Published online July 14, 2004

3724–3733 *Nucleic Acids Research, 2004, Vol. 32, No. 12*
doi:10.1093/nar/gkh686

Prevalence of intron gain over intron loss in the evolution of paralogous gene families

Vladimir N. Babenko, Igor B. Rogozin, Sergei L. Mekhedov and Eugene V. Koonin*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bldg 38A, Bethesda, MD 20894, USA

Received May 12, 2004; Revised June 5, 2004; Accepted June 16, 2004

ABSTRACT

The mechanisms and evolutionary dynamics of intron insertion and loss in eukaryotic genes remain poorly understood. Reconstruction of parsimonious scenarios of gene structure evolution in paralogous gene families in animals and plants revealed numerous gains and losses of introns. In all analyzed lineages, the number of acquired new introns was substantially greater than the number of lost ancestral introns. This trend held even for lineages in which vertical evolution of genes involved more intron losses than gains, suggesting that gene duplication boosts intron insertion. However, dating gene duplications and the associated intron gains and losses based on the molecular clock assumption showed that very few, if any, introns were gained during the last ~100 million years of animal and plant evolution, in agreement with previous conclusions reached through analysis of orthologous gene sets. These results are generally compatible with the emerging notion of intensive insertion and loss of introns during transitional epochs in contrast to the relative quiet of the intervening evolutionary spans.

INTRODUCTION

Eukaryotic protein-coding genes typically contain multiple introns that are spliced out of the pre-mRNA by a distinct, large RNA–protein complex, the spliceosome, which is conserved throughout the eukaryotic world (1–3). The conservation of the protein machinery involved in splicing suggests that introns invaded eukaryotic genes at an early stage of evolution, perhaps concomitantly with the origin of eukaryotes (4). Indeed, systematic comparisons of orthologous eukaryotic genes indicated that many intron positions are conserved over extremely long evolutionary spans (5,6). Specifically, up to 25% of introns in highly conserved eukaryotic genes appear to be inherited from the common ancestor of plants and animals, and some intron positions are shared even by crown group eukaryotes and protists (6). However, the same studies also indicated that extensive loss of ancient introns and insertion of new ones occurred during eukaryotic evolution. A parsimonious reconstruction of the evolutionary scenario for introns in highly conserved eukaryotic genes revealed a highly non-uniform distribution of intron gains and losses among the branches of the eukaryotic phylogenetic tree (6). There were many more intron gains than losses, e.g. in the chordate and plant lineages, whereas, in fungi and arthropods, losses prevailed over gains. In a striking contrast, comparative analysis of orthologous genes among vertebrates revealed no evidence of intron gain and a small number of losses (7). Combining the results of these studies, it appears that massive intron gain and loss might accompany major evolutionary transitions, whereas little if any intron gain and the limited amount of intron loss occur in the intermediate epochs. In other words, the distribution of intron gains and losses appears to be strongly non-uniform not only between evolutionary lineages but also over the history of a particular lineage.

The recent large-scale studies on intron evolution outlined above involved comparative analysis of orthologous genes in different eukaryotic lineages. Orthologs are evolutionary counterparts that derive from a single ancestral gene in the common ancestor of the compared species (8–10). It is common for gene duplications in one or both of the compared lineages to occur subsequent to speciation, in which case the orthologous relationship connects the resulting sets of paralogous (i.e. related via duplication) genes rather than individual genes (11,12). Such sets of paralogous genes derived from a single gene in the common ancestor's genome have been designated as co-orthologous gene sets (10), whereas the paralogs themselves are said to comprise a lineage-specific expansion (LSEs) (13,14).

The studies on evolution of the exon–intron structure of eukaryotic genes cited above analyzed only one-to-one relationships between genes either by ignoring co-orthologous gene sets or by identifying and analyzing the most conserved members of such sets. However, the dynamics of intron gain and loss in LSEs of paralogous genes seems to be of special interest. Gene duplication with subsequent divergence is the principal mechanism of emergence of new genes during evolution (15–18). Paralogs comprise substantial fractions of genes in all genomes, but are particularly prominent in multicellular eukaryotes, where the majority of genes have at least one paralog (12,19,20). Gene duplications occurred throughout the evolution of life: some duplications predate the last universal common ancestor of all modern life forms, whereas many are relatively recent (17). These relatively young duplications form LSEs, which appear to be one of the major forms of adaptation, especially in eukaryotes (14,20). During evolution

*To whom correspondence should be addressed: Tel: +1 301 435 5913; Fax: +1 301 435 7794; Email: koonin@ncbi.nlm.nih.gov

of LSEs, genes undergo functional diversification, which tends to be accompanied by the acceleration of sequence evolution. A question of major interest is whether or not this mode of evolution also involves accelerated change in gene structure, i.e. extensive gain and/or loss of introns. This seems to be particularly pertinent given the apparent intensification of intron gain and loss during evolutionary transitions, which could be thought of as analogous to phases of functional diversification in LSEs. From a mechanistic viewpoint, different mechanisms of gene duplication have been described, with some duplications apparently emerging via reverse transcription of the respective mRNAs (21). Such duplicates obviously start off as intronless genes, and it would be particularly interesting to detect intron insertions into these genes. Furthermore, some of the LSEs include genes evolved via very recent duplications, and comparison of the structures of such genes on genome scale could potentially reveal very young introns and consequently shed light on the mechanism(s) of intron insertion, which so far remains elusive (22).

Evolution of the exon–intron structure of paralogous gene families has not been extensively studied on genomic scale. However, several anecdotal studies revealed considerable variability of intron positions among paralogs, along with conservation of some ancient introns (23–26), and some evidence for possible recent gains was presented (26,27).

We took advantage of the recently developed database of orthologous clusters of eukaryotic genes (KOGs), which also includes LSEs (20,28), in an attempt to investigate the dynamics of intron gain and loss in evolving paralogous gene families on genome scale. By analyzing parsimonious scenarios for numerous LSEs from plants and animals, we found that intron gains significantly outnumbered intron losses over the evolutionary time span of several hundred million years separating the analyzed lineages. By calibrating evolutionary trees of LSEs against the known times of divergence from outgroups, we determined, in agreement with the results of previous studies on the evolution of introns in orthologous gene sets, that there were no or very few intron gains during the last ~100 million years (Myr) of animal and plant evolution.

