16	Cysteine protease <sup>f</sup>	Protease
17	Acetylcholinesterase <sup>g</sup>	Esterase
18	Ubicon	Covalent attachment of ubiquitin to proteins
Structural proteins:		
19	Dynamin	Microtubule-associated force protein
20	Tubulin	Cellular matrix
21	Dynein <sup>h</sup>	Flagellum
22	Filament	Cellular matrix
23	BAND	Interface between membrane and cytoskeleton
Cellular processes (regulate processes in cells):		
24	ARF	Protein trafficking
25	Ras	Protein trafficking
26	Arrestin	Signal transduction pathways
27	Opsin <sup>i</sup>	Signal transduction
28	Hsp70	Chaperone, protein assembly, stress response
29	Hsp90	Chaperone, protein assembly, stress response
30	G-proteins	Signal transduction
31	ANP	Hormones involved in homeostasis
32	Serpin	Serine protease inhibitors
33	Cyclins	Control of cell division
Secretory (effect processes outside of cells):		
34	Matrixin	Extracellular metalloproteinase
35	WNT	Cell-cell interaction during development

<sup>a</sup> Stock et al. (1996).<sup>b</sup> Atchley et al. (1994).<sup>c</sup> Theissen et al. (1996).<sup>d</sup> Le Novere and Changeux (1995).<sup>e</sup> Detera-Wadleigh and Fanning (1994).<sup>f</sup> Hughes (1994a).<sup>g</sup> Fryxell (1995).<sup>h</sup> Porter et al. (1996).<sup>i</sup> Yokoyama (1995).

tree shown in figure 3. A reconciled tree is the smallest possible tree (measured by numbers of branches) that contains a complete record of the species and gene tree (Page and Charleston 1997). In effect, the gene tree is mapped onto a known (and irrefutable) species tree to reveal the orthology of sampled genes (Page and Charleston 1997). The advantage of COMPONENT analysis is that the relative timing and number of gene duplications can be easily extracted from reconciled trees.

Two additional features of COMPONENT analysis deserve mention. First, reconciled trees determined using COMPONENT require completely bifurcating trees, and the mapping algorithm arbitrarily resolves any polytomies. In this analysis, all polytomies in maximum-parsimony trees were manually resolved so as to maximize the congruence of branching relationships of the gene tree with the species tree; thus, resolution of polytomies was done to minimize the number of duplications required to reconcile the gene and species trees. In all cases, unresolved nodes in maximum-parsimony trees were limited to the relationships among mammals within a single gene clade. For example, in the PAX gene tree, there was one unresolved node for the PAX 8 genes from mouse, rat, dog, and human. The effect of this was to minimize the number of duplications inferred to have occurred along the lineage leading to mammals (defined by the split between mouse and human). Second, gene conversion events appear as duplication events; however, because gene conversion is typically recent, the duplications are assigned to terminal branches and were therefore ignored in this analysis.

Maximum-parsimony analysis typically resulted in more than one minimum-length tree. Gene trees with the least number of gene duplications and losses were retained for analysis. For the distance methods, only one fully bifurcating tree was constructed. Because of the large number of unresolved nodes in a subset of the ML trees, these were used only as a resource for evaluating support for nodes in trees generated by the other methods when there were disagreements between species trees and gene trees.

The shape of phylogenetic trees provides information about the history of diversification and extinction of lineages (Raup et al. 1973; Gould et al. 1977; Mooers and Heard 1997). In our case, tree shape is influenced by the degree that rates of gene duplication vary across gene lineages. The more variable the rate, the greater the imbalance, or asymmetry, of the tree. If gene family diversity is a product of genome duplication events, then gene trees should tend to be more symmetrical than if genes duplicate independently but with identical probabilities (see fig. 2C). Thus, the repeated genome duplication hypothesis can be tested by asking whether observed gene tree shapes tend toward symmetry. Many metrics of tree balance or tree symmetry have been proposed (for review, see Kirkpatrick

and Slatkin 1993; Mooers and Heard 1997), all of which succeed in discriminating between symmetrical and asymmetrical tree topologies yet differ in ways that have yet to be rigorously or even heuristically defined. For the purposes of this analysis, Colless's (1982) imbalance measurement  $I_c$  was used:  $I_c = 2(\sum |TR - TL|) / ([n - 1][n - 2])$ , where  $n$  is the number of tips, the summation is for all internal nodes, and at each node the right and left branches subtend  $TR$  and  $TL$  tips. The variable  $I_c$  was calculated only for the gene tree irrespective of the species tree (in other words, speciation events were not included in the analysis of tree shape), and duplication events inferred to have occurred on terminal branches were also omitted. Values of  $I_c$  expected for a model based on identical and independent probability of gene duplication across lineages (equivalent to the Markov null model developed by Simberloff et al. [1981]) were calculated following Heard (1992). A potential complication is that many of the gene trees were complex before the origin of vertebrates and may bias the results against whole-genome duplications if they are asymmetrical. A quick examination of tree shapes indicates that when there are five or more gene lineages before vertebrates and the tree is completely asymmetrical,  $I_c$  values for gene trees will be greater than the lower 95% confidence interval despite the effects of successive genome duplications.

A second confounding factor is gene loss, as it will generally increase asymmetry and favor rejection of genome duplication hypotheses. The effect of gene loss on tree shape was investigated using gene trees with 16–32 gene lineages. For each set of lineages, two original gene trees were constructed that maximized symmetry and asymmetry, respectively. Gene lineages were randomly pruned such that there were 25% and 50% reductions in gene diversity, and  $I_c$  values were calculated for the resulting trees. Means and standard deviations were calculated from five replicates for each set of gene lineages.

Before an exhaustive analysis was performed of the data for all methods of phylogenetic analysis across all gene families, the results from alternative phylogenetic methods were compared to see whether they yielded significantly different results. Half of the gene families (17) were selected and subjected to the four different methods of tree building, and the results compared (see table A1). For estimates of the number and timing of gene duplications, and for the shape of gene trees, similar results were obtained for all tree-building methods. Linear regression of  $X$  on  $Y$ , where  $X$  and  $Y$  are the numbers of gene duplications inferred for each gene family using NJ, FITCH (F), and parsimony (PARS), was not significantly different from unity across all three comparisons (F vs. PARS, 95% consistency index [CI] = 0.79–1.07; F vs. NJ, 95% CI = 0.80–1.09; NJ vs. PARS, 95% CI = 0.83–1.06). The relative

timing of gene duplications inferred using the different methods was also similar (data not shown). For the majority of gene families, pairwise *t*-tests on arcsine-transformed  $I_c$  values (Sokal and Rohlf 1988) show that all methods (PARS, NJ, and FITCH) yielded similar  $I_c$  values (PARS vs. NJ,  $t = 0.35$ ,  $P = .73$ ; PARS vs. FITCH,  $t = -0.54$ ,  $P = .60$ ; NJ vs. FITCH,  $t = -1.03$ ,  $P = .31$ ), and regression analysis revealed that marked discrepancies between estimates of  $I_c$  using different methods were only evident for two gene families (POU, Forkhead). Furthermore, comparison of residuals from regression revealed a lack of correlation between the magnitude of the difference in  $I_c$  values estimated using different methods and the retention index (RI) of gene trees, a measure of consistency of phylogenetic signal.

