Three distinct modes of intron dynamics in the evolution of eukaryotes

Liran Carmel, Yuri I. Wolf, Igor B. Rogozin and Eugene V. Koonin

Genome Res. 2007 17: 1034-1044; originally published online May 10, 2007;
Access the most recent version at doi:10.1101/gr.6438607

Letter

Three distinct modes of intron dynamics in the evolution of eukaryotes

Liran Carmel, Yuri I. Wolf, Igor B. Rogozin, and Eugene V. Koonin[1]

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

Several contrasting scenarios have been proposed for the origin and evolution of spliceosomal introns, a hallmark of eukaryotic genes. A comprehensive probabilistic model to obtain a definitive reconstruction of intron evolution was developed and applied to 391 sets of conserved genes from 19 eukaryotic species. It is inferred that a relatively high intron density was reached early, i.e., the last common ancestor of eukaryotes contained >2.15 introns/kilobase, and the last common ancestor of multicellular life forms harbored ~3.4 introns/kilobase, a greater intron density than in most of the extant fungi and in some animals. The rates of intron gain and intron loss appear to have been dropping during the last ~1.3 billion years, with the decline in the gain rate being much steeper. Eukaryotic lineages exhibit three distinct modes of evolution of the intron–exon structure. The primary, balanced mode, apparently, operates in all lineages. In this mode, intron gain and loss are strongly and positively correlated, in contrast to previous reports on inverse correlation between these processes. The second mode involves an elevated rate of intron loss and is prevalent in several lineages, such as fungi and insects. The third mode, characterized by elevated rate of intron gain, is seen only in deep branches of the tree, indicating that bursts of intron invasion occurred at key points in eukaryotic evolution, such as the origin of animals. Intron dynamics could depend on multiple mechanisms, and in the balanced mode, gain and loss of introns might share common mechanistic features.

[Supplemental material is available online at www.genome.org.]

Spliceosomal introns interrupting protein-coding genes and the concurrent splicing machinery are among the defining features of eukaryotes (Doolittle 1978; Gilbert 1978; Mattick 1994; Deutsch and Long 1999). To date, all eukaryotes with fully sequenced genomes bear introns. Different species vary dramatically in their intron density, ranging from a few introns per genome to over eight per gene (Logsdon 1998; Mourier and Jeffares 2003; Jeffares et al. 2006). Despite this strong foothold in eukaryotic genomes, introns proved astonishingly effective in keeping their secrets. Little is known about the way they first appeared and penetrated genomes, about their subsequent propagation in eukaryotic genomes, about the mechanisms by which they are lost or gained, and about their functional role, if any.

What had become increasingly recognized in recent years is that introns and the splicing machinery evolved at a very early stage of eukaryogenesis. All eukaryotes with sequenced genomes, including parasitic protists with compact genomes, previously suspected of being intronless, have been shown to possess at least a few introns (Nixon et al. 2002; Simpson et al. 2002; Vanacova et al. 2005) and a (nearly) full complement of spliceosomal proteins (Collins and Penny 2005). Thus, the emergence of introns and the splicing machinery seems to antedate the last common ancestor of all extant eukaryotes and might have been linked to the emergence of other signature eukaryotic features, including the nucleus (Martin and Koonin 2006).

Beyond the general notion of the ancient origin of introns and the spliceosome, the evolutionary dynamics of eukaryotic gene structure, which is manifested in intron gain and loss, has been a subject of intense investigation. Generally, the abundance of introns in a genome is thought to be determined by the effective population size and the characteristic mutation rate of the respective species (Lynch and Richardson 2002; Lynch and Conery 2003). However, it has been argued that various selective forces could substantially affect the rates of intron gain and loss (Jeffares et al. 2006). Furthermore, in at least one case study, intron loss in *Drosophila* appears to have been driven by positive selection (Llopart et al. 2002). Comparative genomic studies have revealed impressive conservation of intron positions in diverse animals (Raible et al. 2005) and have shown that the positions of many introns are shared by orthologous genes even in distant eukaryotes, such as animals and plants (Fedorov et al. 2002; Rogozin et al. 2003). However, the evolutionary history of introns in eukaryotes remains a matter of contention (Rogozin et al. 2005b; Rodriguez-Trelles et al. 2006; Roy and Gilbert 2006). In several recent large-scale studies, the evolutionary dynamics of introns was examined over the entire eukaryotic tree. These attempts, however, yielded widely contradicting scenarios. While Qiu et al. (2004) concluded that intron gains were overwhelmingly dominant in eukaryotic evolution, the other studies detected both gains and losses but disagreed on their relative contributions. Analyzing the same set of orthologous genes from eight species (Rogozin et al. 2003), some found an overall excess of gains (Nguyen et al. 2005), others reported a substantial excess of losses (Roy and Gilbert 2005a,b,c, 2006), and yet others did not offer conclusive statements on the relative contributions of gains and losses (Rogozin et al. 2003; Csuros 2005). Each of these studies used a different set of assumptions and simplifications, and employed a different inference technique, making it hard to decide between the conflicting scenarios of intron evolution (Rogozin et al. 2005b). Specifically, Rogozin et al. (2003) used

[1]Corresponding author.
E-mail koonin@ncbi.nlm.nih.gov; fax (301) 480-9241.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6438607. Freely available online through the *Genome Research* Open Access option.