MATERIALS AND METHODS

Data

Clusters of orthologous eukaryotic genes that were represented in at least three of six eukaryotic species with completely sequenced genomes, namely, humans, the nematode *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster*, the two yeasts, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, and the green plant *Arabidopsis thaliana* were extracted from the KOGs database (28) (http://www.ncbi.nlm.nih.gov/COG/new/). All KOGs that included at least three species were analyzed. Multiple alignments of protein sequences were constructed for each KOG using the ClustalW program with default parameters (29). Nucleotide sequence alignments were derived from protein sequence alignments for the analyzed KOGs by using an *ad hoc* program, and intron locations were mapped on the respective nucleotide sequence alignments (6). All introns that did not contain the canonical splice junction dinucleotides at their termini (GT–AG, GC–AG, AT–AC) were removed because non-canonical junctions are the main indicator of errors in intron position identification. For a pair of introns to be considered homologous, they were required to occur in exactly the same position in the aligned sequences of KOG members. Given the inherent problems in the annotation of gene structure and difficulties in aligning poorly conserved regions of protein sequences, we employed the following approach to the analysis of evolutionary conservation of intron positions: all positions containing a deletion or insertion in at least one sequence (i.e. a gap in the alignment) were removed from protein sequence alignments together with adjacent positions. The analysis was repeated with varying numbers of removed positions (from 0 to 5 on each side of the gap) in order to assess the effect of this alignment pruning on the obtained results.

Delineation of LSEs of paralogous genes

The computational pipeline for the identification and analysis of LSEs is schematically shown in Figure 1A. The distance matrices were computed using the PROTDIST program of the PHYLIP package, with the Dayhoff PAM matrix employed to score the distance between a pair of proteins (30). Neighbor-joining trees were constructed using the NEIGHBOR program of the PHYLIP package (30). The bootstrap support for internal branches was calculated from 1000 pseudoreplicates using the SEQBOOT and CONSENSE programs (30). The LSEs were defined as monophyletic sets of paralogous genes that evolved via duplication(s) subsequent to the latest speciation event, which involved the respective lineage. In practical terms, the LSEs were identified as monophyletic clusters of genes from the same species in the neighbor-joining trees (Figure 1B). To ensure reliable identification of the LSE without losing large amounts of data, only LSEs with bootstrap support >70% on each internal branch, including the one that leads from the outgroup to the LSE were accepted for further analysis. The trees were treated as rooted, with the root placed between the LSE and the closest outgroup (Figure 1B).

Reconstruction of intron gain and loss

For the reconstruction of scenarios of intron gain and loss in LSEs, intron positions were represented as a data matrix of intron absence/presence (encoded as 0/1; Figure 1B). The matrices of intron absence/presence along with the rooted neighbor-joining tree of the given LSE were used as the input data for the DOLLOP program of the PHYLIP package (30). This program employs the Dollo parsimony approach, which is based on the assumption that each derived character state (in this case, intron presence) originated only once on the tree (31). The states of intron presence–absence in internal nodes, including the progenitor of the LSE (Figure 1B) as well as the number of intron gains and losses for each branch within the LSE, were derived from the DOLLOP output using an *ad hoc* program. The alignments, matrices of intron presence–absence and phylogenetic trees for all LSEs analyzed in this work are available at ftp://ftp.ncbi.nih.gov/pub/koonin/intron_evolution/LSEs/.

Dating gain and loss of introns

The approximate dates of intron gains and losses were determined under the molecular clock assumption using the

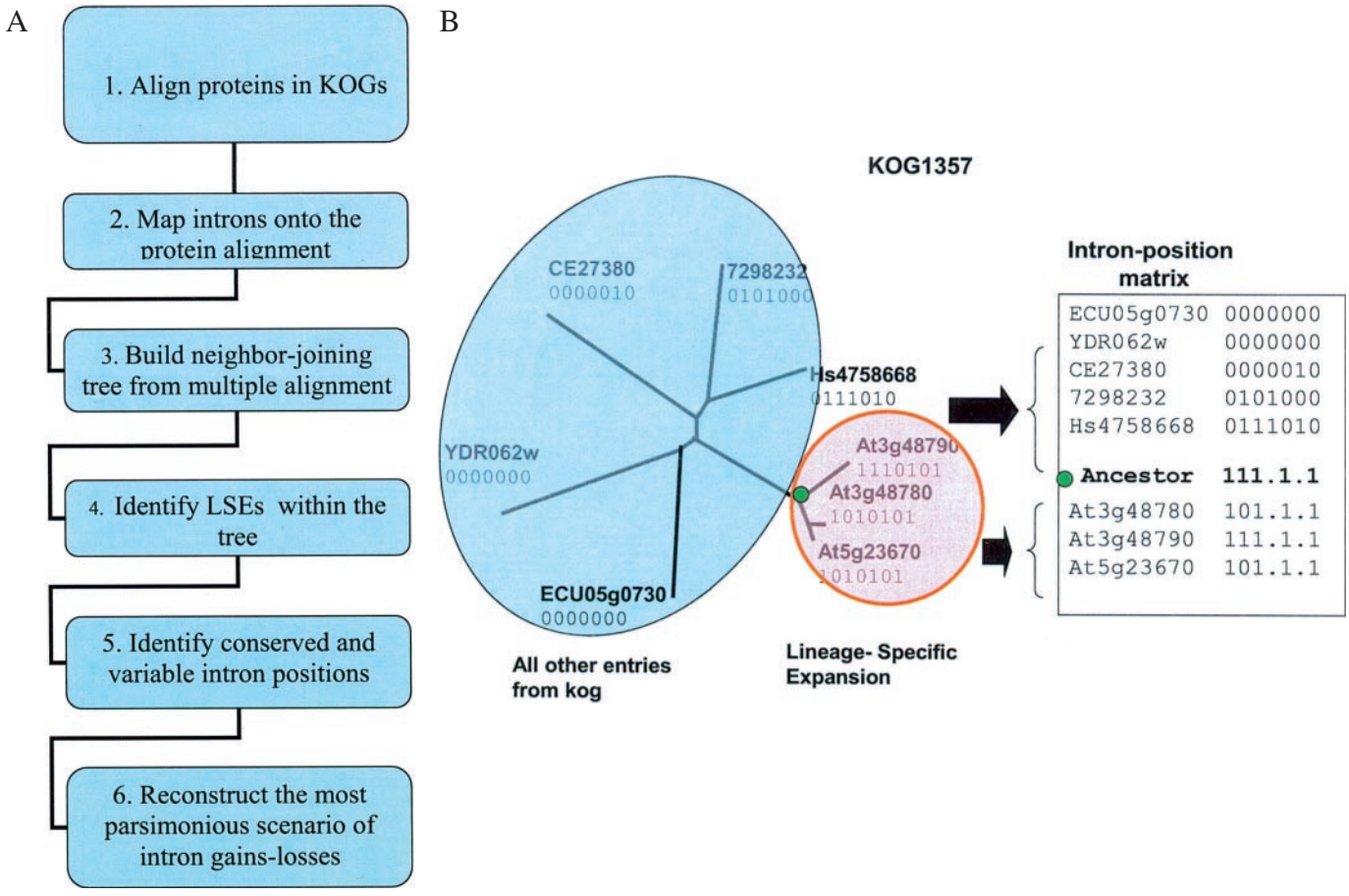


Figure 1. Computational strategy for the analysis of gene structure evolution of LSEs of paralogous genes. (**A**) Flow chart of the procedure. (**B**) Identification of an LSE, construction of the matrix of intron presence (1) and absence (0), and reconstruction of the gene structure of the last common ancestor of the LSE. The procedure is shown with a specific example, KOG1357 (serine palmitoyltransferase). Intron positions that contained introns in some members of the given KOG but not in the LSE, including its inferred ancestor, are denoted by dots. These positions were not part of the analysis of LSE evolution.

