Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution

Xun Gu[1,2], Yufeng Wang[1–3] & Jianying Gu[1,2]

Published online: 28 May 2002, DOI: 10.1038/ng902

The classical (two-round) hypothesis[1] of vertebrate genome duplication proposes two successive whole-genome duplication(s) (polyploidizations) predating the origin of fishes, a view now being seriously challenged[2–7]. As the debate largely concerns the relative merits of the 'big-bang mode' theory[8–13] (large-scale duplication) and the 'continuous mode' theory (constant creation by small-scale duplications)[2–7,14], we tested whether a significant proportion of paralogous genes in the contemporary human genome was indeed generated in the early stage of vertebrate evolution. After an extensive search of major databases, we dated 1,739 gene duplication events from the phylogenetic analysis of 749 vertebrate gene families. We found a pattern characterized by two waves (I, II) and an ancient component. Wave I represents a recent gene family expansion by tandem or segmental duplications[15], whereas wave II, a rapid paralogous gene increase in the early stage of vertebrate evolution, supports the idea of genome duplication(s) (the big-bang mode). Further analysis indicated that large- and small-scale gene duplications both make a significant contribution during the early stage of vertebrate evolution to build the current hierarchy of the human proteome.

Ohno's hypothesis of two-rounds of genome duplication[1] in early vertebrates has recently become widely accepted because it seems to explain why vertebrates have four *HOX* gene clusters whereas invertebrates have only one[10,16]. The impact of this pattern of duplication on vertebrate development has been extensively discussed[10–16]. But the first draft of the human genome sequence[6,7] contains no conclusive evidence for these hypothetical genome duplications—for example, no pervasive duplicated chromosome blocks as are seen in yeast or *Arabidopsis thaliana*[3]. Although these phenomena can be explained as being the consequence of rapid chromosome rearrangements and gene losses after genome duplications[8], this raises the question of whether large-scale duplication(s) occurred in the early stages of vertebrate evolution[2,3,6–7,9].

We addressed this issue by estimating the age distribution of human gene families and testing whether a significant proportion of paralogous genes were generated in the early stage of vertebrate evolution (Fig. 1). We surveyed almost all gene families in vertebrates, resulting in a final data set (749 gene families) suitable for reliable phylogenetic analysis. Because of differential selection pressures on different duplicate genes or lineages, time

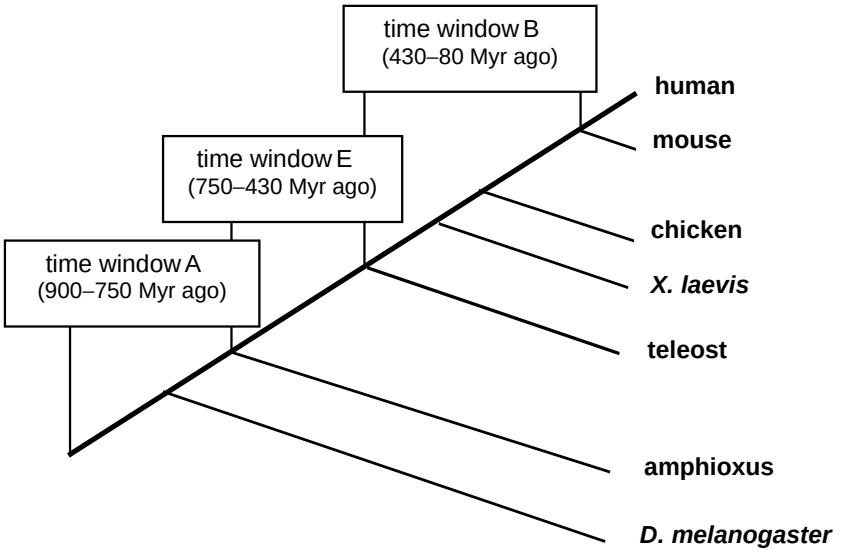

Fig. 1 Phylogenetic tree of major vertebrate groups. Duplication events occurring in the ancestral lineage of humans (bold) can be identified if both duplicates are found in the human genome. The early stage of vertebrates (time window *E*) was defined as the evolutionary period from the emergence of teleosts (approximately 430 Myr ago) to the vertebrate–amphioxus split (around 750 Myr ago). The controversy concerning Ohno's theory[1] is whether two-round genome duplications occurred during this period. In the ancestral lineage of humans, no large-scale duplication occurred in two time windows: (*A*) before the vertebrate–amphioxus split (750–900 Myr ago), and (*B*) from the mammalian radiation to the teleost–mammal split (80–430 Myr ago). Gene duplications identified in these two windows are thus used to estimate the average rate of small-scale duplication.

