Current Biology, Vol. 13, 1512–1517, September 2, 2003, ©2003 Elsevier Science Ltd. All rights reserved. DOI 10.1016/S0960-9822(03)00558-X

Remarkable Interkingdom Conservation of Intron Positions and Massive, Lineage-Specific Intron Loss and Gain in Eukaryotic Evolution

Igor B. Rogozin,[1] Yuri I. Wolf,[1] Alexander V. Sorokin,[1] Boris G. Mirkin,[2] and Eugene V. Koonin[1,*]

[1]National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Bethesda, Maryland 20894

[2]School of Information Systems and Computer Science
Birkbeck College
University of London
Malet Street
London, WC1E 7HX
United Kingdom

Summary

Sequencing of eukaryotic genomes allows one to address major evolutionary problems, such as the evolution of gene structure. We compared the intron positions in 684 orthologous gene sets from 8 complete genomes of animals, plants, fungi, and protists and constructed parsimonious scenarios of evolution of the exon-intron structure for the respective genes. Approximately one-third of the introns in the malaria parasite *Plasmodium falciparum* are shared with at least one crown group eukaryote; this number indicates that these introns have been conserved through >1.5 billion years of evolution that separate *Plasmodium* from the crown group. Paradoxically, humans share many more introns with the plant *Arabidopsis thaliana* than with the fly or nematode. The inferred evolutionary scenario holds that the common ancestor of *Plasmodium* and the crown group and, especially, the common ancestor of animals, plants, and fungi had numerous introns. Most of these ancestral introns, which are retained in the genomes of vertebrates and plants, have been lost in fungi, nematodes, arthropods, and probably *Plasmodium*. In addition, numerous introns have been inserted into vertebrate and plant genes, whereas, in other lineages, intron gain was much less prominent.

Results and Discussion

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]. Positions of some spliceosomal introns are conserved in orthologous genes from plants and animals [2–4]. A recent systematic analysis of pairwise alignments of homologous proteins from animals, fungi, and plants suggested conservation of the positions of 10%–15% of the introns [5]. However, intron densities in eukaryotic species differ widely, the location of introns in orthologous genes does not always coincide even in closely related species [6], likely cases of intron insertion and loss have been documented [7–9], and indications of a high intron turnover rate have been obtained [10]. At least on some occasions, introns seem to have been lost as a result of a gene's recombination with a cDNA produced via reverse transcription of the respective mRNA [7–9].

In general, the evolutionary history of introns, the selective forces that shape intron evolution, and the relative contributions of intron loss and intron insertion in the evolution of eukaryotic genes remain poorly understood. We sought to analyze the evolution of the exon-intron structure of eukaryotic genes on the scale of complete genomes. Such an analysis requires, first, careful identification of orthologous gene sets (sets of genes derived from a single ancestral gene in the last common ancestor of the compared species) and, second, identification of orthologous ("the same") introns in each of these gene sets. We chose a conservative approach to each of these tasks. From the recently produced collection of eukaryotic clusters of orthologous genes (KOGs, after eukaryotic orthologous groups [11, 12]; COG database, http://www.ncbi.nlm.nih.gov/COG/new/), KOGs were extracted that were represented in six eukaryotic species with completely sequenced genomes, namely, humans, the nematode *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster*, two yeasts, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, and the green plant *Arabidopsis thaliana*. Orthologs from two other eukaryotic species, the mosquito *Anopheles gambiae* and the apicomplexan malarial parasite *Plasmodium falciparum*, were added to these KOGs by using the COGNITOR method ([11]; see the Supplemental Experimental Procedures available with this article online for details). For a pair of introns to be considered orthologous, they were required to occur in the exact same position in the aligned sequences of KOG members. Given the problems in the annotation of gene structure and difficulties in aligning poorly conserved regions of protein sequences, we employed two approaches to the analysis of evolutionary conservation of intron positions. Under the first schema, all intron positions were extracted from automatically produced alignments, whereas, under the second one, only positions embedded in well-conserved, unambiguous portions of the alignment were considered (see the Supplemental Experimental Procedures). Altogether, 684 KOGs were examined for intron conservation; these comprised a substantial majority, if not the entirety, of highly conserved eukaryotic genes that are amenable for analysis of exon-intron structure evolution.