neighbor-joining trees of the LSEs. Specifically, the age of an event was calculated as follows:

$$t_e = T\frac{L_t - L_e}{L_t}$$

where

$$L_t = L_0 + L_1$$

and

$$L_e = L_0 + \left(\frac{L_1}{L_2 + L_3}\right)\left(L_2 - \frac{L_4}{2}\right),$$

where t_e is the age of the event in question, T is the time elapsed from the divergence of the analyzed LSEs from the last common ancestor with the closest outgroup, L_e is the shortest path from the root to the event, L_t is the shortest path from the root to a leaf, and $L_0 - L_4$ are various distances in the neighbor-joining tree of the LSE (Figure 2). The normalization coefficient $[L_1/(L_2 + L_3)]$ was introduced to obtain the lower bound of the age of the event, on the premise that the slowest evolving member of the LSE (L_1) retains the original function and hence is likely to conform with the molecular clock (15,32). Since the analysis described here only allowed mapping of an event to a particular branch in the tree but not to a specific point on that branch, it was assumed that the event occurred in the middle of the corresponding time span, hence the term $[L_2 - (L_4/2)]$. The divergence times were from (33,34); in particular, the time of divergence of animals and plants from the common ancestor was accepted at ~1600 Myr, and the time of divergence of chordates from the common ancestors with arthropods and nematodes was accepted at 1000 Myr (given the uncertainties in the tree topology and time estimates, the same date was assumed for both nodes).

RESULTS

Strategy of analysis of exon–intron structure evolution in LSEs of paralogous genes

Many of the KOGs include more than one member from one or more species, and these sets of genes represent probable LSEs vis-à-vis the rest of the analyzed lineages (the KOG database also includes LSEs that have no detectable counterparts in other lineages, but these were not considered here). In order to analyze the intron dynamics associated with gene duplication, we produced approximate reconstructions of the evolutionary history of LSEs starting with their divergence from the

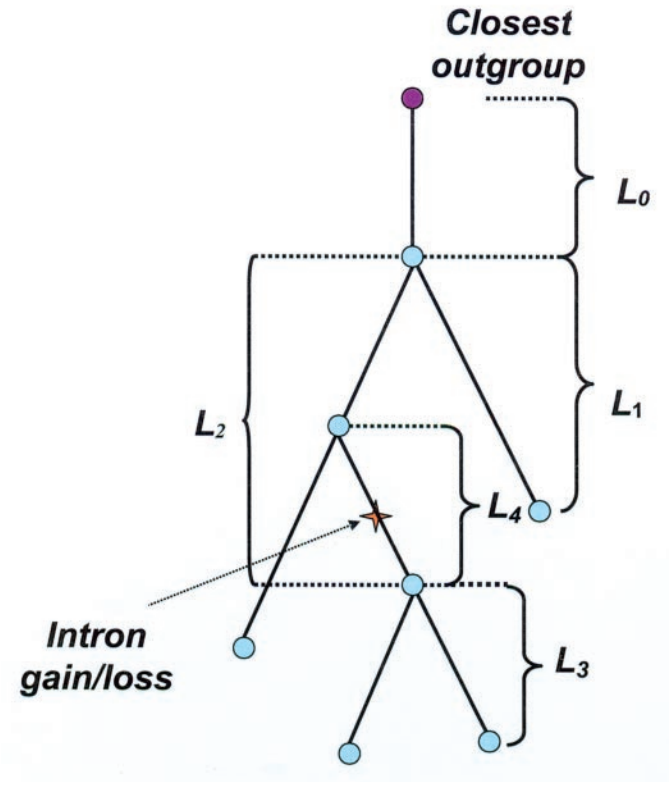


Figure 2. The procedure employed for approximate dating of intron gains and losses. The cartoon shows a neighbor-joining tree for an arbitrary LSE. For designations, see Materials and Methods.

closest identifiable outgroup (ortholog from a different lineage). For this purpose, neighbor-joining trees were constructed from the multiple alignment of each of the analyzed KOGs, and the robust LSEs were identified with respect to the closest outgroup from a different species (Figure 1A and B; for additional details, see Materials and Methods). The intron positions were mapped on the alignments of the LSE members and transformed into matrices of intron presence (1)–absence (0) [Figure 1B and (6)]. These matrices were used as the input for parsimonious reconstruction of the evolutionary scenario for the evolution of exon–intron structure in the given LSE, i.e. the most parsimonious mapping of intron gains and losses on the branches of the evolutionary tree (6). Figure 3 shows examples of parsimonious scenarios of intron gain and loss for LSEs from different species.

For this reconstruction, we applied the Dollo principle, i.e. assumed that the likelihood of independent intron gain in the same position of paralogous genes during the evolution of an LSE was negligible. Intron insertion may not be completely random as implied in the proto-splice site model (35–37). However, if proto-splice sites exist, the requirements to the insertion sequences are relatively weak and, accordingly, proto-splice sites would often occur by chance (38,39). Modeling of intron evolution under this assumption strongly suggests that the great majority of introns found in the same position in homologous genes reflect evolutionary conservation (6). Thus, although a recent study has suggested the possibility of independent insertion of introns into the same position of orthologous genes in distant lineages (40), it appears that such events are rare. Indeed, analysis of intron distribution in ancient paralogs, which appear to have accumulated introns independently, reveal only two possible cases of parallel intron insertion into the same position out of a total of 239 analyzed intron positions (<1%) (23). Thus, the disregard of possible parallel intron insertions in the present study could result in a slight overestimate of the number of losses and the corresponding underestimate of gains.

The dynamics of intron gain and loss during evolution of LSEs

Analysis of the KOGs that included more than two species yielded 1064 LSEs, for which evolutionary reconstruction was performed as depicted in Figure 1. The overall distributions of intron gains and losses inferred to have occurred during the evolution of these LSEs are summarized in Figure 4, and the breakdown of gains and losses by species are given in Table 1. The presented results are for the moderately stringent alignment filtering, with all positions containing gaps removed together with two adjacent positions from each side (see Materials and Methods); very similar results were obtained with both less and more stringent versions of the procedure (see Supplementary Material).