This result indicates that less well-supported trees do not show an increased tendency toward asymmetry (Heard and Mooers 1996). Mooers et al. (1995) note, however, that phylogenetic noise can produce more imbalanced trees; thus, it is possible that some of the imbalance (or asymmetry) of the inferred gene trees reflects a methodological bias and therefore favors rejection of repeated genome duplication hypotheses when they may be true. High concordance of trees constructed using different methods coupled with high retention indices (mean RI = 0.82; see table A1) argues that the effect of methodological bias is minimal, however. Because regression analysis revealed that the results were comparable across methods, the gene family tree requiring the least number of gene duplications was adopted as the best estimate of evolutionary history. For all other gene families, either published trees (see table 1) or topologies determined using neighbor-joining distance methods were used.

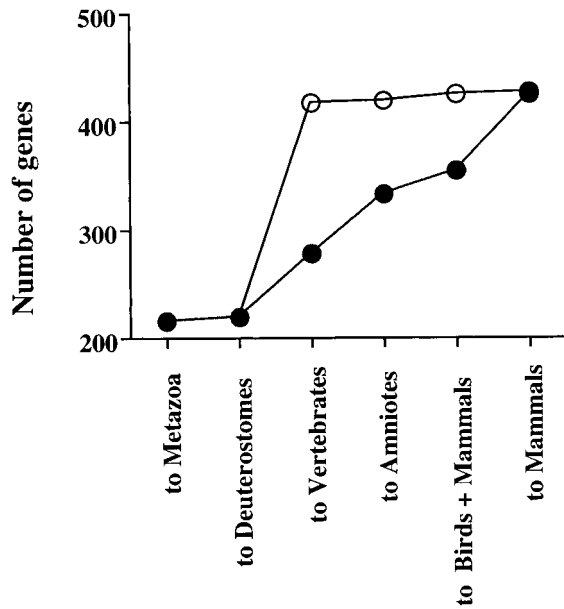
The timing of gene duplications was evaluated under two assumptions. First, genes were assumed to have been sampled evenly across taxa, with the representation of genes available in the database reflecting nature. Obviously this assumption is not true. Nevertheless, when we do the analysis on the current sampling of genes, we make this assumption. Second, genes sampled from higher vertebrates (mammals, birds) were assumed present in lower taxa unless there is evidence to the contrary. The problem with a biased sampling of higher taxa is that many gene duplications are inferred to be more recent than they may actually be because the data are missing from lower taxa. For example, if a particular set of paralogous genes is sampled from a bird and a mammal, but data are missing for one of the genes from frogs and fish, then we infer that the duplication event occurred after the origination of tetrapods and before the origin of amniotes. For most genes, it is likely that there is extensive homology across vertebrate classes. Thus, if the data are missing (rather than the gene is not present) from the lower taxa, then

the duplication event can be pushed back in time, and we have elected to push duplication events to the period before the origin of fishes (see fig. 3). This assumption has been invoked repeatedly (see, e.g., Miklos and Rubin 1996; Spring 1997) and will favor acceptance of hypotheses invoking repeated genome duplications. In most cases, though, it is possible to push most of the ambiguous duplication events back to the base of the tree (before the protostome-deuterostome split). Therefore, our particular method is biased in favor of hypotheses postulating repeated genome duplications before the origin of fishes.

## Results

The pattern of diversification varies considerably across gene families. In some cases, there are a relatively large number of duplications, and most (or all) are inferred to have occurred before the divergence of protostomes and deuterostomes (e.g., dynein, Ras, Ubicon). Some families are relatively small, and most gene duplications (on internal branches) occurred before the origin of fishes (MyoD, Opsin). Other gene families show evidence of gene duplication across the span of vertebrate evolutionary history. Perhaps the most notable feature of the summary is that the timing of gene duplications depends critically on assumptions about sampling. Based on the current sampling of genes, most gene families exhibit gene duplications occurring along all internal branches of the tree. For instance, there are 58, 56, 20, and 72 duplications inferred to have occurred along the lineage leading to vertebrates, tetrapods, amniotes, and mammals, respectively, amounting to a variance in the duplication rate of approximately 491 (table A2). By contrast, if we assume that most gene duplications occurred before the origin of fishes when data from fishes are lacking, then there are 198, two, six, and two inferred duplications along the lineage leading to vertebrates, tetrapods, amniotes, and mammals, respectively, amounting to a variance in duplication rate of approximately 9,451 (table A3). The difference in the timing and pattern of gene family elaboration is evident when the number of genes summed across all gene families is plotted for various intervals of time defined by the internodes in the species tree (fig. 4). Clearly the sampling of gene family diversity across taxa has an enormous influence on whether the predictions of alternative hypotheses are met. Researchers need to concentrate efforts on sampling gene families in lower vertebrates, especially fishes, for resolution of debates focusing on the rate and pattern of gene family elaboration in vertebrates.

Tabulation of orthologous relationships between genes sampled from invertebrates and vertebrates shows little evidence favoring the predictions of the tetralogy hypoth-

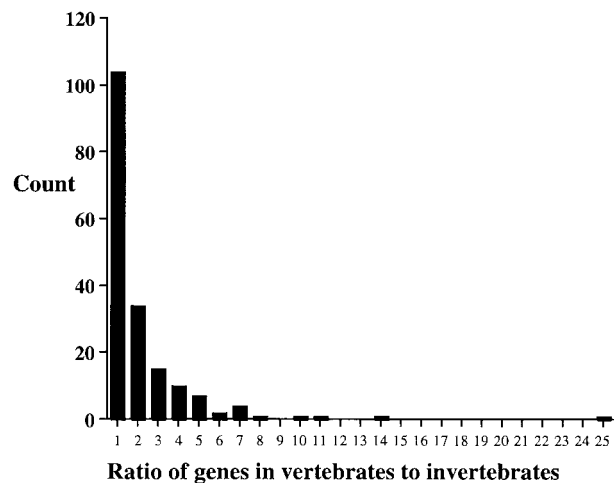


**Figure 4:** Pattern of increase in number of genes for various episodes. Solid circles are the observed numbers of genes summed across all gene families inferred from reconciled trees. Open circles are numbers of genes if we adopt the assumption that all gene duplications happened early in the history of vertebrates (i.e., before the origin of fishes), unless there is evidence to the contrary.

