Open Access

Different evolutionary patterns between young duplicate genes in the human genome

Peng Zhang, Zhenglong Gu and Wen-Hsiung Li

Address: Department of Ecology and Evolution, University of Chicago, East 57th Street, Chicago, IL 60637, USA.

Correspondence: Wen-Hsiung Li. E-mail: whli@uchicago.edu

Published: 1 September 2003

Genome ***Biology*** 2003, **4:**R56

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/9/R56

Received: 15 May 2003
Revised: 24 June 2003
Accepted: 24 July 2003

Abstract

Background: Following gene duplication, two duplicate genes may experience relaxed functional constraints or acquire different mutations, and may also diverge in function. Whether the two copies will evolve in different patterns remains unclear, however, because previous studies have reached conflicting conclusions. In order to resolve this issue, by providing a general picture, we studied 250 independent pairs of young duplicate genes from the whole human genome.

Results: We showed that nearly 60% of the young duplicate gene pairs have evolved at the amino-acid level at significantly different rates from each other. More than 25% of these gene pairs also showed significantly different ratios of nonsynonymous to synonymous rates (K_a/K_s ratios). Moreover, duplicate pairs with different rates of amino-acid substitution also tend to differ in the K_a/K_s ratio, with the fast-evolving copy tending to have a slightly higher K_s than the slow-evolving one. Lastly, a substantial portion of fast-evolving copies have accumulated amino-acid substitutions evenly across the protein sequences, whereas most of the slow-evolving copies exhibit uneven substitution patterns.

Conclusions: Our results suggest that duplicate genes tend to evolve in different patterns following the duplication event. One copy evolves faster than the other and accumulates amino-acid substitutions evenly across the sequence, whereas the other copy evolves more slowly and accumulates amino-acid substitutions unevenly across the sequence. Such different evolutionary patterns may be largely due to different functional constraints on the two copies.

Background

Since Ohno's work [1] gene duplication is widely believed to be the major source of genetic novelties. However, how the two duplicate genes evolve after the duplication event and what the major factors are that determine the fate of duplicate genes remain poorly understood and are currently under intense research.

Lynch and Conery [2] conducted a study of several eukaryotic genomes and concluded that duplicate genes often experience relaxed functional constraints and accumulate mutations at an accelerated rate. However, since their study used within-genome data without an outgroup, it could not reveal differences in evolutionary patterns between two duplicates. Hughes and Hughes [3] used human genes as outgroups to

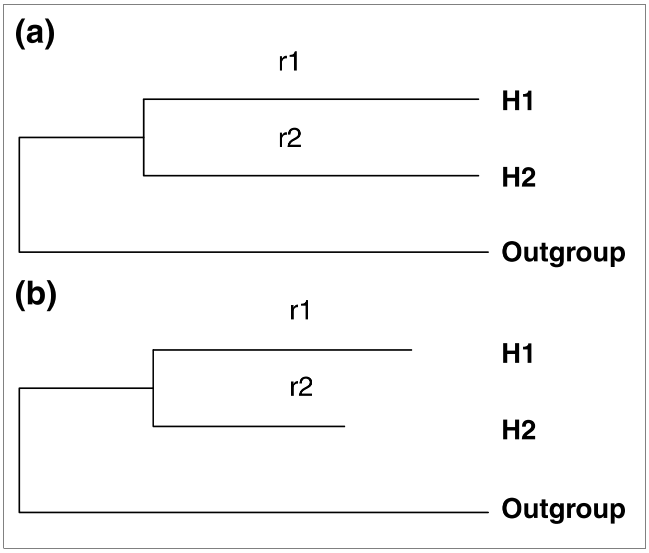

Figure 1
Two models of protein sequence evolution. By comparing the likelihood values of these two models, one can judge whether r1 = r2 (r1 and r2 are the amino-acid substitution rates on branches H1 and H2, respectively). **(a)** The model assumes r1 = r2. **(b)** The model allows r1 and r2 to be different. H1 and H2: two human duplicate genes. Outgroup: the mouse ortholog.