[1]*Department of Zoology and Genetics and* [2]*LHB Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, Iowa 50011, USA.*
[3]*Department of Bioinformatics, American Type Culture Collection (ATCC), Manassas, Virginia, USA. Correspondence should be addressed to X.G. (e-mail: xgu@iastate.edu).*

estimation in terms of the molecular clock was only approximate[11]. We therefore used several protocols to examine the robustness of our results. For illustration, Fig. 2 shows an analysis of the stimulating-hormone receptor gene family. The first and second duplication events were dated using the nearest-neighbor clock and the global clock[11], respectively.

We identified 1,739 duplication events, all of them occurring in the ancestral lineage of humans (Fig. 1). The age distribution of human paralogous genes shows a pattern characterized by two waves (I, II) and an ancient (III) component (Fig. 3). Wave I, which occurred after mammalian radiation, consisted of several large gene families (for example, the immuno-related molecules)[15,17–18]. Wave I was consistent with the recent finding that the activity of segmental or tandem duplications has been enhanced in primate genomes[15]. Wave II showed a rapid increase in the number of paralogous genes in the early stage of vertebrate evolution, as expected from the genome duplication hypothesis and the accompanying paper by McLysaght *et al.*[19]. This wave included a significant proportion of human tissue (development)-specific isoforms, supporting the suggestion that the zebrafish is a good model for functional genomics[20]. The ancient component (III) included duplication events that took place during metazoan evolution or earlier (Fig. 3). These gene duplications are thought to be important for the emergence of major signal transduction pathways[21].

These findings may help to resolve the recent two-round duplication controversy[1–14,16]. To facilitate the analysis, let us consider two competing hypotheses[9]. The big-bang mode hypothesis invokes large-scale duplication(s) to explain the observed rapid increase in the number of paralogous genes. In addition to polyploidization[1,14,16], possible mechanisms involved include partial-genome or single-chromosome duplication. The continuous mode hypothesis claims that paralogous genes are created constantly by small-scale (tandem or segmental) gene duplications[9].

In spite of substantial evidence for recent tandem or segmental duplications[15], most genome duplication models[1,14,16] have not recognized the role of small-scale duplications during the early stage of vertebrate evolution. As the evolutionary scenario should be a combination of the big-bang mode and the continuous mode, the real question concerns their relative importance. In our study, the early stage (E = 430–750 Myr ago) of vertebrate evolution was defined as the evolutionary time period before the emergence of teleosts but after the vertebrate–amphioxus split[22]. It is generally accepted that in the ancestral vertebrate lineage of humans (Fig. 1), there were two time windows in which no large-scale duplication(s) occurred: the ancestral lineage before the vertebrate–amphioxus split (A = 750–900 Myr ago) and the lineage from the mammalian radiation to the teleost–mammal split (B = 80–430 Myr ago). Given the premise that gene duplications during these time windows occur largely a result of the continuous mode, we estimated the average rate of small-scale duplication in time window A to be $\nu_A \approx 0.79$ per Myr, and that in time window B to be $\nu_B \approx 1.25$ per Myr.

Although the rate of small-scale duplication could vary during early-stage vertebrate evolution (time window E), the expected number of duplicate events from the continuous mode (n_c) depends only on the average rate for time window E (ν_E):