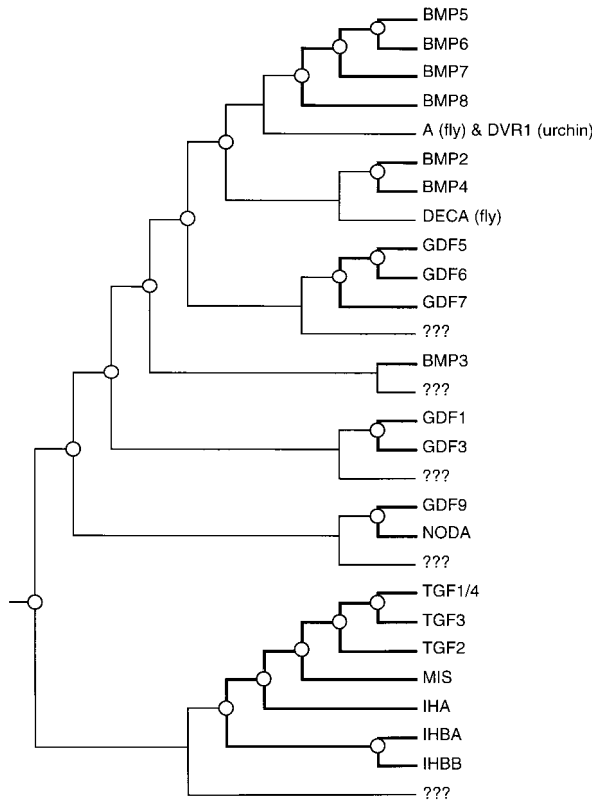
esis—namely, that there are multiple orthologous copies of genes in vertebrates for a single gene in invertebrates. This is evident from the frequency histogram of the ratio of phylogenetically related gene copies in vertebrates and invertebrates (fig. 5), which lacks the multimodality predicted if genome duplications played a major role in the development of the vertebrate genome. The tree of the TGF- $\beta$  gene family provides a good example of the level of complexity evident for many gene family trees. It also shows that relying on simple ratios of gene number in different groups of divergent taxa, tantamount to testing hypotheses of gene family evolution in a nonphylogenetic context, can be misleading (fig. 6). For example, there are four bone morphogen proteins (BMP5-8) that are related to a single gene in the fly and urchin, a result noted by Spring (1997) in support of the tetralogy hypothesis. However, the relationship among the four BMP genes does not match the predictions of the hypothesis. If there were two genome duplications, there should be two distinct gene clades, each with two members (i.e., Bmp5-Bmp6 and Bmp7-Bmp8). The observed clustering could have resulted from three genome duplications followed by multiple gene losses, however—a hypothesis invoked to explain the relationships among Hox genes in vertebrates (Bailey et al.

1997). Spring (1997) also lists the TGF 1–4 genes as a tetralogous group, noting that the failure to identify an ortholog in an invertebrate may be due to recent origination of this gene family. The gene tree suggests that this clade originated early in the evolutionary history of the gene family, however. Similar observations apply for most of the gene trees. Although there are gene families and gene clades within families that match the predictions of genome duplication hypotheses (MyoD, Creatine kinase, and Hsp70), the vast majority of gene clades do not match the predictions of these hypotheses.

Finally, for the 35 gene families sampled, estimates of tree imbalance show a broad scatter of points, with most points occurring within the 95% confidence limits expected if gene lineages duplicate independently and with identical probability (fig. 7). There is no evidence, when all gene families are considered, for a strong signal of genome duplication in the shape of gene trees. Moreover, none of the tree shapes are more symmetrical than the 95% confidence interval for random trees, and the observed  $I_c$  values are all larger than expected  $I_c$  values assuming two successive rounds of genome duplication and the survival of all four gene lineages. If, however, there has been significant gene loss since the hypothesized genome duplications, then tree shapes should be less symmetrical than expected if all gene lineages survived. Even if there is gene loss, most  $I_c$  values will be less than the expectations for trees in which duplications occur independently and with identical probability, except when the



**Figure 5:** The frequency distribution of the ratio of vertebrate-to-invertebrate orthologs for all gene families, except Dynein and Ras. (For most of the gene clades of these latter gene families, the ratio is 1 : 1.) Values <1 were omitted. Hypotheses postulating successive genome duplications predict a peak in the distribution at 4 (or at 3; see fig. 2B).

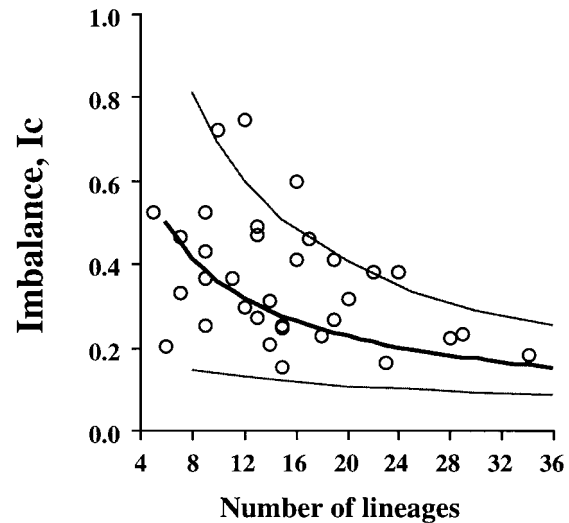


**Figure 6:** Inferred gene tree for the part of the TGF-beta family of transcription factors showing representative lineages for vertebrates (*thick lines*) and invertebrates (*thin lines*). Circles mark putative gene duplication events. Question marks denote that the gene inferred to exist based on reconciliation is missing because of either failure to sample the gene or gene extinction.

original gene trees are highly asymmetrical before genome duplications (table 2). This is not evident from the data. Values of  $I_c$  for almost two-thirds of the gene trees are greater than the expected value for random trees (fig. 7). Nevertheless, many of the larger gene trees may tend toward asymmetry because the gene tree, before putative genome duplications, was itself asymmetrical. This is evident for gene families with 20 or more paralogous genes. The TGF-B gene family illustrates this situation, as does many other of the larger gene families (ARE, Tubulin, Ras, G-proteins, POU, Forkhead). Thus, although there are clear limitations of tree shape analysis for testing hypotheses of genome diversification, gene trees still should have been more symmetric than that expected for gene trees resulting from duplications occurring independently and with identical probability across lineages. This is clearly not the case.

## Discussion

The predictions of alternative hypotheses of gene family diversification during the period of evolutionary history encompassing the origin of vertebrates were evaluated. Results from estimates of rates of gene duplication are equivocal and depend critically on the sampling of taxa. Limited sampling of fishes fails to provide adequate data to establish whether there was a fourfold increase in the rate of gene duplication before the origin of fishes. Although evidence indicates an increase in numbers of genes before the origin of fishes, the data are insufficient to fully refute the hypothesis that there has been a more or less constant rate of gene duplication (and a continuous increase in gene numbers) over the course of vertebrate evolution. In fact, based on the available data, there is little evidence for punctuation in gene family diversity at any time in the history of vertebrates. There is also little evidence for multiple rounds of genome duplications from analyses of gene relatedness for vertebrates and invertebrates. According to the architects of the tetralogy hypothesis, we should expect to see a multimodal distribution in the ratio of orthologous genes for vertebrates and invertebrates. This observation suggests either that gene family diversity arose through independent gene duplications or that postulated ancient



**Figure 7:** Measures of tree asymmetry for various gene families. Tree shape for each gene tree was calculated using Colless's  $I_c$  value. Circles are observed tree imbalance, and the thick line shows expected  $I_c$  values for random tree shapes. Thin lines are the upper and lower 95% confidence limits for the null hypothesis (i.e., Markov null model). Values of  $I_c$  for gene family trees under hypotheses postulating successive rounds of genome duplication should be less than the expected value for random trees, and most should be lower than the 95% confident limit, unless gene loss has been extensive (see table 2).