examine 17 pairs of duplicated frog genes, most of which have been duplicated recently. They found that the two duplicate copies of a gene had evolved at approximately the same rate. However, human genes are only distantly related to frog genes and may not be suitable outgroups for young duplicate frog genes. In contrast, Robinson-Rechavi and Laudet [4] and Van de Peer *et al.* [5], who used human or mammalian genes as outgroups to zebrafish genes, found evidence of unequal evolutionary rates between duplicate genes in zebrafish, although the percentage of pairs with such a pattern differed between the two studies. Furthermore, in a study of young duplicate genes in humans and rodents, Kondrashov *et al.* [6] found only two out of 49 duplicate genes showed different rates of evolution. Thus, the issue remains to be resolved.

We are interested in the questions of whether duplicated genes in general undergo different evolutionary patterns and what the possible causes for this could be. To address these questions, young human duplicate genes, which are defined as duplicate genes with $K_s < 0.3$, are excellent materials for several reasons. Firstly, in young duplicates the K_s (the number of synonymous substitutions per synonymous site) and K_a (the number of nonsynonymous substitutions per nonsynonymous site) values are small and can be estimated more accurately than in older duplicates. Secondly, human genes usually have no strong codon usage bias, so the K_s values are not strongly distorted by this effect. Thirdly, the mouse genome provides a suitable outgroup. With the use of outgroup sequences, the maximum likelihood method [7] can be applied, which allows the comparison of various rate models [8]. Different models for amino-acid sequence evolution, with outgroups incorporated, can be compared to judge whether the amino-acid substitution rates are the same in two duplicate copies (Figure 1). Similarly, the models for coding sequences can also be compared to judge if the K_a/K_s ratios are the same in the two copies after a gene duplication. Traditionally, K_a/K_s is taken as an index for the strength of functional constraints. Different K_a/K_s ratios usually suggest different functional constraints on two duplicate copies.

Another way to examine if two duplicate copies have experienced different functional constraints is to see whether the distribution of substitutions along their sequences are the same. If a duplicate copy is free of functional constraints, then amino-acid substitutions should occur evenly across the sequence. On the other hand, if a duplicate copy is still under considerable functional constraints, then functionally important regions should be subject to stronger constraints than functionally less important regions, and will accumulate fewer substitutions, thus yielding an uneven substitution pattern. Tang and Lewontin [9] described a statistical method to give a quantitative measure for distinguishing between even and uneven substitution patterns. The rationale of this method is that if substitutions occur evenly across a sequence, a cluster of short spaces (lengths between two consecutive substitutions) should not be extremely long. To test the significance, the longest stretch where every space is short is compared to a simulated distribution generated under the hypothesis of even substitution pattern. We combined Tang and Lewontin's [9] method and the maximum likelihood method of locating substitutions [7] to test the evenness of the substitution patterns of two duplicate genes.

Results and discussion

Amino-acid substitution rates

We examined first whether the amino-acid substitution rates in two duplicate copies are the same. We found that among the 250 pairs of young human duplicates studied, 145 pairs showed significant evidence (at the 5% level) that one copy had evolved faster than the other at the amino-acid level. Among them, 130 pairs had significantly different rates at the 1% significance level.

Hughes and Hughes [3] found similar evolutionary rates in the 17 frog duplicates they studied, probably because the human outgroup they used was too distant to make the statistical test powerful. Using fairly closely related outgroups, our results suggest that the majority of young human duplicates evolve at different rates. Our results are consistent with those of Van de Peer *et al.* [5], but the number of gene pairs with significantly unequal rates is much higher than Robinson-Rechavi and Laudet [4] (four out of 19) and Kondrashov *et al.* [6] (two out of 49) found.