The 684 analyzed KOGs contained 21,434 introns in 16,577 unique positions (10,066 introns in 7,236 positions when only the conserved portions of alignments were analyzed); 5,981 introns were conserved in two or more genomes (4,619 in conserved regions). The majority of the conserved introns were present in only two

*Correspondence: koonin@ncbi.nlm.nih.gov

Table 1. Conservation of Intron Positions in Eight Eukaryotic Species

Number of Species[a]		1	2	3	4	5	6	7	8
Number of introns – total	Observed[b]	13,406	2,047	719	275	104	25	1	0
	Expected	21,368	33	0	0	0	0	0	0
	Expected – 10%	20,083	662	8	0	0	0	0	0
Number of introns – conserved blocks	Observed[b]	5,446	1,122	411	163	74	19	1	0
	Expected	9,982	42	0	0	0	0	0	0
	Expected – 10%	8,613	689	25	0	0	0	0	0
Number of introns – conserved blocks, ±1	Observed	9,808	123	2	0	0	0	0	0
	Expected	9,834	116	0	0	0	0	0	0
Number of introns – conserved blocks, ±2	Observed	9,956	55	0	0	0	0	0	0
	Expected	9,838	114	0	0	0	0	0	0
Number of introns – conserved blocks, ±3	Observed	9,920	70	2	0	0	0	0	0
	Expected	9,844	111	0	0	0	0	0	0
Number of introns – conserved blocks, ±4	Observed	9,973	42	3	0	0	0	0	0
	Expected	9,848	109	0	0	0	0	0	0

[a] ±1, ±2, ±3, and ±4 indicate that intron positions separated by exactly 1, 2, 3, or 4 nucleotides, respectively, were analyzed.
[b] The probability that intron sharing in different species was due to chance is P(Monte Carlo) < 0.0001 (applies both to the analysis of all alignment positions and to the test with 10% of the positions allowed for intron insertion).

species, but a substantial number was found in three genomes and several introns were shared by 4–7 species (Table 1). A simulation of the intron distribution in the analyzed set of KOGs by random intron shuffling showed that ~1% of the observed number of introns shared by two species was expected to occur by chance, but that no introns were expected to occupy the same position in three or more species (Table 1). It has been proposed that introns insert into coding sequences not at random, but primarily into "proto-splice sites" [13], although the proto-splice model has been questioned as being inconsistent with the observed distribution of intron phases [14]. Considering this potential nonrandomness of intron insertion, we repeated the simulation and allowed intron insertion in 10% of the positions in the analyzed genes. This led to an increase in the expected number of coinciding introns in two or more species, but the excess of observed introns in the same position remained obvious and highly statistically significant (Table 1). As an additional control, we examined the expected and observed distribution of introns that were separated by 1–4 base pairs. A slight, nonsignificant excess of the observed over the expected number of introns was seen only for the distance of one base pair; this excess could be due to intron sliding [15, 16]. In contrast, there was a deficit in the number of introns separated by two, three, or four nucleotides, presumably due to the large excess of introns in the same position (Table 1). Taken together, these observations show that the great majority of introns located in the same position in orthologous genes from different eukaryotic lineages are orthologous, i.e., they originate from an ancestral intron in the same position in the respective gene of the last common ancestor of the compared species.