**Table 2:** Summary of the influence of gene loss on estimates of tree shape

	Number of genes lost/initial number of genes									
	0/8	0/12	0/16	0/20	3/12	4/16	5/20	8/16	12/24	16/32
Sym	≪	≪	≪	≪	≪	≪	≪	≪	<	≪
Asym		≪	≪	<	<	<	>	<	>	>

Note: Double arrow indicates that the average  $I_c$  value is less than the lower 95% confidence interval expected for random-shaped trees (see figure 7); less than sign indicates that average  $I_c$  value is less than the  $I_c$  value expected for random-shaped trees; greater than sign indicates average  $I_c$  value is greater than the  $I_c$  value expected for random-shaped trees. Sym and Asym refer to the shape of the original tree before pruning, where Sym and Asym are the most and least symmetric trees possible, respectively (see "Methods").

genome duplications have been erased by the continuous process of gene birth and death characteristic of gene family evolution (Nei et al. 1997). Over time, the process of gene birth and death will slowly erode gene orthology across taxa, even though the corresponding genes themselves may encode functional homologous proteins. Nevertheless, the long list of putative tetralogs (Spring 1997; and see fig. 2B) suggests that we should observe a second peak in the distribution of the ratio of vertebrate to invertebrate genes near three or four.

Finally, analysis of tree shape refutes the whole-genome duplication hypotheses because there is no indication for a trend toward symmetry of gene trees. Instead, all of the gene trees conform to the predictions of a model in which gene lineages duplicate independently and with identical probability. Lack of expected tree shape was also evident from Bailey et al.'s (1997) phylogenetic analysis of the Hox gene cluster. They suggested that there were three rounds of genome duplication followed by the loss of four whole clusters. Under this scenario, gene loss must have been extensive, a hypothesis that would explain the lack of a tendency toward symmetry of gene trees. Thus, there may have been successive rounds of genome duplication in some ancestral lineage of vertebrates, and the signature of this event has been wiped out by subsequent gene duplication and extinction. This is evident to some degree from our analysis of the effects of gene loss on tree shape. When gene loss was reasonably extensive (50% reduction), the average  $I_c$  value was similar to tree shapes of random trees, although the values were lower than the expectation for a random tree. If gene loss was extensive, or if there have been repeated gene duplication events overlaying early genome duplications, then hypotheses invoking multiple genome duplications early in the evolution of vertebrates may not be testable using the phylogenetic methods employed here.

Overall, the signal from our phylogenetic analysis suggests that gene families evolve by a process in which duplications of individual gene lineages are independent of

other gene duplications, and all gene lineages have similar probabilities of duplication. I failed to find any compelling evidence favoring the current, widely accepted hypothesis that the contemporary genome organization of vertebrates reflects two ancient tetraploidization events preceding the origin of fishes. The estimated three- to fourfold difference in numbers of coding genes between lower metazoans (flies, nematodes, sea urchins) and vertebrates that is implied by the tetralogy paradigm seems a gross simplification of differences in gene complexity across divergent taxa. For example, there are large, multigene families in vertebrates that are not present in lower metazoans (i.e., genes associated with the immune response). We did not include any of these genes in our analysis because of the shallow phylogenetic history of many of these gene families. Moreover, many paralogous genes trace their ancestry within species and are therefore not shared among divergent taxa (e.g., Hughes 1995). Similarly, some large gene families in mammals are less complex in lower vertebrates (e.g., olfactory receptor genes; Ngai et al. 1993; Glusman et al. 1996). Furthermore, there are many examples in which there are single homologs across all metazoans, which implies that for some cellular processes, there is no selective advantage of having multiple, divergent paralogs.

It is important to bear in mind that the sampling of the evolutionary history of genomic evolution is far from sufficient. To more fully understand the origin and evolution of genetic complexity, we need to adopt a broad comparative approach. With respect to the evolution of vertebrates, taking a step backward in time and sampling cephalochordates, agnathans, and Chondrichthyan fishes (Araki et al. 1996; Karabinos and Reimer 1997; Schlake et al. 1997) is a step in the right direction. Until we design experiments and collect data to explicitly test alternative hypotheses of gene family evolution, I suspect that we will be able to entertain many different evolutionary hypotheses (Skrabanek and Wolfe 1998).

*Implications for Gene Order*

There are conflicting reports regarding the phylogenetic relationships among putative paralogous genes that are members of syntenic groups on different chromosomes. Phylogenetic analysis of putative paralogous chromosomal regions (4p16, 5q33-35, 8p12-21, and 10q24-26) was taken as support for the hypothesis of two large-scale duplications (Pebusque et al. 1998). It is difficult to reconcile the hypothesis of two large-scale duplications with the two gene trees. First, the two gene trees do not show concordance of inferred relationships among the chromosomes. The ankryin tree supports an (8,(10,4)) clustering, whereas EGR supports a (2,(5,(8,10))) relationship. (Presumably the ankryin paralog on chromosome 5 was lost.) It is conceivable that the EGR4 gene translocated from the fourth to the second chromosome after duplication; however, if the ankryin tree reflects the relationships among chromosomes, then the EGR4 gene should have clustered as the sister taxa to the gene on chromosome 10. Finally, assuming that the trees are accurate portraits of relationships, then the 4 : 1 ratio evident for the EGR gene (which would be taken as further evidence of the tetralogy hypothesis) does not reflect true tetralogy. The putative ortholog of EGR1-3 in invertebrates is missing, and the one copy sampled from invertebrates appears orthologous to only one gene in vertebrates. If we root the tree with invertebrates (and therefore eliminate the problem of the missing gene), the tree shape is asymmetrical and does not match the hypothesis of two large-scale duplications. A similar analysis by Hughes (1998) for chromosomes 6, 9, and 1 failed to yield evidence favoring whole-genome or large-scale chromosomal duplication.

If there were not two, or more, large-scale duplications (either block or whole-genome duplications), then how do we explain putative paralogy of chromosomes (Lundin 1993)? Hughes (1998) entertains the hypothesis that gene clustering is favored by natural selection. This hypothesis implies that translocation is a frequent event in the genome, and increasing evidence indicates that this is true (e.g., Rowen et al. 1996; Saxena et al. 1996). Hughes (1998) speculates that selection may group genes with similar patterns of expression and length of encoded proteins. Whatever the cause of synteny across chromosomes, the results reported here support Hughes's (1998, p. 866) argument that "the existence of two clusters of homologous genes on different chromosomes cannot be taken as evidence of a simultaneous or 'block' duplication event." Because observed patterns reflect the action of mutation and natural selection, careful tests of hypotheses must distinguish between the effects of these two processes.