The most notable observations coming out of these comparisons are that (i) with the only exception of the intron-poor yeast genomes, the gene structure in LSEs appears to be dynamic, with >50% of the LSEs having gained and/or lost at least one intron and (ii) the number of intron gains in the evolution of LSEs was approximately three times greater than the number of intron losses. The dynamics of intron gain and loss was particularly pronounced in the relatively intron-poor *C.elegans* and *D.melanogaster*, where a significant majority of intron positions experienced at least one gain or loss event (Table 1). Evolution of LSEs seems to have been considerably more conservative in the intron-rich species *H.sapiens* and *A.thaliana*, with a significantly greater fraction of intron positions without gains or losses (Table 1).

The distribution of the number of intron gains in LSEs had a long tail, which included LSEs with numerous gains; in contrast, no LSE was found to have lost more than seven introns in the course of its evolution (Figure 4A). The prevalence of intron gain over intron loss becomes particularly obvious when LSEs are classified by the fraction of gains in the total number of events: in more than half of the LSEs, evolution involved only intron gains and no losses, whereas the number of LSEs with losses but no gains was approximately three times less (Figure 4B).

Dating gain and loss of introns in LSEs

One of the primary incentives in studying the evolution of introns in LSEs is the potential of detecting recent intron insertions. We compared the evolutionary distances from the closest outgroup to the common ancestor of the expansion to the distances within the expansion itself for all the analyzed LSEs (Figure 2; for details, see Materials and Methods). Using the results of these comparisons and the available estimates of the divergence times of the major eukaryotic lineages (33,41), we roughly dated the duplications and, accordingly, intron gains and losses within the LSEs under the molecular clock assumption (42,43) (for details, see Materials and Methods).

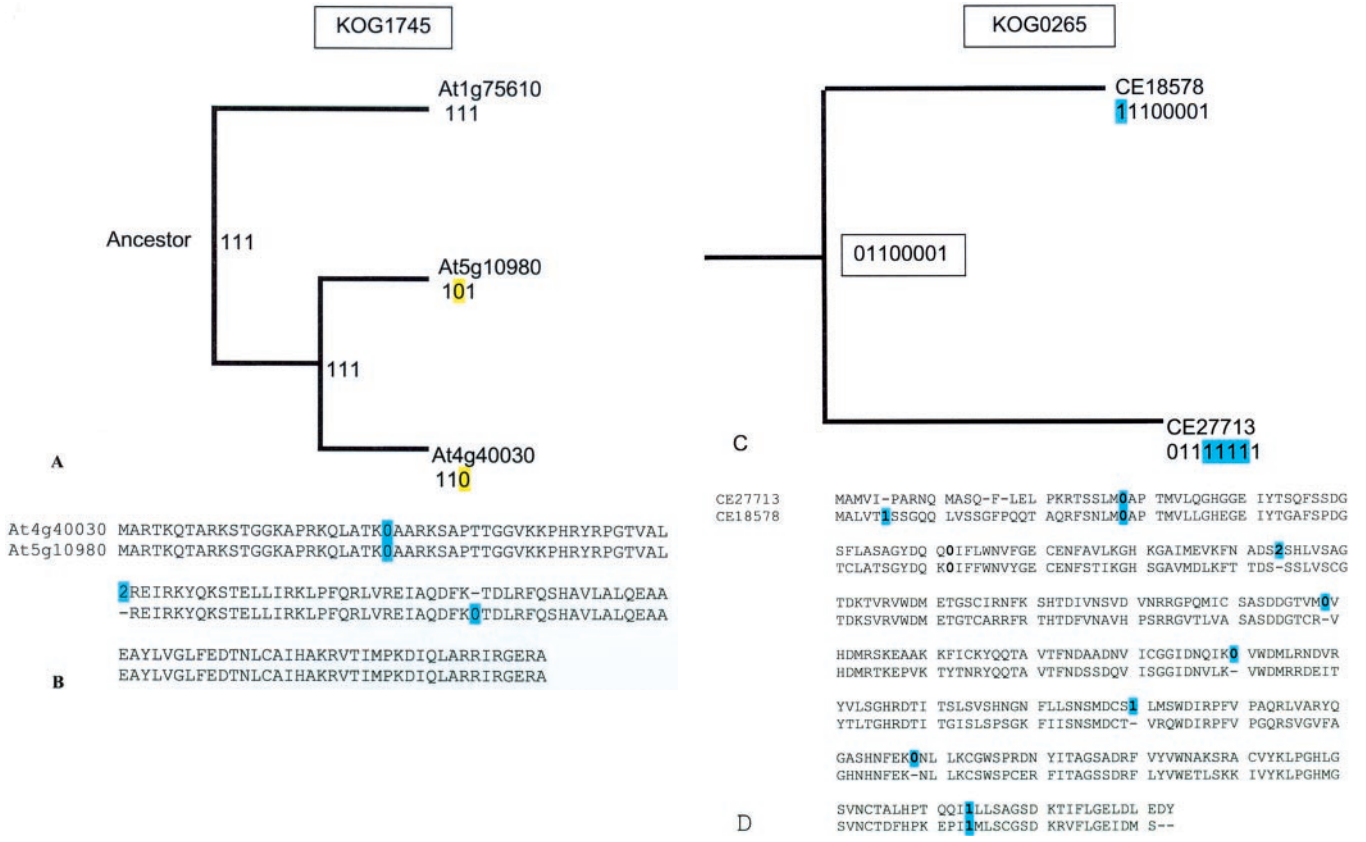


Figure 3. Examples of reconstructed evolutionary scenarios of intron gain and loss in individual LSEs. (**A**) Neighbor-joining tree for the histone H3 family of *A.thaliana*, with intron presence–absence indicated for each of the paralogs and the reconstructed ancestral forms. Intron losses are shown with yellow shading. (**B**) Protein sequence alignment of *A.thaliana* histone H3 paralogs with intron positions shaded and intron phase indicated. (**C**) Neighbor-joining tree for the U5 snRNP-specific protein-like factor family of *C.elegans* [the designations are as in (A)]. Intron gains are shaded in blue. (**D**) Protein sequence alignment of *C.elegans* U5 snRNP-specific protein-like factor paralogs with intron positions shaded and intron phase indicated.