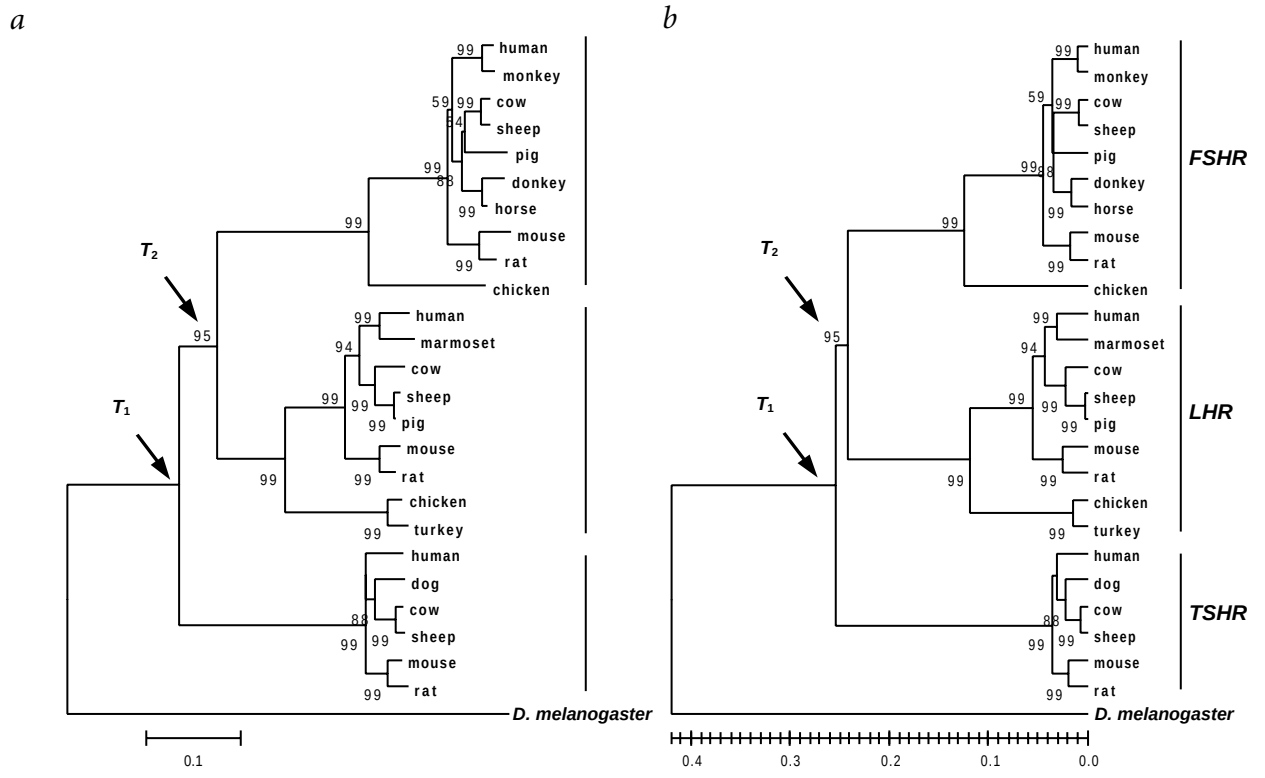

Fig. 2. Molecular evolutionary analysis of the stimulating-hormone receptor gene family. In vertebrates, this has three member genes: follicle-stimulating hormone receptor (*FSHR*), luteinizing hormone receptor (*LHR*) and thyroid-stimulating hormone receptor (*TSHR*). ***a***, The phylogenetic tree inferred by the neighbor-joining method[24] with Poisson distance. Boot-strapping values of more than 50% were presented. T_1 and T_2 are the time points of the first and second gene duplications, respectively. ***b***, The linearized neighbor-joining tree[28] used in converting evolutionary distances to the (relative) molecular time scale. By the global clock[25], we estimated T_1=491 Myr ago and T_2=451 Myr ago. By the nearest-neighbor clock, we found that the phylogenetic interval for both T_1 and T_2 is (T_{bird}, T_{fly}): that is, from the mammal–bird split to the fly–vertebrate split. Further calculation results in T_1 = 481 Myr ago and T_2 = 442 Myr ago, respectively.