The matrix of shared introns in all pairs of analyzed eukaryotic genomes revealed a striking, unexpected pattern (Table 2). The number of conserved introns did not drop monotonically with the increase of the evolutionary distance between the compared organisms. On the contrary, human genes shared the greatest number of introns not with any of the three animals but with the plant *Arabidopsis*; in the conserved regions (the more accurate results given the uncertainties in alignment in other parts of genes), 24% of the analyzed human introns were shared with *Arabidopsis* (~27% of the *Arabidopsis* introns) compared to ~12%–17% of the introns shared by humans with the fly, mosquito, and the worm (Table 2). The difference becomes even more dramatic when the numbers of introns conserved in *Arabidopsis* and each of the three animal species are compared: approximately three times more plant introns have a counterpart in humans than in the fly or the worm (Table 2). Although *S. pombe* and *Plasmodium* have few introns compared to plants or animals, the same asymmetry was observed for these organisms: the numbers of introns shared with *Arabidopsis* and humans are close and are 2–3 times greater than the number of introns shared with the insects or the worm (Table 2).

Plasmodium belongs to the alveolate kingdom of protists, generally believed to have branched off the trunk of the eukaryotic tree prior to the divergence of the lineages that comprise the eukaryotic crown group, including animals, plants, and fungi [17]. Therefore, it is particularly notable that *Plasmodium* shares 143 (nearly one-third) of the 450 introns present in the conserved regions of the analyzed genes with at least one crown group species (Table 2). Furthermore, the results suggested the possibility that the common ancestor of the crown group had an intron-rich genome; the majority of the ancestral introns seem to have survived in plants and vertebrates but have been lost in yeasts, nematodes, and arthropods.

To explore the evolutionary dynamics of introns, we turned to phylogenetic analysis. For this purpose, intron positions were represented as a data matrix of intron absence/presence (encoded as 0/1; Figure 1). The intron absence/presence data are conducive to evolutionary parsimony analysis; the Dollo parsimony, which is based on the assumption that each derived character state (in this case, intron presence) originated only once on the tree, seems to be most appropriate in this case. However, the reconstructed Dollo parsimony tree obviously does not mimic the evolutionary history of the analyzed species, with humans and *Arabidopsis* forming a strongly supported cluster embedded within the metazoan

Table 2. Conservation of Intron Positions among Eukaryotes

	Pf	Sc	Sp	At	Ce	Dm	Ag	Hs
Pf	**450/971**	2	48	137	50	46	54	145
Sc	1	**22/46**	7	3	3	3	4	6
Sp	34	6	**450/839**	209	98	114	111	308
At	97	2	147	**2933/5589**	353	255	254	1148
Ce	33	2	63	240	**1468/3465**	315	312	948
Dm	32	1	72	161	179	**723/1826**	787	802
Ag	36	1	62	158	176	382	**675/1768**	771
Hs	104	3	207	787	557	433	403	**3345/6930**

The diagonal (numbers in bold) shows the total number of introns in the 684 analyzed genes (denominator) from the given species and the number of introns in conserved regions of alignments (numerator). For each pair of species, the total number of shared introns is shown above the diagonal, and the number of introns in conserved regions is shown below the diagonal. Species abbreviations: At, *Arabidopsis thaliana*; Ce, *Caenorhabditis elegans*; Dm, *Drosophila melanogaster*; Hs, *Homo sapiens*; Ag, *Anopheles gambiae*; Pf, *Plasmodium falciparum*; Sc, *Saccharomyces cerevisiae*; Sp, *Schizosaccharomyces pombe*.

clade and another anomalous cluster formed by *S. pombe* and *Plasmodium* (Figure 2). Other phylogenetic approaches, including unweighted maximum parsimony and several distance methods, reproduced the topology seen in Figure 2 (not shown). The topology of the constructed trees supports the notion, already suggested by the data in Table 2, that ancestral introns have been, to a large extent, conserved in plants and vertebrates but have been extensively eliminated in fungi, nematodes, and arthropods. These observations show that intron locations are not suitable markers for phylogenetic analysis at long evolutionary distances.