The resulting approximate age distributions of intron gains and losses for humans and *Arabidopsis* are shown in Figure 5. There were no detectable intron gains in LSEs in the human lineage on the time scale of mammalian evolution (~100 Myr) and a limited number of intron during the last ~400 Myr of evolution (Figure 5A). The majority of intron gains in the chordate lineage appear to have occurred within the first ~500 Myr after divergence from the common ancestors of chordates and arthropods or nematodes (Figure 5A). Similar patterns were observed in *C.elegans* and *Arabidopsis* although a few 'younger' intron gains dating to ~100 Myr ago were detected (Figure 5B and C). The time distribution of intron losses was substantially different from that of intron gain. Although very few intron losses was mapped to the last ~200 Myr of mammalian and nematode evolution, there was a clear peak of losses at ~400–500 Myr in each of these lineages (Figure 5D and E). The intron losses in *Arabidopsis* were spread much more evenly over time, although no recent losses (~100 Myr) were detected (Figure 5F). Several estimates that suggest more recent divergence times (~600 Myr ago) for the major animal phyla have been published previously (44–46). Should we use these dates for estimating the timing of intron gain and loss in paralogous families, the scale in Figure 5 would compress but the conclusions on the non-uniform distribution of these events over time and the paucity of gains and losses during the last ~100 Myr would not have been affected.

DISCUSSION AND CONCLUSIONS

Recent sequencing of multiple eukaryotic genomes enabled genome-wide reconstruction of gene structure evolution, and these reconstructions yielded different results depending on the examined evolutionary scale. Comparisons of orthologous genes from distantly related eukaryotes, such as representatives of different animal phyla and different crown-group kingdoms, suggested substantial gain in introns (6). In some lineages, such as chordates or plants, many more introns seem to have been gained than lost, according to this scenario. In a stark contrast, comparative analysis of vertebrate genomes suggests that very few (if any) new introns have been gained during >500 Myr of vertebrate evolution; intron loss apparently occurred during this time span but was not extensive (7). The only way out of this conundrum seems to be the hypothesis that intron gains and losses occurred during limited time spans, perhaps coinciding with major evolutionary transitions, such as radiation of phyla and classes.

The genome-wide analyses that led to the above conclusions and conjectures were based on comparisons of orthologous genes. The mode of evolution of paralogous gene families substantially differs from that of the vertical evolution of orthologs (17,18). It is well recognized that gene duplication is a major source of functional innovation (15,47–49) and, in accord with this notion, it has been found that evolution of

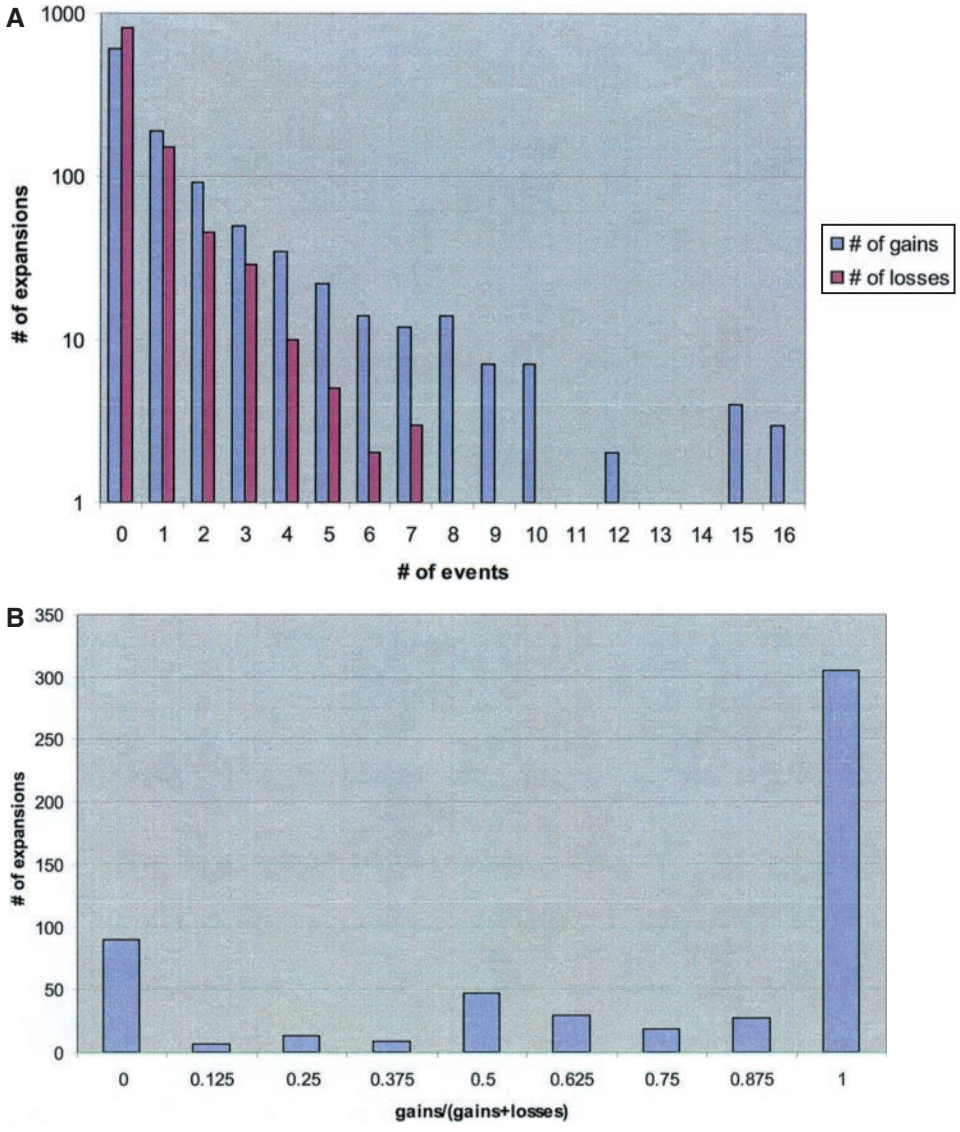
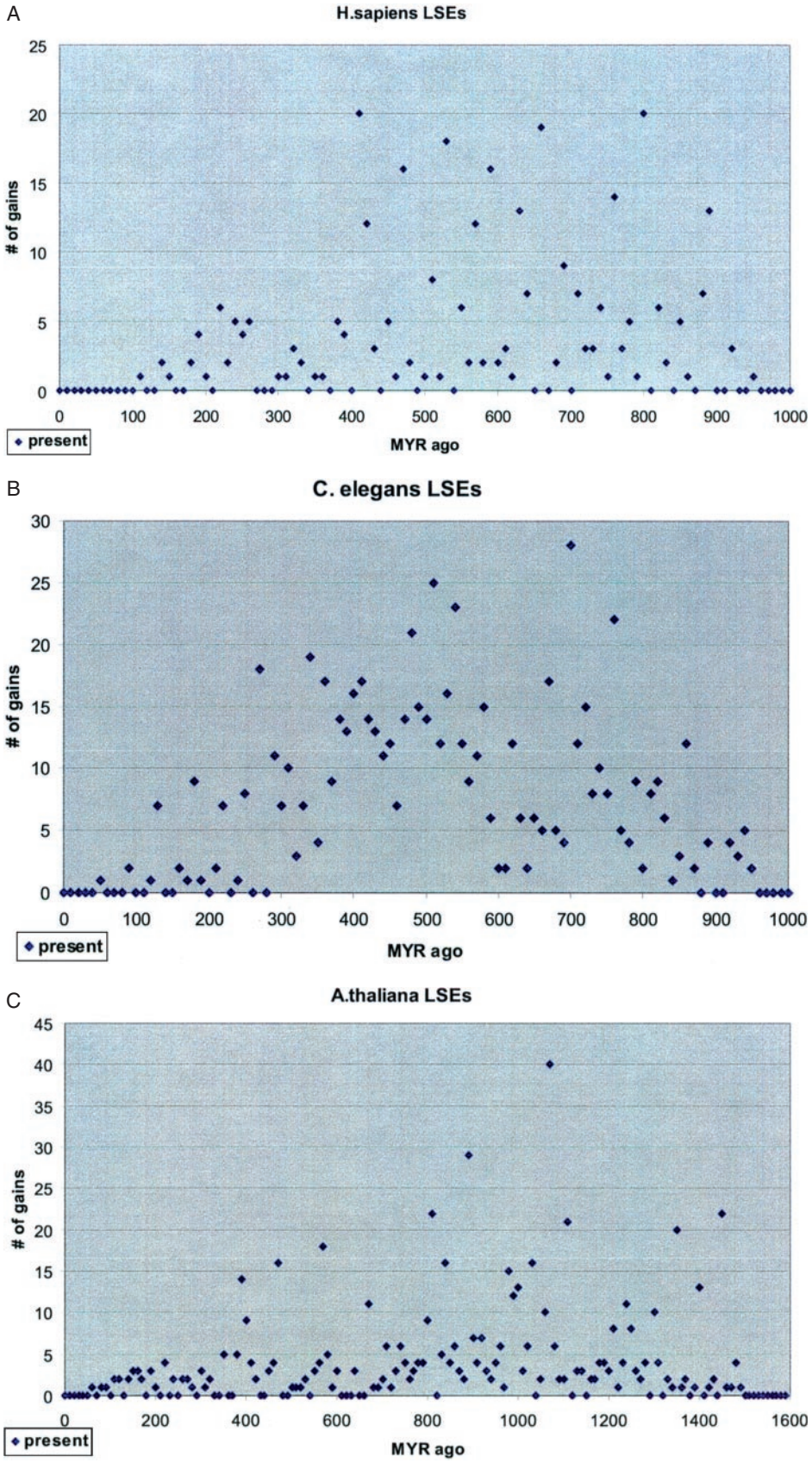