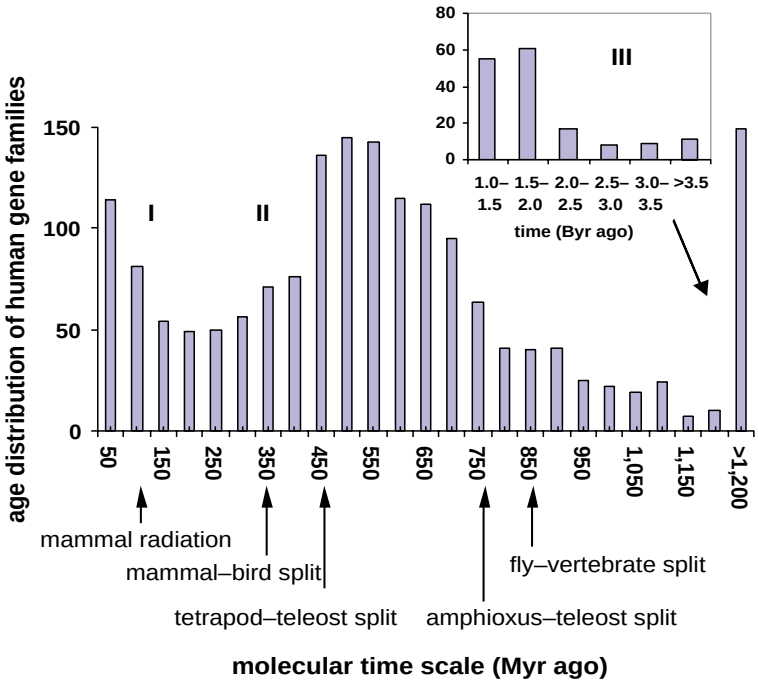

Fig. 3 Age distribution of human gene families. The molecular time scale is measured as Myr ago. Each bin for the histogram is 50 Myr, except for the most ancient one (>1,000 Myr ago). The relative positions of several major speciation events in the time scale are indicated. 1 Byr, 1 billion years (1,000 Myr).

thus $n_c = v_E \times T_E$, where $T_E = 750–430 = 320$ Myr. Because these three time windows (*B*, *E* and *A*) are adjacent, we assumed that v_E would not have differed greatly from the average rate of time window *A* or *B*, or $v_E \approx 0.79–1.25$ per Myr. It follows that, in the early (*E*) stage of vertebrate evolution, the expected number of duplication events arising from the continuous mode is 254–438, which is 30–52% of the total number (847) of duplication events. Notably, this indicates that during the early stage of vertebrate evolution, the big-bang mode and continuous mode both made a significant contribution to building the hierarchy of vertebrate gene families.

Our study used about one-quarter of human gene families[17], which contains the largest number of vertebrate gene families that can be used for estimating the molecular time scale of duplication events. Gene families were excluded mostly because they lacked orthologous genes. We extensively examined the robustness of the estimated age distribution of gene families. The size distribution of gene families (Fig. 4*a*) was almost identical to that of the human genome sequence[6]. The coefficient of correlation between time estimates by the nearest-neighbor clock or global clock was 0.93, with the average difference being approximately 5% (Fig. 4*b*). The cumulative age distributions estimated by these two methods were similar (Fig. 4*c*), indicating that the phylogenies were generally consistent with the molecular clock dating. In addition, the use of different calibration systems, including fossil data, evolutionary distances and speciation events, had almost no effect on the shape of the age distribution (Fig. 3), although a marginal shift in the time scale (of 2–5% on average) was observed. For example, a rapid increase in the number of duplicate genes was always seen, regardless of whether the vertebrate–*Drosophila* split was used for calibration (Fig. 4*d*).

Although long-term vertebrate evolution may have erased many historical details, our study suggests a plausible evolutionary scenario. Since the origin of chordates, small-scale (tandem and segmental) duplications have provided a continuous flux to create new paralogous genes. Before the mammalian radiation, the average rate of small-scale duplications was about three to five events per genome per million years (assuming that a quarter of human gene families are included). More recently (in mammals or primates), this rate may have increased up to approximately 10 events per genome per million years (calculated from wave I), which is consistent with the previous estimate made using a different approach[23]. In the early stage of vertebrate evolution, large-scale duplication(s) had a substantial impact on the evolution of tissue specificity. Overall, both mechanisms (continuous and big-bang) have made significant contributions to the contemporary hierarchy of human gene families.