Table 1

Substitution rates versus K_a/K_s ratios in duplicate genes

	Different amino-acid substitution rates*	Equal amino-acid substitution rate†	Total
Different K_a/K_s ratios‡	54	11	65
Equal K_a/K_s ratio§	91	94	185
Total	145	105	250

A 2 × 2 chi-square test. $\chi^2 = 12.78$, df = 1, $p < 0.001$. The null hypothesis is that the number of pairs with different K_a/K_s ratios is independent of the number of pairs with different amino-acid substitution rates. The values are the observed number of pairs for each category; for example, there are 54 pairs with both different K_a/K_s ratios and different amino-acid substitution rates. The amino-acid substitution rates (or the K_a/K_s ratios) in the two duplicate genes are considered different only if the difference is statistically significant. *Gene pairs with different amino-acid substitution rates between the two duplicates. †Gene pairs with equal amino-acid substitution rates between the two duplicates. ‡Gene pairs with different K_a/K_s ratios between the two duplicates. §Gene pairs with equal K_a/K_s ratios between the two duplicates.

K_a/K_s ratio

To discover if the functional constraints were the same on two duplicate genes, we examined the K_a/K_s ratio on each branch leading to the two copies. Among the 250 pairs, 65 pairs showed significantly different K_a/K_s ratios at the 5% level and 31 pairs showed a significant difference at the 1% level. As mentioned earlier, the K_a/K_s ratio is an important index of functional constraints. The smaller the K_a/K_s ratio is, the stronger the functional constraints are. Our result suggests that after gene duplication, a substantial proportion (65/250 = 26%) of the duplicate pairs have experienced different functional constraints.

Among the 65 pairs that have different K_a/K_s ratios, 54 pairs also differ between the two copies in their amino-acid substitution rates. Among the 185 pairs that showed no significant difference in K_a/K_s ratios, less than 50% showed significantly different amino-acid substitution rates. A 2 × 2 chi-square test (Table 1, $\chi^2 = 22.675$, df = 1, $p < 0.001$) reveals a significant correlation between different K_a/K_s ratios and different amino-acid substitution rates. Therefore, duplicate pairs with different K_a/K_s ratios tend to evolve at different rates, suggesting that different functional constraints might be largely responsible for the unequal evolutionary rates, although, as mentioned below, some duplicate genes have apparently undergone positive selection.

One reason why we could not detect as many pairs with different K_a/K_s ratios as pairs with different amino-acid substitution rates could be, in part, because fast-evolving sequences tend to have a higher K_s than slow-evolving ones. To see whether this was true, we calculated the K_s difference between the two copies of each pair - the K_s of the fast-evolving copy minus the K_s of the slow-evolving copy. Figure 2 shows that most of the pairs have a positive K_s difference, which means that in most of the pairs the fast-evolving copy has a higher K_s than the slow-evolving copy.

Two duplicate copies may differ significantly in the number of amino-acid substitutions, which reflects a significant difference between two K_a values at the nucleotide level. However, the two K_s values are also different (usually the copy with a higher K_a also has a higher K_s) which reduces the chance for the K_a/K_s ratios of the two copies to be significantly different. This weak correlation between K_a and K_s is consistent with several previous studies [10-13] and may be largely explained by the fact that silent sites in some genes are also under purifying selection (that is, codon usage bias) [14]. So, although we found nearly 60% of pairs with different amino-acid substitution rates, we found far fewer pairs having different K_a/K_s ratios.

We also looked for evidence of positive selection. Figure 3a shows that most of the genes have a K_a/K_s ratio of less than one, although there are still 113 genes with a K_a/K_s ratio greater than one. $K_a/K_s > 1$ suggests positive selection but evidence for positive selection requires the ratio to be significantly greater than one. In the genes with $K_a/K_s > 1$, many results are just slightly greater than one and only seven genes are found to have the K_a/K_s ratio significantly greater than one. However, this does not imply that only seven pairs of duplicate genes were subject to positive selection because, in many cases, the number of substitutions between two young