Figure 4. Distribution of the LSEs by the number of intron gains and losses. (**A**) Intron gains and losses in the LSEs. (**B**) Fraction of intron gains among the evolutionary events affecting gene structure in the LSEs. The numbers on the horizontal axis show the midpoints of the corresponding intervals.

Table 1. Intron dynamics in the evolution of LSEs of paralogs

Species	No. of genes	No. of LSEs	No. of LSEs without intron gains or losses	No. of intron positions without gain or loss	No. of intron positions with gains and/or losses	No. of gains	No. of losses
Hs	607	278	186	1035	354	274	80
Ce	508	228	49	281	674	517	157
Dm	202	93	20	66	179	127	52
At	1072	419	236	1297	557	434	139
Sp	61	29	11	22	48	34	14
Sc	34	17	15	15	2	1	1
Total	2484	1064	517	2716	1814	1387	443

paralogs significantly accelerated immediately after duplication due to weakened purifying selection (17,18). Given these differences in the modes of evolution of orthologous and paralogous genes, we were interested to determine whether or not duplications were also accompanied by accelerated gain and/or loss of introns. This question was all the more pertinent given that considerable variability of intron positions and apparent recent acquisition of introns have been reported for several paralogous gene families in animals and plants (25,27,50).



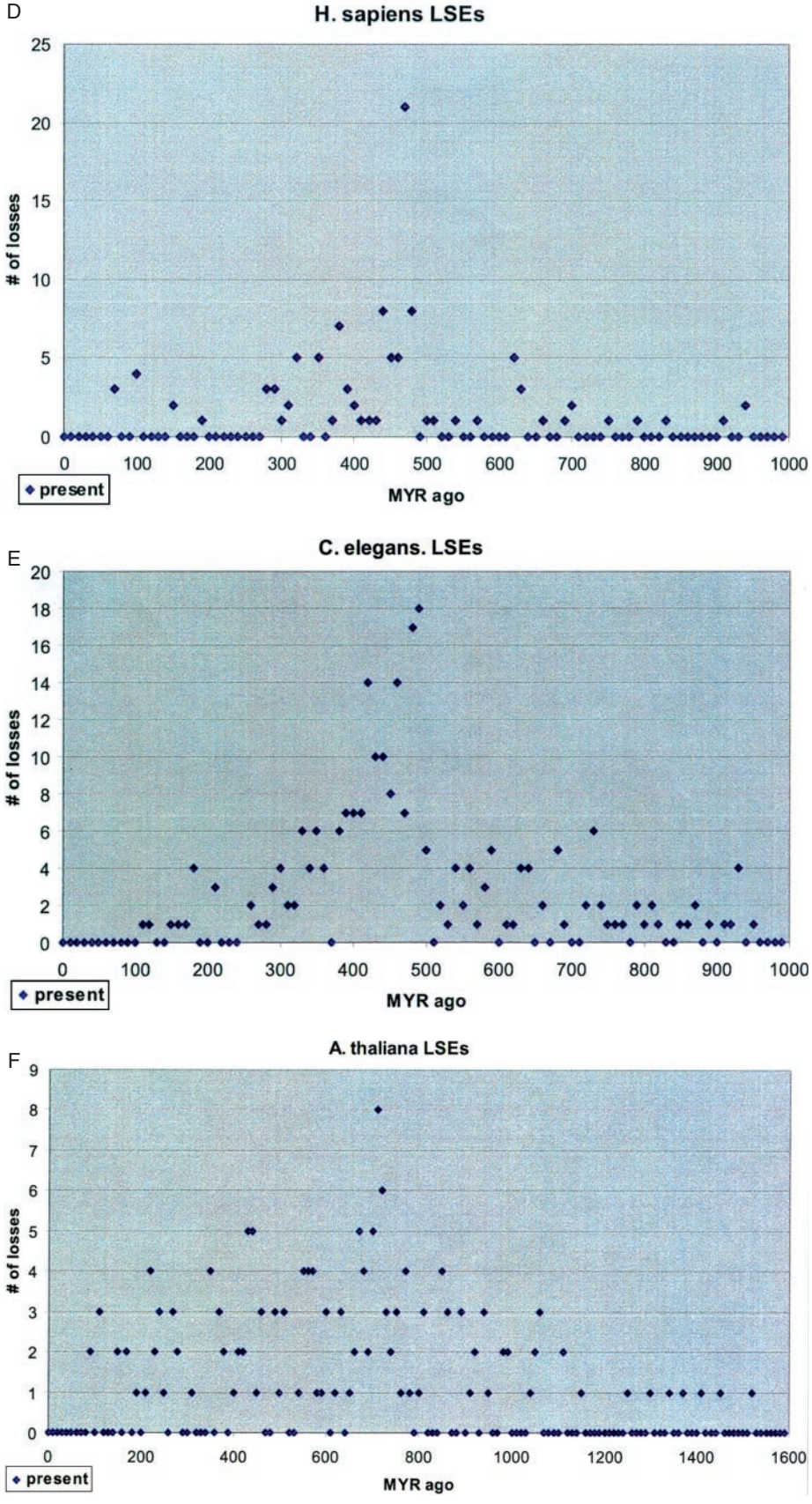


Figure 5. Estimated age distributions of intron gains and losses in LSEs. (**A**) *H.sapiens*, gains; (**B**) *C.elegans*, gains; (**C**) *A.thaliana*, gains; (**D**) *H.sapiens*, losses; (**E**) *C.elegans*, losses; and (**F**) *A.thaliana*, losses. The dots on the horizontal axis show branches of the respective estimated age with no intron gains (losses).

In the genome-wide reconstruction reported here, we observed two clear trends: (i) the evolution of the gene structure in LSEs is notably dynamic: the majority of LSEs seem to have gained or lost at least one and, typically, several introns within the time span of ~1000 Myr separating different animal phyla from their last common ancestor and (ii) the evolution of LSEs in all examined lineages involved more intron gains than losses. The latter trend of the LSE evolution held both for lineages, such as insects, that are characterized by preferential loss of introns in sets of orthologous genes, and those that apparently gained many more introns in the same genes than they have lost, such as chordates (6). This seems to be compatible with the notion that change in the selective pressure that is associated with gene duplication also triggers insertion of new introns to a greater extent than loss of ancestral ones. However, due to the sparse sampling of genomes included in this study, the LSEs analyzed here had an extremely broad age distribution, and the great majority of intron gains were associated with the ancient duplications. In particular, our present results are generally compatible with the reported (near) lack of intron gains in the chordate evolution (7) (considering that the latter work includes analysis of a limited number of orthologous genes). The distribution of intron gains and losses over time seems to have been highly non-uniform. Early evolution of animals, which involved the radiation of the phyla and classes, apparently was accompanied by the proliferation of introns after duplications as well as upon speciation, whereas, during the subsequent evolution of chordates, in spite of numerous gene duplications, few new introns emerged. The evolution of LSEs in plants (as judged from the analysis of the *Arabidopsis* genome) followed a similar pattern, with only a few more recent intron insertions. Interestingly, in animals, intron losses seemed to be distributed more non-uniformly over time than intron gains. This apparent non-uniformity is generally compatible with the notion that changes in gene structure occur preferentially during times of major evolutionary transitions. In particular, the peak of intron loss in human LSEs at ~500 Myr might be tentatively associated with the divergence of vertebrate classes. Perhaps the simplest interpretation of the possible link between bursts of intron gain and loss and divergence of major eukaryotic lineages is offered by the neutralist hypothesis of evolution of genomic complexity recently proposed by Lynch and Conery (51,52). According to this concept, transitional evolutionary epochs were associated with the population bottlenecks, which led to the weakened purifying selection and fixation of otherwise deleterious mutations via neutral drift; gain of new introns and loss of ancestral ones might have been among such mutations.