Having shown that evolution of introns in eukaryotic genes did not follow the species tree, we applied parsimony in the opposite direction: given a species tree topology, we constructed the most parsimonious (and, by inference, most likely) scenario for intron evolution, i.e., the distribution of intron gain and loss events over

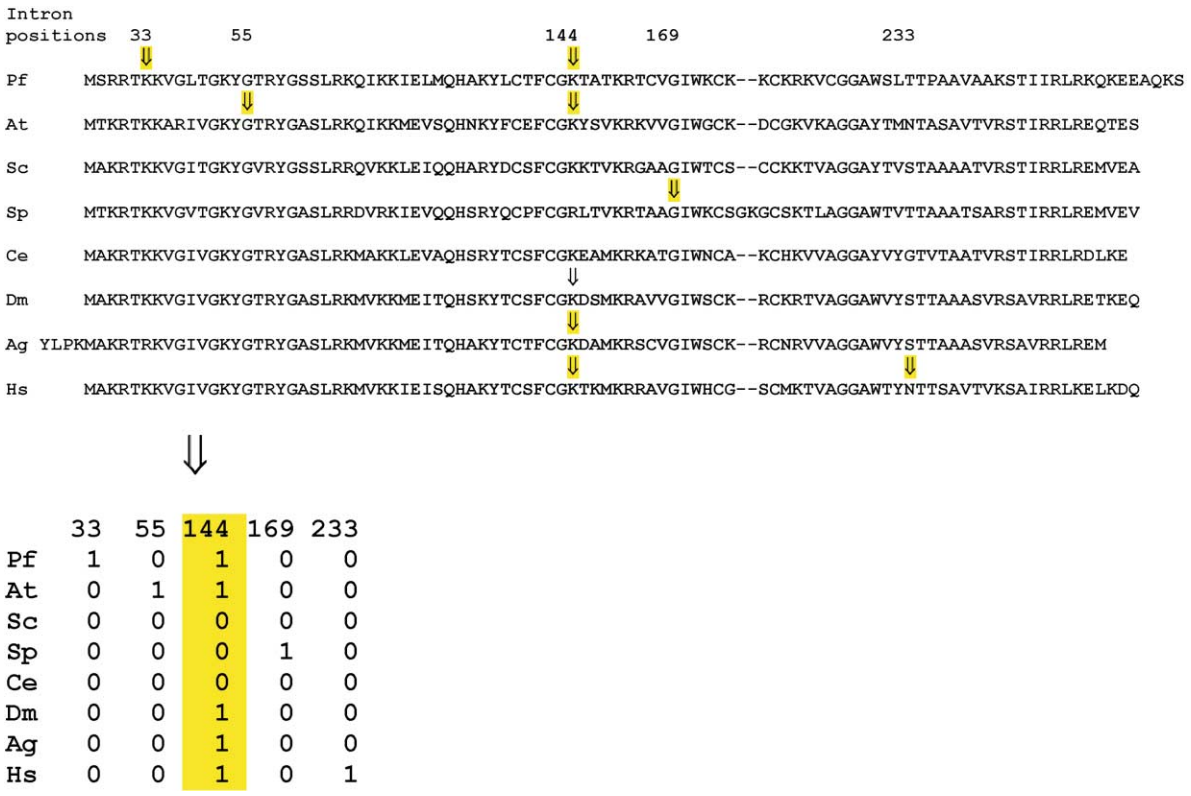

Figure 1. Examples of Conservation and Variability of Intron Positions in Orthologous Eukaryotic Genes

The data are for KOG0402 (ribosomal protein L37). The intron positions are shown directly on the alignment, and the conversion of the intron alignment mapping into an absence/presence matrix is illustrated. "1" indicates the presence of an intron, and "0" indicates the absence of an intron in the given alignment position (shown on the top of the table). The highly conserved intron positions are highlighted. Species abbreviations: At, *Arabidopsis thaliana*; Ce, *Caenorhabditis elegans*; Dm, *Drosophila melanogaster*; Hs, *Homo sapiens*; Ag, *Anopheles gambiae*; Pf, *Plasmodium falciparum*; Sc, *Saccharomyces cerevisiae*; Sp, *Schizosaccharomyces pombe*.