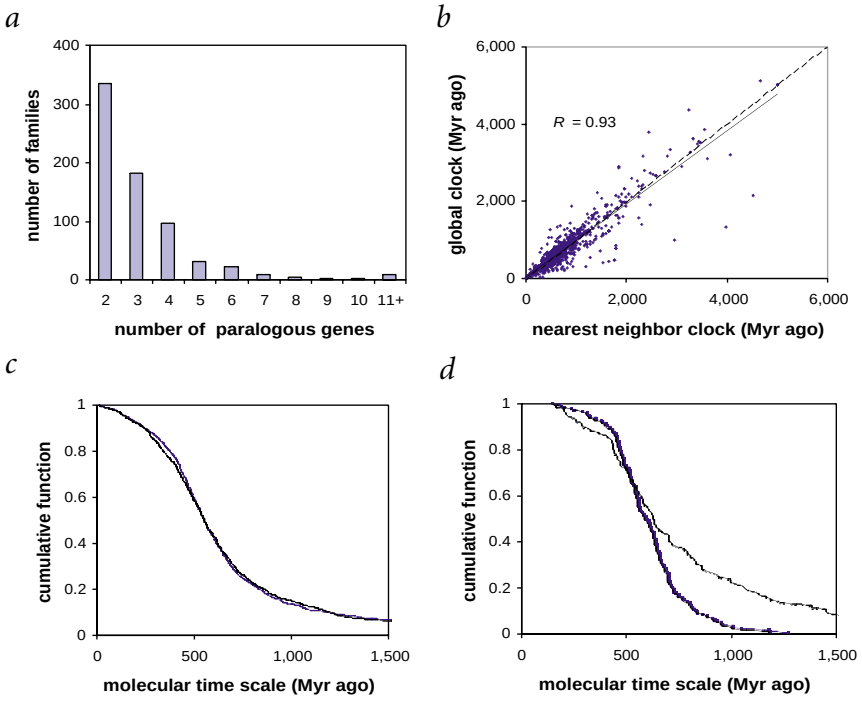

Fig. 4 Robustness of the age distribution of human paralogous genes. ***a***, Size distribution of 707 gene families used in the current study, which is similar to that in ref. 6. ***b***, Linear regression of 1,708 duplication time estimates between the nearest-neighbor clock and the global clock. The coefficient of correlation is 0.93. ***c***, Cumulative distribution of the age distribution estimated by nearest-neighbor clock or global clock was very similar. ***d***, Cumulative distributions estimated from a subset (283 gene families). The bold curve uses the inferred orthologous gene in fruit fly as a calibration, as this seems to be more sigmoid-shaped than the curve seen when sequences from fruit fly are excluded. Nevertheless, they share a qualitatively similar pattern.

The new model can be simply expressed as $mR + C$: that is, m round large-scale duplication(s) plus continuous small-scale creations. Any notion implying $m = 0$ or $C = 0$ is not sufficient to explain the data (Fig. 3). The hypothesis of multiple rounds ($m = 3$ or more) seems unnecessary because the inconsistent phylogenetic pattern[9] can be easily explained by the steady flux of small-scale duplications. Although our data showed a significant proportion of gene families with three or four members generated 430–750 Myr ago, we are cautious about whether $2R + C$, a natural extension of the classical two-round ($2R$) model, provides a better interpretation than the more parsimonious $1R + C$ model. We are developing a model-based likelihood ratio test to determine whether $m = 1$ is sufficient (data not shown). To resolve these questions, it will be necessary to sequence the genomes of a wide variety of vertebrates and to identify genome duplication(s) in other vertebrate branches, such as the teleosts or frogs[3].

Methods

Data collection and mining. We constructed a data-mining pipeline to find all vertebrate gene families suitable for use in estimating the age distribution of duplications. Amino-acid sequences of more than 3,000 vertebrate gene families were obtained from the HOVERGEN databases (http://pbil.univ-lyon1.fr/), each of which have at least two human homologous genes (see Web Table A online for details of the gene families studied). We then excluded redundant sequences by a procedure combining multiple alignment[24], phylogenetic analysis[25] and a literature search. Amino-acid sequences used in Fig. 2 are derived from fam000474 (HOVERGENE) for vertebrate genes and from AAB07030 (GenBank) for the *D. melanogaster* gene. Two *X. laevis* genes (AJ249845 and AJ249846) were not included because of their short sequence length, but the result remained virtually the same if they were used (data not shown).