*Directional Evolution and Increased Genomic Complexity*

There is growing evidence for an increase in the rate of gene duplication before the origin of fishes. If we refute the importance of whole-genome duplications, then episodic advances in gene family complexity can only occur as a consequence of directional selection favoring duplication mutations. Surveys of gene mutations in animals reveal an abundance of regional DNA duplications in which portions of genes, whole genes, small regions of chromosomes, and whole chromosomes are duplicated (Brown et al. 1990, 1996; Currie and Sullivan 1994; Irwin 1995; Tomarev et al. 1995; Jobling et al. 1996; Rowen et al. 1996; Saxena et al. 1996; Gilbert et al. 1997; Kenck et al. 1997; Potier et al. 1998; Ritchie et al. 1998; SanMiguel et al. 1998). Estimated rates of unequal crossing over and replicative transposition are sufficiently high that apparent episodic increase in gene number at the base of vertebrates can be accounted for by independent regional duplications (e.g., Shapira and Finnerty 1986; Fryxell 1996; SanMiguel et al. 1998). Although invoking a large number of duplication events is not as parsimonious as whole-genome duplication, evidence for high rates of gene duplication suggests that rapid changes in genome size and complexity are possible.

Duplicate genes are likely to be functionally redundant and probably have limited consequences for fitness, relative to gene deletion events (which are expected to occur with similar frequency by unequal crossing over). This asymmetry of fitness effects between the gain and loss of genes by unequal crossing-over, coupled with the replicative bias of transposition-mediated gene duplication, will increase the number of genes over time (unless there is strong selection against increasing genome size). Fixation of duplicate loci may have occurred as a consequence of indirect selection (e.g., on cell size; Szarski 1983; Roth et al. 1997) or because of direct selection on fitness effects (Ohno 1970; Hughes 1994b; Kondo et al. 1996; Sidow 1996; Nowak et al. 1997; Brown et al. 1998). It is also possible that fixation of duplicate loci was enhanced by prior duplication events, a cascading, pleiotropic process that may have resulted in higher duplication rates. Episodic advance in gene family diversity may have also been assisted if the underlying mutation rates for gene duplications was higher in ancestral vertebrates than observed in *Drosophila* and humans.

**Conclusions**

Evaluation of alternative hypotheses about the evolution of genome complexity in vertebrates using phylogenetic methods indicates that, at this time, there is little evidence favoring the widespread belief that the contemporary ver-

tebrate genome organization resulted from multiple tetraploidization events before the origin of fishes. The data fail to refute an alternative hypothesis—namely, that gene family complexity increases through independent gene duplications. My argument is not that whole-genome duplications did not play a role in establishing the contemporary vertebrate gene organization but that the relative importance of successive rounds of tetraploidization is not apparent from phylogenetic analysis of available data. If there were successive genome duplications, subsequent expansion and contraction of gene families has effectively

silenced the genomic “big bang” implied by the tetralogy hypothesis.

### Acknowledgments

M. Ronshaugen performed many of the analyses reported in this article en route to an M.S. I am grateful to A. De Queiroz, S. Edwards, J. Wilcox, G. Wray, and two anonymous reviews for comments on the manuscript. This work was supported by the National Science Foundation grant DEB-9628094.

## APPENDIX

**Table A1:** Summary statistics from the complete phylogenetic analysis of 17 gene families

Gene	NTax	NChar	NJ			Parsimony			Likelihood			Least squares		
			TL	CI	RI	TL	CI	RI	TL	CI	RI	TL	CI	RI
ARF <sup>a</sup>	43	208	1,451	.63	.70	1,438	.63	.70	1,537	.59	.65	1,450	.63	.70
Arre	28	363	1,228	.75	.85	1,222	.75	.86	1,228	.75	.85	1,228	.75	.85
PAX	29	129	221	.82	.91	221	.82	.91	224	.81	.91	221	.82	.91
POU <sup>a</sup>	40	78	198	.84	.93	192	.87	.94	214	.78	.89	197	.84	.93
TGF <sup>a</sup>	76	111	939	.61	.87	920	.62	.88	1,114	.48	.80	943	.61	.87
WNT <sup>a</sup>	66	444	1,469	.54	.73	1,449	.54	.74	1,577	.5	.69	1,475	.53	.73
Hsp70 <sup>a</sup>	63	729	4,363	.55	.70	ND	ND	ND	4,818	.49	.63	4,395	.54	.70
Hsp90 <sup>a</sup>	31	778	2,251	.66	.79	ND	ND	ND	2,690	.63	.76	2,542	.66	.79
NGF	31	128	354	.81	.92	350	.82	.92	368	.78	.90	354	.81	.92
CK <sup>a</sup>	33	405	1,655	.70	.78	1,650	.70	.79	1,708	.68	.76	1,660	.69	.78
FGF <sup>a</sup>	41	155	694	.75	.90	686	.76	.90	710	.73	.89	689	.75	.90
G-pro <sup>a</sup>	36	450	2,308	.70	.80	2,305	.71	.80	2,336	.70	.78	2,319	.70	.79
FkHd <sup>a</sup>	30	119	464	.73	.738	ND	ND	ND	498	.68	.67	476	.72	.73
Dyna <sup>a</sup>	23	230	623	.82	.87	612	.83	.88	638	.80	.86	630	.81	.86
BAND <sup>a</sup>	23	154	638	.79	.82	633	.80	.82	651	.77	.80	633	.80	.82
Matrix <sup>a</sup>	39	224	1,228	.65	.73	ND	ND	ND	1,278	.63	.70	1,213	.66	.74
ANP <sup>a</sup>	26	109	569	.80	.84	566	.80	.85	597	.76	.81	570	.80	.84

Note: Ntax = number of taxa; Nchar = number of amino acids in alignment; NJ = neighbor joining; TL = total length of tree; CI = consistency index; RI = retention index. Values for gene trees obtained from the literature, and values for the gene trees for which only neighbor-joining trees were estimated, are not included. ND indicates that there were a large number of equally parsimonious trees, and we therefore omitted the data from the analysis.

<sup>a</sup> Maximum likelihood PUZZLE tree contained polytomies.

Table A2: Tally of the number of gene duplications for periods of evolutionary history depicted in figure 3

Gene	Time periods (from figure 3)					
	1	2	3	4	5	6
PAX	3	0	2	0	2	4
GATF	4	0	0	4	0	1
ANP	...	...	3	1	0	0
Arrestin	0	0	2	1	0	0
ARF	8	0	0	0	0	2
BAND	5	1	0	0	0	2
FGF	...	...	4	4	0	2
NGF	...	...	3	2	0	0
Gprot	3	0	0	0	0	10
Matrixin	1	1	0	2	0	4
ETS	10	0	0	4	0	9
Forkhead	10	0	0	1	0	2
Dynamin	3	0	1	0	0	3
TGF	9	0	1	6	2	5
POU	12	0	0	1	1	1
CK	1	2	1	0	1	0
WNT	9	0	0	2	0	0
Opsin	0	0	4	0	0	0
Nicotinic rec	4	0	10	1	0	0
MADS	7	0	0	0	0	0
Hsp90	1	0	0	0	1	0
DLX	3	0	5	1	0	0
Cyclin	7	0	0	2	0	2
Achol	7	0	5	0	2	0
Tubulin	12	0	1	1	3	0
Serpin	3	0	0	0	3	9
Ras	24	0	3	0	1	2
Trypsinogen	2	0	5	0	0	6
Hsp70	4	0	2	0	1	0
Filament	1	0	4	14	2	6
MyoD	1	0	2	1	0	0
Ubicon	16	0	0	0	0	1
Cysteine protease	8	0	0	0	0	1
Steroid receptors	15	0	0	8	1	0
Dynein	19	0	0	0	0	0
Duplications	212	4	58	56	20	72
Number of genes	213	217	275	331	351	423
Duration	NA	250	100	100	100	100
Rate (dups/my)	NA	.016	.58	.56	.2	.72
Rate (dup/my/gene)	NA	8E-05	.003	.002	6E-04	.002