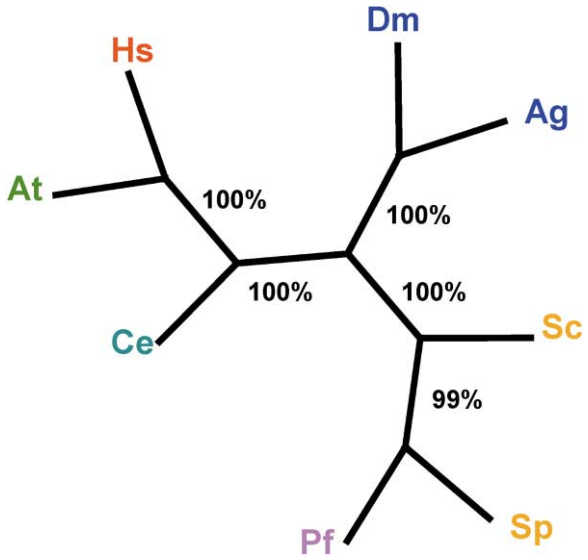

Figure 2. A Maximum Parsimony Tree Based on the Concatenated Intron Absence/Presence Data

Only the data for conserved alignment regions were analyzed. The unrooted tree was constructed by using Dollo parsimony. Only one most parsimonious tree was obtained; the numbers at the interior branches are bootstrap values with 1000 replicates. The species abbreviations are as in Figure 1.

the tree branches. Recent, large-scale phylogenetic studies favor a topology of the eukaryotic crown group tree, in which metazoa and fungi form a clade to the exclusion of plants [18, 19], and evidence is also accumulating in support of the coelomate clade (vertebrates with arthropods), as opposed to the ecdysozoan clade [20], among the animals [17, 19] (Y.I.W., I.B.R., and E.V.K., unpublished data). We constructed the most parsimonious scenario of evolution for intron positions assuming this tree topology (Figures 3A and 3B); qualitatively similar results were obtained for complete alignments (Figure 3A) and the conserved regions only (Figure 3B). The resulting scenario suggests an intron-rich ancestor of the crown group, with minimal intron loss in the animal ancestor but massive losses in yeasts (particularly *S. cerevisiae*), worm, and insects (Figure 3). Of further note are the differences in the relative rates of intron gain and loss in the terminal branches: there is a dramatic excess of gains over losses in humans and *S. pombe* and an equally notable excess of losses in insects and *S. cerevisiae*, whereas *C. elegans* shows nearly equal numbers of gains and losses. All introns shared by *Plasmodium* and any of the crown group species (at least 210; Figure 3A) are assigned to the last common ancestor of alveolates and the crown group, which lived some 1.5–2.0 billion years ago [17]. At present, loss of ancestral introns in *Plasmodium* cannot be detected because *Plasmodium* is the outgroup with respect to all other analyzed species; hence, we produced a conservative estimate of the number of the most ancient introns in the analyzed gene set. It is likely to be a substantial underestimate given that *Plasmodium* is a parasite with a highly degraded genome and low intron density [21]. In the apicomplexan *Toxoplasma gondii*, a relative of *Plasmodium*, the intron density was somewhat higher than in *Drosophila* [6]. Furthermore, an alternative tree topology, in which *Apicomplexa* clusters with plants, has been proposed on the basis of phylogenetic analysis of several protein families [19]. When this topology was employed for constructing an evolutionary scenario, the last common ancestor of the eukaryotic species with sequenced genomes comes out particularly intron rich (Figure 3C).