Next, we excluded a gene family from further analysis if it contained only mammal sequences and the duplication event(s) was very ancient. This rule is based on previous experience indicating that dating ancient duplication events using only mammals for calibration is highly unreliable[11,26–28]. The cut-off point was defined according to the average ratio (z) of paralogous distance to orthologous distance (for example, human/mouse). A gene duplication is excluded if $z > 5.5$, roughly corresponding to the duplication event that occurred before the teleost–tetrapod split. When the cutoff varied between 4 and 7, the difference margin was approximately ±2%. We obtained 979 gene families meeting these criteria. We also searched for invertebrate homologues (typically *Drosophila melanogaster* and *Caenorhabditis elegans*) by BLAST-P analysis. We identified by phylogenetic analysis the potential *D. melanogaster* orthologs of approximately 41% of gene families assessed. *C. elegans* homologs were found in approximately the same amount of assessed gene families, but these sequences were not used for estimating the age distribution of gene families because the divergence time between nematode and vertebrate is unclear.

Inference of duplication events. Based on the inferred phylogeny, we identified subfamilies in each gene family, along with the corresponding duplication events (Fig. 2). Vertebrate genes in the same subfamily of the human gene (boot-strapping value >60%) were considered orthologous. To minimize the error rate, we excluded gene families when the gene tree of the subfamily was sharply in conflict with the uncontested vertebrate phylogeny. We believe that the effect of orthologous misidentification has been reduced to a minimum.

In the final data set with 749 gene families, we identified 1,739 duplication events. Among the 2,488 orthologous subfamilies, 100% included human and at least one other mammal (91% rodents) sequences, 35% had bird sequences, 31% frog, 28% teleost and 41% *D. melanogaster*.

Estimation of molecule time scale. In the global clock method[26], a linearized neighbor-joining tree[29] is used to convert the (average) Poisson distance of protein sequences to the molecular time scale. We adopted primate–rodent (80 Myr ago), mammal–bird (310 Myr ago), mammal–amphibian (350 Myr ago), tetrapod–teleost (430 Myr ago) and vertebrate–*Drosophila* splits (830 Myr ago)[16] for the calibrations.

Applying the nearest-neighbor clock involved two steps. First, we determined the phylogenetic interval of a duplicate event by the closest speciation(s). '(Teleost–tetrapod, *Drosophila*–vertebrate)', for example, meant that a duplication event occurred before the teleost–tetrapod split but after the *Drosophila*–vertebrate split. This information can be one sided, for example '(teleost-tetrapod, ?)'; where there are insufficient sequence data. The phylogenetic interval is only based on the topology. Thus, as long as the phylogeny is largely correct, it is robust with respect to non-constant evolutionary rate caused by differential selective constraints in genes and lineages. Second, we estimated the age of the duplication event from the (fossil record) calibrations of the phylogenetic interval, taking an average if there were multiple fossil-record data points (Fig. 2).

Statistical error and sampling bias. Gene families selected for our study were required to meet some criteria for the reliable estimation of duplication time. As described above, the selection procedure was applied carefully to minimize any potential sampling bias. We examined gene families that were *not* selected and found no evidence for systematic bias against either ancient or recent gene duplications. For example, for gene families that were excluded in the third round, the ratio of paralogous to orthologous (human/rodent) distances was randomly distributed (data not shown).

Some gene families have dramatically expanded in primates[6,7,15,17] by tandem or segmental duplications, which may be subject to gene conversion or recombination[18]. Dating these very recent duplication events, many occurring less than 10 Myr ago, could be problematic, although this effect would be considerably reduced for ancient gene duplications.

Consistent with previous results[11], the uncertainty of some vertebrate fossil records had a marginal effect on the time estimation of duplicate events. The overall difference between the global and nearest-neighbor clocks was small (Fig. 4), indicating that the phylogenies of gene families were generally consistent with the estimates of gene age. Using other measures of distance (for example PAM (Accepted Point Mutation) or (gamma) distance)[26–28] resulted in age distributions of the same shape with a small shift (<5%) in time scale.

We did not use synonymous distance (the S-clock), for two reasons. First, the S-clock is highly unreliable for ancient gene duplications because of fast saturation when $S > 1$ (refs 23,30). Second, the S-clock[23] is usually used with the assumption of a constant synonymous rate across genes, neglecting variations as large as tenfold among genes.

After gene (genome) duplication, the preservation of duplicates may be gene (locus)–specific as a result of functional specificity. To test whether this would affect our results, we classified gene families into transcription factors (15%), signal transduction–related (20%), cell/development-related (35%), basic metabolic (19%) and others (11%), and found no evidence for different percentages of functional categories in the three adjacent time periods. We therefore concluded that this problem may not be very serious.

Note: Supplementary information is available on the Nature Genetics website.