**Table A3:** Tally of the number of gene duplications for periods of evolutionary history depicted in figure 3

Gene	Time periods (from figure 3)					
	1	2	3	4	5	6
PAX	3	0	5	0	0	0
GATF	4	0	4	0	0	0
ANP	...	...	4	0	0	0
Arrestin	0	0	3	0	0	0
ARF	8	0	2	0	0	0
BAND	5	1	2	0	0	0
FGF	...	...	9	0	0	0
NGF	...	...	3	0	2	0
Gprot	3	0	10	0	0	0
Matrixin	1	1	6	0	0	0
ETS	10	0	12	0	1	0
Forkhead	10	0	2	0	1	0
Dynamin	3	0	5	0	0	0
TGF	9	0	14	0	0	0
POU	12	0	2	0	1	0
CK	1	2	2	0	0	0
WNT	9	0	2	0	0	0
Opsin	0	0	4	0	0	0
Nicotinic rec	4	0	10	1	0	0
MADS	7	0	0	0	0	0
Hsp90	1	0	1	0	0	0
DLX	3	0	5	1	0	0
Cyclin	7	0	4	0	0	0
Achol	7	0	9	0	0	0
Tubulin	12	0	5	0	0	0
Serpin	3	0	14	0	0	0
Ras	24	0	6	0	0	0
Trypsinogen	2	0	10	0	0	1
Hsp70	4	0	2	0	1	0
Filament	1	0	26	0	0	1
MyoD	1	0	3	0	0	0
Ubicon	16	0	1	0	0	0
Cysteine protease	8	0	1	0	0	0
Steroid receptors	15	0	10	0	0	0
Dynein	19	0	0	0	0	0
Duplications	212	4	198	2	6	2
Number of genes	213	217	415	417	423	425
Duration	NA	250	100	100	100	100
Rate (dups/my)	NA	.016	1.98	.02	.06	.02
Rate (dup/my/gene)	NA	7E-05	.009	5E-05	1E-04	5E-05

Note: Timing of gene duplications adjusted to account for missing data (see "Methods").

## Literature Cited

- Allendorf, F. W., and R. G. Danzmann. 1997. Secondary tetrasomic segregation of MDH-B and preferential pairing of homeologues in rainbow trout. *Genetics* 145: 1083–1092.
- Amores, A., A. Force, Y. L. Yan, L. Joly, C. Amemiya, A. Fritz, R. K. Ho, et al. 1998. Zebrafish hox clusters and vertebrate genome evolution. *Science* (Washington, D.C.) 282:1711–1714.
- Aparicio, S., K. Hawker, A. Cottage, Y. Mikawa, L. Zuo, B. Venkatesh, E. Chen, R. Krumlauf, and S. Brenner. 1997. Organization of the *Fugu rubripes* Hox clusters: evidence for continuing evolution of vertebrate Hox complexes. *Nature Genetics* 16:79–83.
- Araki, I., K. Terazawa, and N. Satoh. 1996. Duplication of an amphioxus myogenic bHLH gene is independent of vertebrate myogenic bHLH gene duplication. *Gene* 171: 231–236.
- Atchley, W. R., W. M. Fitch, and M. Bronner-Fraser. 1994. Molecular evolution of the MyoD family of transcription factors. *Proceedings of the National Academy of Sciences of the USA* 91:11522–11526.
- Bailey, W. J., J. Kim, G. P. Wagner, and F. H. Ruddle. 1997. Phylogenetic reconstruction of vertebrate Hox cluster duplication. *Molecular Biology and Evolution* 14: 843–853.
- Bonner, J. T. 1988. *The evolution of complexity by means of natural selection*. Princeton University Press, Princeton, N.J.
- Brenner, S., G. Elgar, R. Sandford, A. Macrae, B. Venkatesh, and S. Aparicio. 1993. Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome. *Nature* (London) 366:265–268.
- Brown, C. J., C. F. Aquadro, and W. W. Anderson. 1990. DNA sequence evolution of the amylase multigene family in *Drosophila pseudoobscura*. *Genetics* 126:131–138.
- Brown, C. J., K. M. Todd, and R. F. Rosenzweig. 1998. Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Molecular Biology and Evolution* 15:931–942.
- Brown, M. A., C. F. Xu, H. Nicolai, B. Griffiths, J. A. Chambers, D. Blackand, and E. Solomon. 1996. The 5' end of the BRCA1 gene lies within a duplicated region of human chromosome 17q21. *Oncogene* 12: 2507–2513.
- Clark, A. G. 1994. Invasion and maintenance of a gene duplication. *Proceedings of the National Academy of Sciences of the USA* 91:2950–2954.
- Colless, D. H. 1982. Review of phylogenetics: the theory and practice of phylogenetic systematics. *Systematic Zoology* 31:100–104.
- Currie, P. D., and D. T. Sullivan. 1994. Structure, expression and duplication of genes which encode phosphoglyceromutase of *Drosophila melanogaster*. *Genetics* 138: 353–368.
- Detera-Wadleigh, S. D., and T. G. Fanning. 1994. Phylogeny of the steroid receptor superfamily. *Molecular Phylogenetics and Evolution* 3:192–205.
- Eddy, S. R. 1996. Hidden Markov models. *Current Opinions in Structural Biology* 6:361–365.
- Felsenstein, J. 1993. PHYLIP (phylogeny inference package), version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.
- Felsenstein, J., and G. A. Churchill. 1996. A hidden Markov model approach to variation among sites in the rate of evolution. *Molecular Biology and Evolution* 13:93–104.
- Fields, C., M. D. Adams, O. White, and J. C. Venter. 1994. How many genes in the human genome? *Nature Genetics* 7:345–346.
- Fryxell, K. J. 1995. The evolutionary divergence of neurotransmitter receptors and second-messenger pathways. *Journal of Molecular Evolution* 41:85–97.
- . 1996. The coevolution of gene family trees. *Trends in Genetics* 50:98–103.
- Garcia-Meunier, P., M. Etienne-Julan, P. Fort, M. Piechaczyk, and F. Bonhomme. 1993. Concerted evolution in the GAPDH family of retrotransposed pseudogenes. *Mammalian Genome* 4:695–703.
- Gaut, B. S., and J. F. Doebley. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proceedings of the National Academy of Sciences of the USA* 94:6809–6814.
- Gilbert, J.-M., E. Mouchel-Vielh, and J. S. Deutsch. 1997. Engrailed duplication events during the evolution of barnacles. *Journal of Molecular Evolution* 44:585–594.
- Glusman, G., S. Clifton, B. Roe, and D. Lancet. 1996. Sequence analysis in the olfactory receptor gene cluster on human chromosome 17: recombinatorial events affecting receptor diversity. *Genomics* 37:147–160.
- Gordon, R. 1994. Evolution escapes rugged fitness landscapes by gene or genome doubling: the blessing of higher dimensionality. *Computers in Chemistry* 18: 325–331.
- Gould, S. J., D. M. Raup, J. J. Sepkowski, T. J. Schopf, and D. S. Simberloff. 1977. The shape of evolution: a comparison of real and random clades. *Paleobiology* 3: 23–40.
- Guigo, R., I. Muchnik, and T. F. Smith. 1996. Reconstruction of ancient molecular phylogeny. *Molecular Phylogenetics and Evolution* 6:189–213.
- Heard, S. B. 1992. Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. *Evolution* 46:1818–1826.
- Heard, S. B., and A. Ø. Mooers. 1996. Imperfect infor-