The present analysis pushes the origin of numerous spliceosomal introns back to the stage of eukaryotic evolution, 1.5–2.0 billion years ago, which precedes the origin of multicellularity and shows that as many as 25%–30% of the introns in vertebrates and plants are apparently inherited from the common ancestor of the crown group. Why have so many ancestral introns survived almost 2 billion years of evolution? One intriguing possibility is that conserved introns are functionally important, but there is currently little evidence in support of this hypothesis. The second explanation is that the elimination of introns, particularly in essential genes, such as those included in the present study, would often lead to gene inactivation and, accordingly, would be lethal. We found that ancient introns, compared to more recently inserted ones, tend to be located in less variable portions of genes (see Table S1 in the Supplemental Data); this finding seems to be compatible with the above hypothesis. In some lineages, particularly the yeasts, but also insects and nematodes, the disadvantage of intron loss apparently had been overcome by selective pressure for genome streamlining or as a result of retrotransposition sweeps or both. However, absent such specific circumstances favoring intron elimination, many ancestral introns might have survived simply because losing them is costly.

Evolution of spliceosomal introns had been long considered in the context of the "intron-early" versus "intron-late" debate. The "intron-early" hypothesis suggests that introns existed before the divergence of prokaryotes and eukaryotes [22, 23]. In contrast, the "intron-late" hypothesis holds that introns have been inserted into eukaryotic genes after this divergence [6, 10, 24]. The present observations do not bear directly on the substance of this debate, but they do show that many introns not only emerged shortly after the origin of eukaryotes, but retained their positions in some eukaryotic lineages. Interestingly, and contrary to the predictions of the "intron-early" hypothesis [25], we found that the excess of phase 0 introns (those inserted between codons) over phase 1 and 2 introns was even greater in relatively recently acquired introns than in ancient ones (see Table S2 in the Supplemental Data). The evidence presented here and elsewhere [5, 10] appears to be compatible with a "many introns very early in eukaryotic evolution" view. The recent discovery of introns in several protists, which might be the deepest branches in eukaryotic phylogeny [26–28], is compatible with this view. It even seems possible that invasion of protein-coding genes by ancestors of introns was part of the dramatic and still mysterious series of events that led to the origin of the eukaryotic cell.

The remarkable conservation of many intron positions and the extensive loss of introns in some of the eu-