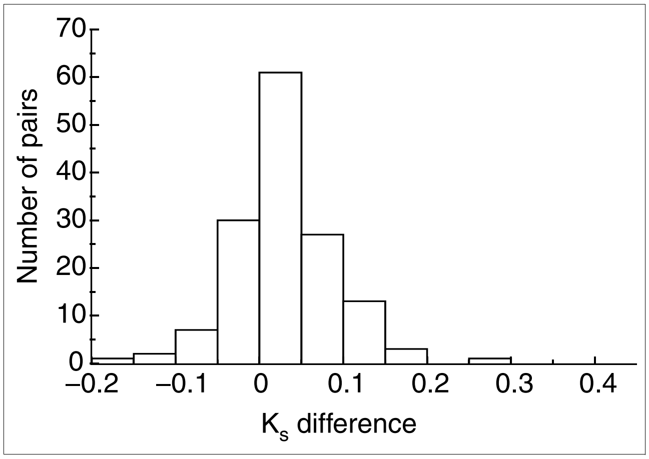

Figure 2
Distribution of K_s differences between duplicate genes for gene pairs with different amino-acid substitution rates. The *x* axis is the K_s difference between duplicate genes (the K_s of the fast-evolving copy minus the K_s of the slow-evolving one). The *y* axis is the number of gene pairs within a K_s bin. This figure shows that most pairs have a positive K_s difference, which suggests that the fast-evolving copy usually has a higher K_s than the slow-evolving copy.

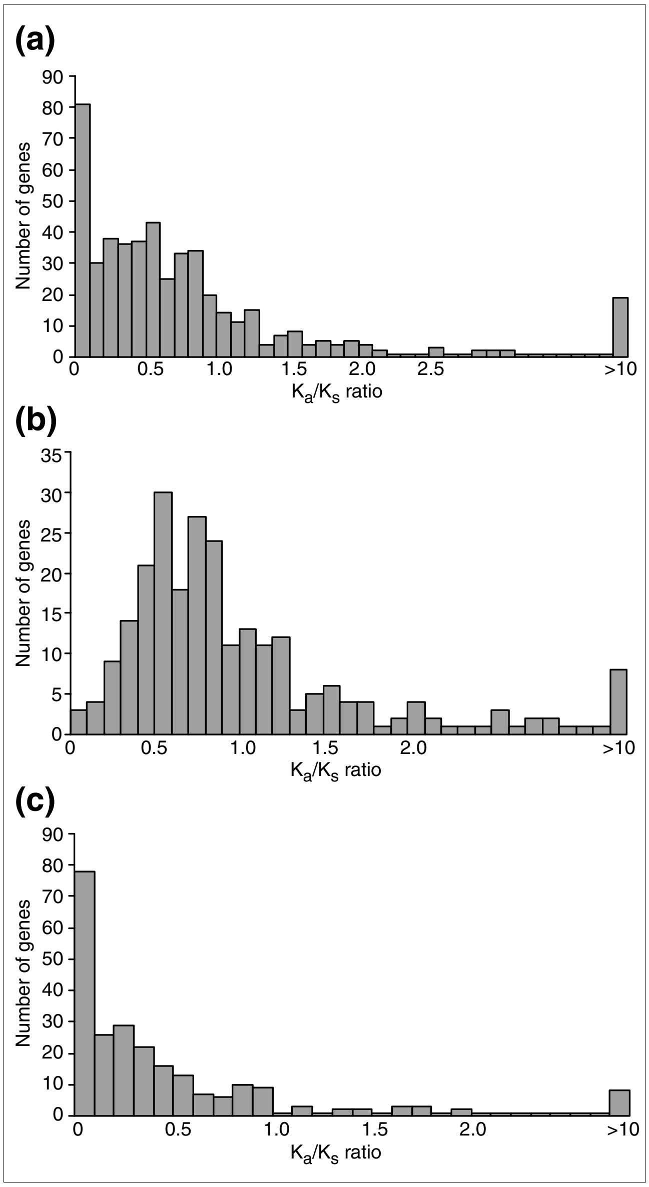

Figure 3

The K_a/K_s ratio distribution of young human duplicates. The *x* axis is the K_a/K_s ratio on the branch leading to one human duplicated gene. The *y* axis is the number of genes within a K_a/K_s bin. **(a)** All genes from the 250 pairs, a total of 500 sequences. **(b)** The fast-evolving duplicate copies of 250 pairs, a total of 250 sequences. **(c)** The slow-evolving duplicate copies of 250 pairs, a total of 250 sequences.

duplicates may be too small for the test to be statistically significant, even if some of the substitutions have occurred by positive selection.