- mation and the balance of cladograms and phenograms. *Systematic Biology* 45:115–118.
- Holland, P. W., J. Garcia-Fernandez, N. A. Williams, and A. Sidow. 1994. Gene duplications and the origin of vertebrate development. *Development Supplement* 1994:125–133.
- Holland, P. W., H. Holland, and J. Garcia-Fernandez. 1996. Hox genes and chordate evolution. *Developmental Biology* 173:382–395.
- Holloway, A. J., N. G. Della, C. F. Fletcher, D. A. Lagespada, N. G. Copeland, N. A. Jenkins, and D. D. Bowtell. 1997. Chromosomal mapping of five highly conserved murine homologues of the *Drosophila* ring finger gene seven-in-absentia. *Genomics* 41:160–168.
- Huelsenbeck, J. P., and D. M. Hillis. 1993. Success of phylogenetic methods in the four-taxon case. *Systematic Biology* 42:247–264.
- Hughes, A. L. 1994a. Evolution of cystein proteinases in eukaryotes. *Molecular Phylogenetics and Evolution* 3:310–321.
- . 1994b. The evolution of functionally novel proteins after gene duplication. *Proceedings of the Royal Society of London B, Biological Sciences* 256:119–224.
- . 1995. The evolution of the type I interferon gene family in mammals. *Journal of Molecular Evolution* 41:539–548.
- . 1998. Phylogenetic test of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9 and 1. *Molecular Biology and Evolution* 15:854–870.
- Irwin, D. M. 1995. Evolution of the bovine lysozyme gene family: changes in gene expression and reversion of function. *Journal of Molecular Evolution* 41:299–312.
- Iwabe, N., K. Kuma, and T. Miyata. 1996. Evolution of gene families and relationship with organismal evolution: rapid divergence of tissue-specific genes in the early evolution of chordates. *Molecular Biology and Evolution* 13:483–493.
- Jobling, M. A., V. Samara, A. Pandya, N. Fretwell, B. Bernasconi, R. J. Mitchell, T. Gerelsaikhani et al. 1996. Recurrent duplication and deletion polymorphisms on the long arm of the Y chromosome in normal males. *Human Molecular Genetics* 5:1767–1775.
- Karabinos, A., and D. Reimer. 1997. The single calmodulin gene of the cephalochordate *Branchiostoma*. *Gene* 195:229–233.
- Kasahara, M., J. Nakaya, Y. Satta, and N. Takahata. 1997. Chromosomal duplication and the emergence of the adaptive immune system. *Trends in Genetics* 13:90–92.
- Kenck, C., P. Bugert, M. Wilhelm, and G. Kovacs. 1997. Duplication of an approximately 1.5 Mb DNA segment at chromosome 5q22 indicates the locus of a new tumour gene in nonpapillary renal cell carcinomas. *Oncogene* 14:1093–1098.
- Kirkpatrick, M., and M. Slatkin. 1993. Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution* 47:1171–1181.
- Kondo, H., M. Ino, A. Suzuki, H. Ishizaki, and M. Iwami. 1996. Multiple gene copies for bombyxin, an insulin-related peptide of the silkworm *Bombyx mori*: structural signs for gene rearrangement and duplication responsible for generation of multiple molecular forms of bombyxin. *Journal of Molecular Biology* 259:926–937.
- Leipoldt, M. 1983. Towards an understanding of the molecular mechanisms regulating gene expression during diploidization in phylogenetically polyploid lower vertebrates. *Human Genetics* 65:11–18.
- Le Novere, N., and J.-P. Changeux. 1995. Molecular evolution of the nicotinic acetylcholine receptor: an example of a multigene family in excitable cells. *Journal of Molecular Evolution* 40:155–172.
- 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.
- McClure, M. A., T. K. Vasi, and M. W. Fitch. 1994. Comparative analysis of multiple protein-sequence alignment methods. *Molecular Biological Evolution* 11:571–592.
- McShea, D. 1996. Metazoan complexity and evolution: is there a trend? *Evolution* 50:477–492.
- Miklos, G. L., and G. M. Rubin. 1996. The role of the genome project in determining gene function: insights from model organisms. *Cell* 86:521–529.
- Mooers, A. Ø., and S. B. Heard. 1997. Inferring evolutionary process from phylogenetic tree shape. *Quarterly Review of Biology* 72:31–54.
- Mooers, A. Ø., R. D. M. Page, A. Purvis, and P. H. Harvey. 1995. Phylogenetic noise leads to imbalanced tree reconstructions. *Systematic Biology* 44:332–432.
- Nadeau, J. H., and D. Sankoff. 1997. Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics* 147:1259–1266.
- Nei, M., X. Gu, and T. Sitnikova. 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proceedings of the National Academy of Sciences of the USA* 94:7799–7806.
- Ngai, J., M. M. Dowling, L. Buck, R. Axel, and A. Chess. 1993. The family of gene encoding odorant receptors in the channel catfish. *Cell* 72:657–666.
- Nowak, M. A., M. C. Boerlijst, J. Cooke, and J. M. Smith. 1997. Evolution of genetic redundancy. *Nature (London)* 388:167–171.
- Ohno, S. 1970. *Evolution by gene duplication*. Springer, New York.