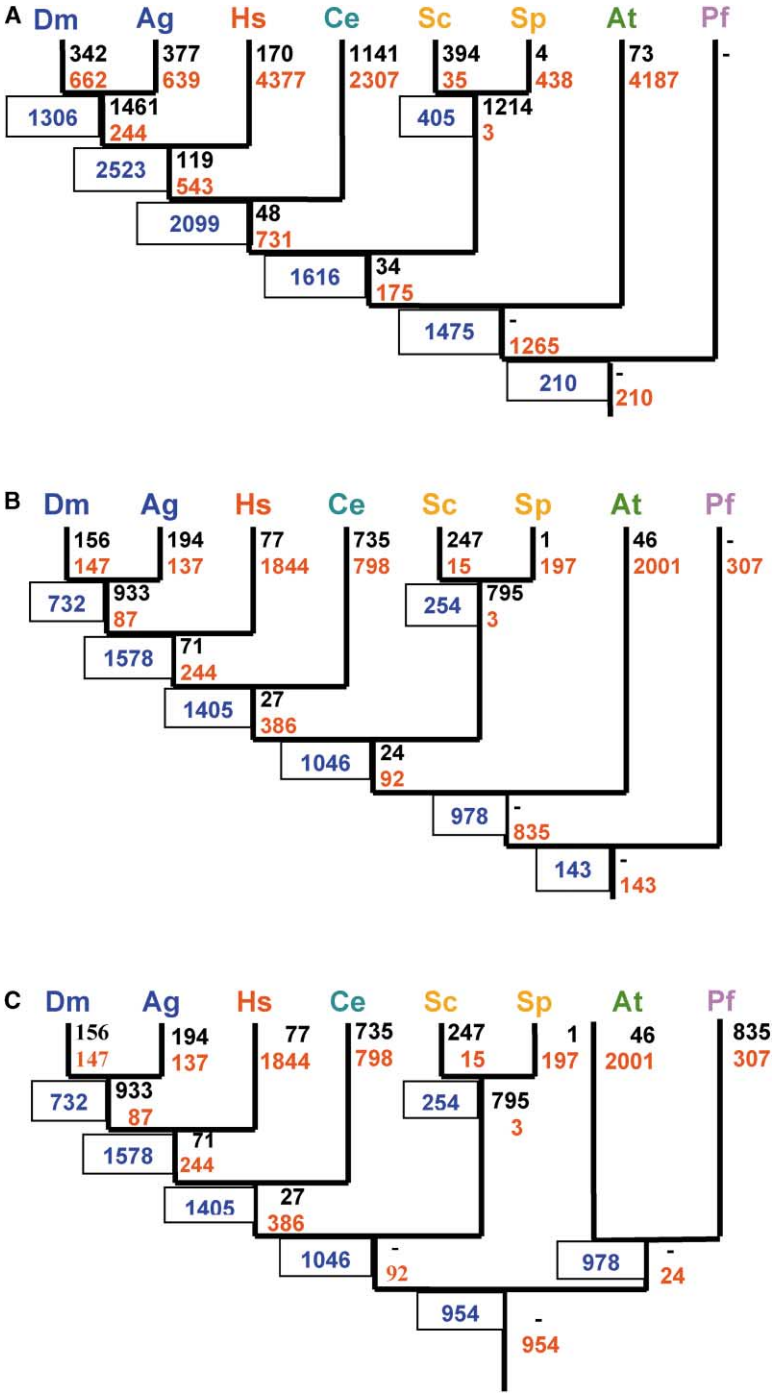

Figure 3. The Parsimonious Evolutionary Scenario of Intron Gain/Loss for the Most Likely Topology of the Eukaryotic Phylogenetic Tree

Intron gains and losses are mapped to each species and each internal branch. Intron losses are shown in black and intron gains are shown in red; the dashes show branches for which losses could not be inferred from the available data. The (minimal) number of introns inferred to have existed in the analyzed set of genes in the respective ancestral forms is indicated in a box next to each internal node of the tree. The species abbreviations are as in Figure 1.

(A) Data for complete alignments.

(B) Data for conserved alignment regions.

(C) Scenario for an alternative tree topology, with *Plasmodium* forming a clade with *Arabidopsis*. Data for conserved alignment regions.

karyotic lineages should not overshadow the observation that, in other lineages, such as vertebrates and plants, the majority of introns apparently have been gained relatively recently (Figure 3). The lineage-specific trends of intron loss and gain might reflect more general tendencies for genome compaction and genome expansion, the underlying causality of which remains to be understood.

Supplemental Data

The list of KOGs used for this analysis, amino acid sequence alignments, and intron absence/presence matrices for each KOG are available at the authors' website at ftp://ftp.ncbi.nlm.nih.gov/pub/koonin/intron_evolution. Supplemental Data including the Experimental Procedures and supplemental tables are available at http://www.current-biology.com/cgi/content/full/13/17/1512/DC1.

Acknowledgments

We thank Roman Tatusov for his contribution to the construction of the KOGs, Masatoshi Nei, Fyodor Kondrashov, Mikhail Gelfand, Peter Gogarten, Andrei Mironov, and Mark Borodovsky for simulating discussions, and Fyodor Kondrashov and I. King Jordan for critical reading of the manuscript.