Most of the fast-evolving duplicate copies have higher K_a/K_s ratios (Figure 3b) than slow-evolving duplicate copies (Figure 3c). This supports the view that after gene duplication, one duplicate copy may have undergone purifying selection, while the functional constraints on the other copy may have been relaxed to some extent.

Different substitution patterns

Among the 145 fast-evolving human young duplicates, 109 have an even amino-acid substitution pattern across the sequence between the human and mouse orthologs. In other words, these 109 sequences show no large highly-conserved regions. On the other hand, 65 of the 145 slow-evolving copies show evidence of an uneven substitution pattern between human and mouse orthologs, which suggests that they have some slow-evolving regions and some fast-evolving regions at the protein level.

In order to infer the position of each amino-acid substitution in the sequence, we inferred the ancestral sequences by using PAML (Phylogenetic Analysis by Maximum Likelihood) [7] standard settings, which assume constant rates across sites. It is possible, therefore, that our estimated substitutions may be more evenly distributed than they actually are. However, because we are comparing the percentage of sequences with even patterns in fast-evolving copies to those in slow-evolving copies, this potential bias should be on both sides of the comparison and should not change our conclusion.

Figure 4 with a chi-square test ($\chi^2 = 12.78$, df = 1, $p < 0.01$) shows that fast-evolving duplicates have a significantly higher proportion of sequences with an even substitution pattern. This finding suggests that most of the fast-evolving copies have more relaxed functional constraints than slow-evolving copies and tend to accumulate substitutions evenly across the sequence. The suggestion of relaxed functional constraints for young duplicates is consistent with the observation of Lynch and Conery [2]. Of course, we cannot exclude the possibility that some of the amino-acid substitutions in fast-evolving copies might have been due to positive Darwinian selection.

Our finding is very different from that of Kondrashov *et al.* [6], who found only two pairs with unequal evolutionary rates out of 49 pairs studied in mammals. Since they also focused on young duplications ($0.05 < K_s < 0.5$) and the approach they used to identify duplicate genes is similar to ours, this may be due to the different datasets used. Since the neutral pattern found in the fast-evolving copies in our study is to some extent similar to the evolution of pseudogenes, we examined the possibility of the inclusion of many pseudogenes in our sample.

The gene predictions in the Ensembl database [15] we used always produce a translation for each gene and a stringent criterion (near full-length similarity) was used in our grouping method; consequently, our dataset does not include pseudogenes due to premature codons. Since we limited our set of duplicated genes to $K_s > 0.05$, a pseudogene in our sample would be likely to have lost its function only very recently, otherwise it would have gained one or more premature stop codons since the time of nonfunctionalization. In the Ensembl database we used only those genes ('known' genes)

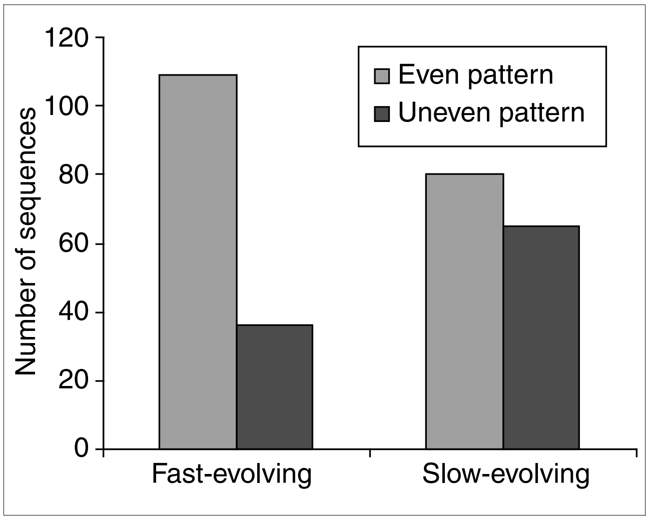

Figure 4
Comparison between fast-evolving copies and slow-evolving copies. The figure shows that fast-evolving copies have more cases with substitutions distributed evenly along the sequence than slow-evolving copies. Fast-evolving: the copy that has evolved faster than the other in each duplicate pair. Slow-evolving: the copy that has evolved slower than the other in each duplicate pair. Even pattern: a sequence that has evenly-distributed substitutions along the sequence. Uneven pattern: a sequence that has unevenly-distributed substitutions.