An interesting possibility that we considered when incepting the present study was the detection of potential reverse-transcription-mediated duplication events. In the reconstructed scenarios of LSE evolution, such events would be manifested as reconstructed intronless evolutionary intermediates, which then might accumulate new introns. However, among the analyzed LSEs, we failed to identify a single reliable case of such an intronless 'bottleneck', indicating that reverse transcription contributed little to gene duplication, at least in the evolution of families of paralogous genes containing introns [proliferation of intronless genes, such as those for seven-transmembrane receptors in vertebrates, is well documented (21)].

To conclude, the findings reported here reveal a strong link between gene duplication leading to evolution of LSEs and enhanced intron insertion. However, we also found, supporting and extending previous observations made by comparing structures of orthologous genes, that at least the last ~100–200 Myr have been quite uneventful with respect to intron comings and goings, in both animals and plants. Together with previous results, these findings contribute to the emerging notion of intensive insertion and loss of introns during transitional epochs in contrast to the relative quiet of the intervening evolutionary spans.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Yuri Wolf, Alex Kondrashov and Michael Galperin for helpful discussions.

REFERENCES

1. Logsdon,J.M.,Jr (1998) The recent origins of spliceosomal introns revisited. *Curr. Opin. Genet. Dev.*, **8**, 637–648.
2. Lamond,A.I. (1999) RNA splicing. Running rings around RNA. *Nature*, **397**, 655–656.
3. Dacks,J.B. and Doolittle,W.F. (2001) Reconstructing/deconstructing the earliest eukaryotes: how comparative genomics can help. *Cell*, **107**, 419–425.
4. Lynch,M. and Richardson,A.O. (2002) The evolution of spliceosomal introns. *Curr. Opin. Genet. Dev.*, **12**, 701–710.
5. Fedorov,A., Merican,A.F. and Gilbert,W. (2002) Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proc. Natl Acad. Sci. USA*, **99**, 16128–16133.
6. Rogozin,I.B., Wolf,Y.I., Sorokin,A.V., Mirkin,B.G. and Koonin,E.V. (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol.*, **13**, 1512–1517.
7. Roy,S.W., Fedorov,A. and Gilbert,W. (2003) Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc. Natl Acad. Sci. USA*, **100**, 7158–7162.
8. Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–106.
9. Fitch,W.M. (2000) Homology: a personal view on some of the problems. *Trends Genet.*, **16**, 227–231.
10. Sonnhammer,E.L. and Koonin,E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.*, **18**, 619–620.
11. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
12. Remm,M., Storm,C.E. and Sonnhammer,E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.
13. Jordan,I.K., Makarova,K.S., Spouge,J.L., Wolf,Y.I. and Koonin,E.V. (2001) Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res.*, **11**, 555–565.
14. Lespinet,O., Wolf,Y.I., Koonin,E.V. and Aravind,L. (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.*, **12**, 1048–1059.
15. Ohno,S. (1970) *Evolution by gene duplication*. Springer-Verlag, Berlin, Heidelberg, NY.
16. Hughes,A.L. (1994) The evolution of functionally novel proteins after gene duplication. *Proc. R. Soc. Lond. B Biol. Sci.*, **256**, 119–124.
17. Lynch,M. and Conery,J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.
18. Kondrashov,F.A., Rogozin,I.B., Wolf,Y.I. and Koonin,E.V. (2002) Selection in the evolution of gene duplications. *Genome Biol.*, **3**, R8.
19. Chervitz,S.A., Aravind,L., Sherlock,G., Ball,C.A., Koonin,E.V., Dwight,S.S., Harris,M.A., Dolinski,K., Mohr,S., Smith,T. *et al.* (1998)

Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science*, **282**, 2022–2028.
20. Koonin,E.V., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Krylov,D.M., Makarova,K.S., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N., Rao,B.S. *et al*. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol*., **5**, R7.
21. Glusman,G., Sosinsky,A., Ben-Asher,E., Avidan,N., Sonkin,D., Bahar,A., Rosenthal,A., Clifton,S., Roe,B., Ferraz,C. *et al*. (2000) Sequence, structure, and evolution of a complete human olfactory receptor gene cluster. *Genomics*, **63**, 227–245.
22. Fedorov,A., Roy,S., Fedorova,L. and Gilbert,W. (2003) Mystery of intron gain. *Genome Res*., **13**, 2236–2241.
23. Cho,G. and Doolittle,R.F. (1997) Intron distribution in ancient paralogs supports random insertion and not random loss. *J. Mol. Evol*., **44**, 573–584.
24. Rzhetsky,A., Ayala,F.J., Hsu,L.C., Chang,C. and Yoshida,A. (1997) Exon/intron structure of aldehyde dehydrogenase genes supports the ''introns-late'' theory. *Proc. Natl Acad. Sci. USA*, **94**, 6820–6825.
25. Lecharny,A., Boudet,N., Gy,I., Aubourg,S. and Kreis,M. (2003) Introns in, introns out in plant gene families: a genomic approach of the dynamics of gene structure. *J. Struct. Funct. Genomics*, **3**, 111–116.
26. Boudet,N., Aubourg,S., Toffano-Nioche,C., Kreis,M. and Lecharny,A. (2001) Evolution of intron/exon structure of DEAD helicase family genes in *Arabidopsis*, *Caenorhabditis*, and *Drosophila*. *Genome Res*., **11**, 2101–2114.
27. Hankeln,T., Friedl,H., Ebersberger,I., Martin,J. and Schmidt,E.R. (1997) A variable intron distribution in globin genes of Chironomus: evidence for recent intron gain. *Gene*, **205**, 151–160.
28. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al*. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
29. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.
30. Felsenstein,J. (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol*., **266**, 418–427.
31. Farris,J.S. (1977) Phylogenetic analysis under Dollo's Law. *Syst. Zool*., **26**, 77–88.
32. Kimura,M. (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.
33. Hedges,S.B. and Kumar,S. (2003) Genomic clocks and evolutionary timescales. *Trends Genet*., **19**, 200–206.
34. Hedges,S.B., Chen,H., Kumar,S., Wang,D.Y., Thompson,A.S. and Watanabe,H. (2001) A genomic timescale for the origin of eukaryotes. *BMC Evol. Biol*., **1**, 4.
35. Dibb,N.J. and Newman,A.J. (1989) Evidence that introns arose at proto-splice sites. *EMBO J*., **8**, 2015–2021.
36. Dibb,N.J. (1991) Proto-splice site model of intron origin. *J. Theor. Biol*., **151**, 405–416.
37. Sadusky,T., Newman,A.J. and Dibb,N.J. (2004) Exon junction sequences as cryptic splice sites: implications for intron origin. *Curr. Biol*., **14**, 505–509.
38. Stephens,R.M. and Schneider,T.D. (1992) Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *J. Mol. Biol*., **228**, 1124–1136.
39. Sverdlov,A.V., Rogozin,I.B., Babenko,V.N. and Koonin,E.V. (2003) Evidence of splice signal migration from exon to intron during intron evolution. *Curr. Biol*., **13**, 2170–2174.
40. Tarrio,R., Rodriguez-Trelles,F. and Ayala,F.J. (2003) A new *Drosophila* spliceosomal intron position is common in plants. *Proc. Natl Acad. Sci. USA*, **100**, 6580–6583.
41. Wang,D.Y., Kumar,S. and Hedges,S.B. (1999) Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc. R. Soc. Lond. B Biol. Sci*., **266**, 163–171.
42. Zuckerkandl,E. and Pauling,L. (1962) Molecular evolution. In Kasha,M. and Pullman,B. (eds), *Horizons in Biochemistry*. Academic Press, NY, pp. 189–225.
43. Bromham,L. and Penny,D. (2003) The modern molecular clock. *Nature Rev. Genet*., **4**, 216–224.
44. Ayala,F.J., Rzhetsky,A. and Ayala,F.J. (1998) Origin of the metazoan phyla: molecular clocks confirm paleontological estimates. *Proc. Natl Acad. Sci. USA*, **95**, 606–611.
45. Aris-Brosou,S. and Yang,Z. (2003) Bayesian models of episodic evolution support a late precambrian explosive diversification of the Metazoa. *Mol. Biol. Evol*., **20**, 1947–1954.
46. Peterson,K.J., Lyons,J.B., Nowak,K.S., Takacs,C.M., Wargo,M.J. and McPeek,M.A. (2004) Estimating metazoan divergence times with a molecular clock. *Proc. Natl Acad. Sci. USA*, **101**, 6536–6541.
47. Haldane,J.B.S. (1933) The part played by recurrent mutation in evolution. *Am. Nature*, **67**, 5–19.
48. Zuckerkandl,E. and Pauling,L. (1965) Molecules as documents of evolutionary history. *J. Theor. Biol*., **8**, 357–366.
49. Lynch,M. and Force,A. (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics*, **154**, 459–473.
50. Rogers,J.H. (1989) How were introns inserted into nuclear genes? *Trends Genet*., **5**, 213–216.
51. Lynch,M. and Conery,J.S. (2003) The origins of genome complexity. *Science*, **302**, 1401–1404.
52. Koonin,E.V. (2004) A non-adaptationist perspective on evolution of genomic complexity or the continued dethroning of man. *Cell Cycle*, **3**, 280–285.