Received: June 12, 2003
Revised: June 30, 2003
Accepted: July 4, 2003
Published: September 2, 2003

References

1. Lamond, A.I. (1999). RNA splicing. Running rings around RNA. Nature *397*, 655–656.
2. Marchionni, M., and Gilbert, W. (1986). The triosephosphate isomerase gene from maize: introns antedate the plant-animal divergence. Cell *46*, 133–141.
3. Logsdon, J.M., Jr., Tyshenko, M.G., Dixon, C., D-Jafari, J., Walker, V.K., and Palmer, J.D. (1995). Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory. Proc. Natl. Acad. Sci. USA *92*, 8507–8511.
4. 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.
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. Logsdon, J.M., Jr. (1998). The recent origins of spliceosomal introns revisited. Curr. Opin. Genet. Dev. *8*, 637–648.
7. 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.
8. Logsdon, J.M., Jr., Stoltzfus, A., and Doolittle, W.F. (1998). Molecular evolution: recent cases of spliceosomal intron gain? Curr. Biol. *8*, R560–R563.
9. Feiber, A.L., Rangarajan, J., and Vaughn, J.C. (2002). The evolution of single-copy *Drosophila* nuclear 4f-rnp genes: spliceosomal intron losses create polymorphic alleles. J. Mol. Evol. *55*, 401–413.
10. Lynch, M., and Richardson, A.O. (2002). The evolution of spliceosomal introns. Curr. Opin. Genet. Dev. *12*, 701–710.
11. Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997). A genomic perspective on protein families. Science *278*, 631–637.
12. 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, in press.
13. Dibb, N.J. (1991). Proto-splice site model of intron origin. J. Theor. Biol. *151*, 405–416.
14. Long, M., and Rosenberg, C. (2000). Testing the "proto-splice sites" model of intron origin: evidence from analysis of intron phase correlations. Mol. Biol. Evol. *17*, 1789–1796.
15. Stoltzfus, A., Logsdon, J.M., Jr., Palmer, J.D., and Doolittle, W.F. (1997). Intron "sliding" and the diversity of intron positions. Proc. Natl. Acad. Sci. USA *94*, 10739–10744.
16. Rogozin, I.B., Lyons-Weiler, J., and Koonin, E.V. (2000). Intron sliding in conserved gene families. Trends Genet. *16*, 430–432.
17. Hedges, S.B. (2002). The origin and evolution of model organisms. Nat. Rev. Genet. *3*, 838–849.
18. Kumar, S., and Rzhetsky, A. (1996). Evolutionary relationships of eukaryotic kingdoms. J. Mol. Evol. *42*, 183–193.
19. Baldauf, S.L., Roger, A.J., Wenk-Siefert, I., and Doolittle, W.F. (2000). A kingdom-level phylogeny of eukaryotes based on combined protein data. Science *290*, 972–977.
20. Aguinaldo, A.M., Turbeville, J.M., Linford, L.S., Rivera, M.C., Garey, J.R., Raff, R.A., and Lake, J.A. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. Nature *387*, 489–493.
21. Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., et al. (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature *419*, 498–511.
22. Gilbert, W. (1987). The exon theory of genes. Cold Spring Harb. Symp. Quant. Biol. *52*, 901–905.
23. Gilbert, W., and Glynias, M. (1993). On the ancient nature of introns. Gene *135*, 137–144.
24. Cavalier-Smith, T. (1985). Selfish DNA and the origin of introns. Nature *315*, 283–284.
25. de Souza, S.J., Long, M., Klein, R.J., Roy, S., Lin, S., and Gilbert, W. (1998). Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. Proc. Natl. Acad. Sci. USA *95*, 5094–5099.
26. Nixon, J.E., Wang, A., Morrison, H.G., McArthur, A.G., Sogin, M.L., Loftus, B.J., and Samuelson, J. (2002). A spliceosomal intron in *Giardia lamblia*. Proc. Natl. Acad. Sci. USA *99*, 3701–3705.
27. Simpson, A.G., MacQuarrie, E.K., and Roger, A.J. (2002). Eukaryotic evolution: early origin of canonical introns. Nature *419*, 270.
28. Archibald, J.M., O'Kelly, C.J., and Doolittle, W.F. (2002). The chaperonin genes of jakobid and jakobid-like flagellates: implications for eukaryotic evolution. Mol. Biol. Evol. *19*, 422–431.