with experimental support and those genes ('novel' genes) with high similarity to known genes in human and other organisms. Genes purely from Genscan predictions were not used in this analysis. These approaches would have effectively reduced the portion of pseudogenes in our dataset. If the functional constraints on a gene are largely relaxed, the evolutionary pattern of this gene may be similar to that of pseudogenes. So it is possible that some of the fast-evolving genes may be on their way to become pseudogenes, although it is still possible that they may evolve new functions. Kondrashov *et al.* [6] used a cDNA-based dataset and found only a few duplicated pairs with different evolutionary rates, which may have represented those genes that survived well through selection and were still functioning. In other words, the cDNA-based genes which they used are normally expressed, meaning these genes may still be under strong selection pressure. Our dataset might be more appropriate for providing a general scenario of how two duplicate genes evolve after gene duplication.

Conclusion

We used conservative criteria to select young human duplicate pairs and applied a stringent statistical method to test whether two duplicate copies exhibit different evolutionary patterns. Our results suggest that, in most cases, during the early stage of evolution following gene duplication, the two duplicates evolve at different rates, which could affect the fate of the two copies. Different functional constraints on the two copies may have been largely responsible for the different rates. One copy may have relaxed functional constraints, while the other could still be under strong constraints. The stringent statistical tests used in this study might have underestimated the proportion of pairs with this pattern, but this could only strengthen our argument.

Materials and methods

Processing data and selecting independent young human duplicate genes

Human genes were downloaded from the Ensembl human database [15] version 11.31 (28 February 2003). The original dataset is available from the authors on request. Only known and novel genes were used in this analysis; those sequences containing repetitive elements detected by FASTA ($E = 10^{-5}$) searching against Repbase [16-18] were removed from the dataset. If two genes overlapped at a chromosomal position, the gene with the longer protein was retained. The protein sequences selected were grouped into families by the method used by Gu *et al.* [19]. Within each gene family, the selection of independent duplicate gene pairs proceeded with increasing K_s. That is, within each gene family, we selected the gene pair with the smallest K_s and excluded it from the family and then selected the gene pair with the smallest K_s from among the remaining genes. We repeated this until no gene pairs could be selected. This method ensured that a gene in one pair would not appear in another pair. Among the duplicate genes we selected, those pairs with $0.05 < K_s < 0.3$ were used in this study. We used $K_s < 0.3$ as a cutoff to define young human duplicates. With the K_s between human duplicate genes less than 0.3, the duplication would have occurred less than 50 million years ago. Pairs with $K_s < 0.05$ have too few substitutions to make a statistical test meaningful. For each pair we selected, both copies were used to search the mouse database. Mouse genes were also obtained from the Ensembl database [15] version 11.3 (28 February 2003) and were cleaned using the same procedure that was used to clean the human database. The pairs in which the two copies had the same best hits and a human-mouse $K_s < 1$ were kept for this study. We chose $K_s < 1$ as a cutoff point because a distant outgroup makes it harder to detect rate differences. A set of 250 young human duplicate pairs and their mouse orthologs were thus retained.

Statistical methods to compare evolutionary patterns between two copies

To calculate if the evolutionary rates and the K_a/K_s ratio are the same between the two duplicate copies, the likelihood-ratio test [8] was applied to each pair selected. To test the hypothesis of equal evolutionary rates between the two duplicate copies at the amino-acid level, a two-rate model and a free-rate model were compared. The two models differ in that the two-rate model assumes the same evolutionary rate on the two branches leading to the two duplicates but allows the rate on the outgroup branch to be different, while the free-rate model does not impose any equal rates among branches.