Acknowledgments

We thank W.H. Li and W. Nordstrom for their comments. This study is supported by an NIH grant to X.G. X.G. is the 2001 DuPont Young Professor.

Competing interests statement

The authors declare that they have no competing interests.

Received 10 April; accepted 2 May 2002

1. Ohno, S. *Evolution by Gene Duplication* (George Allen and Unwin, London, 1970).
2. Hughes, A.L. 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 (1999).
3. Wolfe, K.H. Yesterday's polyploids and the mystery of diploidization. *Nature Rev Genet* **2**, 333–341 (2001).
4. Meyer, A. & Schartl, M. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr. Opin. Cell Biol.* **11**, 699–704 (1999).
5. Martin, A. Is tetralogy true? Lack of support for the 'one-to-four rule'. *Mol. Biol. Evol.* **18**, 89–93 (2001).

6. Lander, E.S. *et al*. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
7. Venter, J.C. *et al*. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
8. Nadeau, J.H. & Sankoff, D. Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics* **147**, 1259–1266 (1997).
9. Hughes, A.L., da Silva, J. & Friedman, R. Ancient genome duplications did not structure the human Hox-bearing chromosomes. *Genome Res.* **11**, 771–780 (2001).
10. Schughart, K., Kappen, C. & Ruddle, F.H. Duplication of large genomic regions during the evolution of vertebrate homeobox genes. *Proc. Natl Acad. Sci. USA* **86**, 7067–7071 (1989).
11. Wang, Y. & Gu, X. Evolutionary patterns of gene families generated in the early stage of vertebrates. *J. Mol. Evol.* **51**, 88–96 (2000).
12. Lundin, L.G. Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics* **16**, 1–19 (1993).
13. Lopreato, G.F. *et al*. Evolution and divergence of sodium channel genes in vertebrates. *Proc. Natl Acad. Sci. USA* **98**, 7588–7592 (2001).
14. Spring, J. Vertebrate evolution by interspecific hybridization—are we polyploid? *FEBS Lett.* **400**, 2–8 (1997).
15. Eichler, E.E. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet.* **17**, 661–669 (2001).
16. Sidow, A. Gen(om)e duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev.* **6**, 715–722 (1996).
17. Li, W.H., Gu, Z., Wang, H. & Nekrutenko, A. Evolutionary analyses of the human genome. *Nature* **409**, 847–849 (2001).
18. Nei, M., Gu, X. & Stinikova, T. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc. Natl Acad. Sci. USA*. **94**, 7799–7806 (1997).
19. McLysaght, A., Hokamp, K. & Wolfe, K.H. Extensive genomic duplication during early chordate evolution. *Nature Genet*. **31**, 200–204 (2002); advance online publication, 28 May 2002 (DOI:10.1038/ng884).
20. Fishman, M.C. Genomics. Zebrafish—the canonical vertebrate. *Science* **294**, 1290–1291 (2001).
21. Iwabe, N., Kuma, K. & Miyata, T. Evolution of gene families and relationship with organismal evolution: rapid divergence of tissue-specific genes in the early evolution of chordates. *Mol. Biol. Evol.* **13**, 483–493 (1996).
22. Nikoh, N. *et al*. An estimate of divergence time of Parazoa and Eumetazoa and that of Cephalochordata and Vertebrata by aldolase and triose phosphate isomerase clocks. *J. Mol. Evol*. **45**, 97–106 (1997).
23. Lynch, M. & Conery, J.S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).
24. Thompson, J.D., Higgins, D.G. & Gibson, T.J. 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 (1994).
25. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*. **4**, 406–425 (1987).
26. Gu, X. Early metazoan divergence was about 830 million years ago. *J. Mol. Evol.* **47**, 369–371 (1998).
27. Kumar, S. & Hedges, S.B. A molecular timescale for vertebrate evolution. *Nature* **392**, 917–920 (1998).
28. Nei, M., Xu, P. & Glazko, G. Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proc. Natl Acad. Sci. USA* **98**, 2497–2502 (2001).
29. Takezaki, N., Rzhetsky, A. & Nei, M. Phylogenetic test of the molecular clock and linearized trees. *Mol. Biol. Evol*. **12**, 823–833 (1995).
30. Long, M. *et al*. Gene duplication and evolution [Technical Comments]. *Science* **293**, 1551 (2001).