- . 1998. The notion of the Cambrian pananimalia genome and a genomic difference that separated vertebrates from invertebrates. *Progress in Molecular and Subcellular Biology* 21:97–117.
- Ohno, S., U. Wolf, and N. B. Atkin. 1967. Evolution from fish to mammals by gene duplication. *Hereditas* 59: 169–187.
- Page, R. D. 1993a. COMPONENT: tree comparison software for Microsoft Windows, version 2.0. Natural History Museum, London.
- . 1993b. Genes, organisms, and areas: the problem of multiple lineages. *Systematic Biology* 42:77–84.
- Page, R. D., and M. A. Chareleston. 1997. Reconciled trees, incongruent genes and species trees. In B. Mirkin, F. R. McMorris, F. S. Roberts, and A. Rzhetsky, eds. *Mathematical hierarchies in biology*. Vol. 21. American Mathematical Society, Providence, R.I.
- Pebusque, M.-J., F. Coulier, D. Birnbaum, and P. Pontarotti. 1998. Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *Molecular Biology and Evolution* 15: 1145–1159.
- Porter, M. E., J. A. Knott, S. H. Myser, and S. J. Farlow. 1996. The dynein gene family in *Chlamydomonas reinhardtii*. *Genetics* 144:569–585.
- Postlethwait, H. H., Y. L. Yan, M. A. Gates, S. Horne, A. Amores, A. Brownlie, A. Donovan, et al. 1998. Vertebrate genome evolution and the zebrafish gene map. 1998. *Nature Genetics* 18:345–349.
- Potier, M. C., A. Dutriaux, R. Orti, J. Groet, N. Gibelin, G. Karadima, G. Lutfalla, et al. 1998. Two sequence-ready contigs spanning the two copies of a 200-kb duplication on human 21q: partial sequence and polymorphisms. *Genomics* 51:417–426.
- Ramey, S. 1997. A performance comparison between the Cray-YMP2 and the Origin 2000. In B. Stevenson, ed. *Terabit*. National Supercomputing Center for Energy and the Environment, Las Vegas, Nev.
- Raup, D. M., S. J. Gould, T. J. Schopf, and D. Simberloff. 1973. Stochastic models of phylogeny and the evolution of diversity. *Journal of Geology* 81:525–542.
- Ritchie, R. J., M. G. Mattei, and M. Lalande. 1998. A large polymorphic repeat in the pericentromeric region of human chromosome 15q contains three partial gene duplications. *Human Molecular Genetics* 7:1253–1260.
- Roth, G., J. Blanke, and D. B. Wake. 1997. Genome size, secondary simplification, and the evolution of the brain in salamanders. *Brain Behavior and Evolution* 50:50–59.
- Rowen, L., B. F. Koop, and L. Hood. 1996. The complete 685kb DNA sequence of the human b T cell receptor locus. *Science (Washington, D.C.)* 272:1755–1762.
- Ruvinsky, I., and L. M. Silver. 1997. Newly identified paralogous groups on mouse chromosomes 5 and 11 reveal the age of a T-box cluster duplication. *Genomics* 40: 262–266.
- SanMiguel, P., B. S. Gaut, A. Tikhonov, Y. Nakajima, and J. L. Bennetzen. 1998. The paleontology of intergene retrotransposons of maize. *Nature Genetics* 20:43–45.
- Saxena, R., L. G. Brown, T. Hawkins, R. K. Alagappan, H. Skaletsky, M. P. Reeve, R. Reijo, et al. 1996. The DAZ gene cluster on the human Y chromosome arose from an autosomal gene that was transposed, repeatedly amplified and pruned. *Nature Genetics* 14:292–299.
- Schlake, T., M. Schorpp, M. Nehls, and T. Boehm. 1997. The nude gene encodes a sequence-specific DNA binding protein with homologs in organisms that lack an anticipatory immune system. *Proceedings of the National Academy of Sciences of the USA* 94:3842–3847.
- Semenov, M. V., and M. Snyder. 1997. Human dishevelled genes constitute a DHR-containing multigene family. *Genomics* 42:302–310.
- Shain, D. H., T. Neuman, and M. X. Zuber. 1997. Embryonic expression and evolution of duplicated E-protein genes in *Xenopus laevis*: parallels with ancestral E-protein genes. *Genetics* 146:345–353.
- Shapira, S. K., and V. G. Finnerty. 1986. The use of genetic complementation in the study of eukaryotic macromolecular evolution: rate of spontaneous gene duplication at two loci of *Drosophila melanogaster*. *Journal of Molecular Evolution* 23:159–167.
- Sharman, A. C., and P. W. Holland. 1998. Estimation of Hox gene cluster number in lampreys. *International Journal of Developmental Biology* 42:617–620.
- Sidow, A. 1996. Gen(om)e duplications in the evolution of early vertebrates. *Current Opinion in Genetics and Development* 6:715–722.
- Simberloff, D., K. L. Heck, E. D. McCoy, and E. F. Conner. 1981. There have been no statistical tests of cladistic biogeographical hypotheses. Pages 40–63 in G. Nelson and D. E. Rosen, eds. *Vicariance biogeography: a critique*. Columbia University Press, New York.
- Simmen, M. W., S. Leitgeb, V. H. Clark, S. J. Jones, and A. Bird. 1998. Gene number in an invertebrate chordate, *Ciona intestinalis*. *Proceedings of the National Academy of Sciences of the USA* 95:4437–4440.
- Skrabanek, L., and K. H. Wolfe. 1998. Eukaryote genome duplication—where's the evidence? *Current Opinions in Genetics and Development* 8:694–700.
- Sokal, R. R., and F. J. Rohlf. 1988. *Biometry*. W. H. Freeman, New York.
- Sonnhammer, E. L. L., S. R. Eddy, and R. Durbin. 1997. Pfam: a comprehensive database of protein families based on alignments. *Proteins* 28:405–420.
- Spring, J. 1997. Vertebrate evolution by interspecific hybridization—are we polyploid? *Federation of the European Biological Society Letters* 400:2–8. (Data avail-

- able at: <http://www.unibas.ch/dib/zoologie/research/tetrabase2.html>.)
- Stock, D. W., D. L. Ellies, Z. Zhao, M. Ekker, F. H. Ruddle, and K. W. Weiss. 1996. The evolution of the vertebrate *Dlx* gene family. *Proceedings of the National Academy of Sciences of the USA* 93:10858–10863.
- Strimmer, K., and A. von Haeseler. 1996. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution* 13:964–969.
- Strimmer, K., N. Goldman, and A. von Haeseler. 1997. PUZZLE: maximum likelihood analysis for nucleotide and amino acid alignments. (Program available at: <http://www.zi.biologie.uni-muenchen.de/~strimmer/puzzle.html>.)
- Swofford, D. L. 1993. PAUP: phylogenetic analysis using parsimony, version 3.1. Illinois Natural History Survey, Champaign.
- Szarski, H. 1983. Cell size and the concept of wasteful and frugal evolutionary strategies. *Journal of Theoretical Biology* 105:201–209.
- Theissen, G., J. T. Kim, and H. Saedler. 1996. Classification and phylogeny of the MADS-box multigene family suggest defined roles of MADS-box gene subfamilies in the morphological evolution of eukaryotes. *Journal of Molecular Evolution* 43:484–516.
- Tomarev, S. I., S. Chung, and J. Piatigorsky. 1995. Glutathione S-transferase and S-crystallins of cephalopods: evolution from active enzyme to lens-refractive proteins. *Journal of Molecular Evolution* 41:1048–1056.
- Walsh, J. B. 1995. How often do duplicated genes evolve new functions? *Genetics* 139:421–428.
- Wolfe, K. H., and D. C. Shields. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature (London)* 387:708–713.
- Yamaguchi, F., and S. Brenner. 1997. Molecular cloning of 5-hydroxytryptamine (5-HT) type I receptor genes from the Japanese puffer fish, *Fugu rubripes*. *Gene* 191: 219–223.
- Yokoyama, S. 1995. Amino acid replacements and wavelength absorption of visual pigments in vertebrates. *Molecular Biology and Evolution* 12:53–61.

Associate Editor: Gregory A. Wray