The codeml program (set seqtype = 2 for amino-acid sequences) in the PAML package was run for each of the two models with all parameters set to default except for the parameter 'model' for amino-acid substitution, which was set to 'Poisson'. We also set this parameter to the 'Jones-Taylor-Thornton model' and the conclusion was basically the same. Two maximum likelihood values for the two models were given, and twice the difference was compared to a chi-square distribution. If significant, the results suggest that the two branches have evolved at unequal rates. To test if the K_a/K_s ratios are different between the coding sequences of the two duplicates, a two-ratio model, which assumes the same K_a/K_s ratio on the branches leading to the two duplicates but an independent K_a/K_s ratio on the branch leading to the outgroup, was compared to the free-ratio model, which assumes an independent K_a/K_s ratio for each branch.

Substitution patterns and differential selection

Using the PAML package [7], the ancestral sequence of each human duplicate gene pair was reconstructed and the position of each substitution was located. Tang and Lewontin's [9] method was then applied to calculate the T statistics of each human sequence. For each human sequence, 100,000 pseudo-sequences were generated under the null hypothesis that the substitutions distribute evenly across the sequences, and the distribution of T statistics of these 100,000 pseudo-sequences was generated. The T statistic of the real sequence was then compared to this distribution. If the T statistic was extremely large or small (the rejection level is 0.05), we considered it to be an even distribution. The program was written in Perl and is available upon request.

Additional data files

The amino-acid alignments (Additional data file 1) and coding sequence alignments (Additional data file 2) are available in PAML format with the online version of this article.

Acknowledgements

This study was supported by NIH grants.

References

1. Ohno S: Evolution by Gene Duplication. Berlin: Springer-Verlag: 1970.
2. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes.** *Science* 2000, **290:**1151-1155.
3. Hughes MK, Hughes AL: **Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*.** *Mol Biol Evol* 1993, **10:**1360-1369.
4. Robinson-Rechavi M, Laudet V: **Evolutionary rates of duplicate genes in fish and mammals.** *Mol Biol Evol* 2001, **18:**681-683.
5. Van de Peer Y, Taylor JS, Braasch I, Meyer A: **The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes.** *J Mol Evol* 2001, **53:**436-446.
6. Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV: **Selection in the evolution of gene duplications.** *Genome Biol* 2002, **3:**research0008.1-0008.9.
7. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *CABIOS* 1997, **13:**555-556.
8. Yang Z, Nielsen R: **Synonymous and nonsynonymous rate variation in nuclear genes of mammals.** *J Mol Evol* 1998, **46:**409-418.
9. Tang H, Lewontin RC: **Locating regions of differential variability in DNA and protein sequences.** *Genetics* 1999, **153:**485-495.
10. Graur D: **Amino acid composition and the evolutionary rates of protein-coding genes.** *J Mol Evol* 1985, **22:**53-62.
11. Wolfe KH, Sharp PM: **Mammalian gene evolution: nucleotide sequence divergence between mouse and rat.** *J Mol Evol* 1993, **37:**441-456.
12. Mouchiroud D, Gautier C, Bernardi G: **Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions.** *J Mol Evol* 1995, **40:**107-113.
13. Ohta T, Ina Y: **Variation in synonymous substitution rates among mammalian genes and the correlation between synonymous and nonsynonymous divergence.** *J Mol Evol* 1995, **41:**717-720.
14. Hellmann I, Zollner S, Enard W, Ebersberger I, Nickel B, Paabo S: **Selection on human genes as revealed by comparisons to chimpanzee cDNA.** *Genome Res* 2003, **13:**831-7.
15. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, *et al.*: **The Ensembl genome database project.** *Nucleic Acids Res* 2002, **30:**38-41.
16. Jurka J: **Repbase Update: a database and an electronic journal of repetitive elements.** *Trends Genet* 2000, **16:**418-420.
17. Jurka J: **Repeats in genomic DNA: mining and meaning.** *Curr Opin Struct Biol* 1998, **8:**333-337.
18. **Repbase** [http://www.girinst.org/Repbase_Update.html]
19. Gu Z, Cavalcanti A, Chen FC, Bouman P, Li WH: **Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast.** *Mol Biol Evol* 2002, **19:**256-262.