17:1034–1044 ©2007 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/07; www.genome.org

Dollo parsimony to infer ancestral states, a method that is known to overestimate gains in terminal branches, hence underestimating the number of introns in ancestral genomes (Csuros 2005; Rogozin et al. 2005a). Roy and Gilbert (2005a,c) employed a simple evolutionary model, in which different lineages are associated with specific loss and gain probabilities, and formulated an inference procedure that combines maximum likelihood (ML) principles and parsimony. However, this hybrid technique has been shown to inflate the estimated number of introns in ancestral forms (Csuros 2005). The model of Roy and Gilbert assumes that the gain and loss rates depend only on the lineage, thus tacitly presuming that all genes have identical rates of intron gain and loss. The inverse approach was adopted by Qiu et al. (2004), who assumed that the rates of intron gain and loss are different between genes, but for a particular gene remain constant across the entire phylogenetic tree. The latter assumption is hard to accept given the apparent dramatic differences in the rates of intron turnover in different lineages (Fedorov et al. 2003; Cho et al. 2004; Roy and Hartl 2006). Recently, two ML techniques have been developed for essentially the same evolutionary model as that of Roy and Gilbert (Csuros 2005; Nguyen et al. 2005). Both methods based their inference on intron-bearing sites only and, consequently, run into the need to estimate the total number of intronless sites in the data. These studies employed technically different but conceptually similar methods to evaluate this number, yielding, as expected, very similar results. Predictably, these estimates were higher than those obtained with Dollo parsimony but lower than those produced by the hybrid technique of Roy and Gilbert.

With the exception of the work of Qiu et al. (2004), all these studies used the eight-species data compiled by Rogozin et al. (2003). However, it seems that richer sampling is required in order to arrive at more definite conclusions. Moreover, as mentioned above, these works assume that all the genes have identical rates of intron gain and loss. While this assumption simplifies the analysis, it has two undesirable effects. First, the model of evolution is necessarily incomplete, as genes undoubtedly differ in their tendencies to gain and lose introns. Second, the description of the evolution is, obviously, restricted to the lineage level. None of these models yields any information on intron dynamics at the gene level.

Here, we develop means to overcome these problems. First, we compiled a considerably larger data set, comprised of 391 sets of orthologous genes from 19 eukaryotic species. This extended data set not only allows for more definite reconstruction of gene structure evolution, but also permits zooming in on specific portions of the eukaryotic tree. Second, we developed a comprehensive probabilistic model of intron evolution that allows for intron gain and loss rate heterogeneity between both lineages and genes. In addition, this model allows for intron gain and loss rate variability among sites within a gene, thus accounting for all sites, including intronless ones, and avoiding difficulties of the other methods. Applying this model to the 19-species data set, we obtained a detailed evolutionary reconstruction of intron–exon structure. A method was developed to decompose this reconstruction into the relative contributions of the lineages and the genes. In this paper, we report the results of the analysis at the more traditional lineage level, whereas the results at the gene level are presented in the accompanying paper (Carmel et al. 2007). We demonstrate that ancestral eukaryotic forms were intron-rich and show that evolution of eukaryotic genes involved numerous gains and losses of introns, with losses being somewhat more common. We identify three distinct modalities of intron gain and loss during eukaryotic evolution. The balanced mode appears to operate in all eukaryotic lineages, and is characterized by approximately proportional intron gain and loss rates, thus refuting previous predictions of an inverse correlation between the two. On top of this universal process, some lineages exhibit elevated loss rate, whereas others exhibit elevated gain rate. Moreover, we show that rates of intron gain and loss are highly nonuniform over evolutionary times, and both rates decreased with time in the last 1.3 billion yr. The decrease in gains was faster than the decrease in losses, resulting in many lineages with very limited intron gain over the last several hundred million years.

Results and Discussion

The model of evolution

Suppose we observe intron positions in orthologous genes from S eukaryotic species. Let the evolutionary relationships between these species be described by a rooted phylogeny of $N = 2S - 1$ nodes. Assigning each node with a state, either one (presence of an intron) or zero (absence of an intron), defines a history of intron evolution at a particular genomic site. We denote by q_t the state of node t in the tree, and by q_t^P the state of its parent node. By convention, we index the root of the tree as zero, and its state is therefore q_0. We index the branches of the tree by the node into which they lead, and use Δ_t for the length of the branch (in time units) leading into node t. Hereafter, we assume that the tree topology, as well as the branch lengths $\Delta_1, \ldots, \Delta_{N-1}$, are known.

We assume that each gene g has an intrinsic intron gain rate per site (η_g) and intron loss rate per site (θ_g), such that the tendency of a gene to gain or retain an intron at a particular site during a time interval Δ is $1 - e^{-\eta_g \Delta}$ and $e^{-\theta_g \Delta}$, respectively. Similarly, each branch t has an intrinsic intron gain rate per site (Ξ_t) and intron loss rate per site (Φ_t), such that the tendency of a branch whose length is Δ_t to gain or retain an intron at a particular site is $1 - e^{-\Xi_t \Delta_t}$ and $e^{-\Phi_t \Delta_t}$, respectively. For convenience, we define the *branch-specific intron gain coefficient* as $\xi_t = 1 - e^{-\Xi_t \Delta_t}$, and the *branch-specific intron loss coefficient* as $\phi_t = 1 - e^{-\Phi_t \Delta_t}$.

The central part of the model is the transition matrix for gene g along branch t, $T_{ij}(g,t) = P(q_t = j | q_t^P = i, g)$, that takes the form

$$T(g,t) = \begin{pmatrix} 1 - \xi_t(1 - e^{-\eta_g \Delta_t}) & \xi_t(1 - e^{-\eta_g \Delta_t}) \\ 1 - (1 - \phi_t)e^{-\theta_g \Delta_t} & (1 - \phi_t)e^{-\theta_g \Delta_t} \end{pmatrix}.$$

Clearly, the probability of each event depends on both the gene and the branch where the event takes place. The probability to gain an intron in gene g along branch t is $\xi_t(1 - e^{-\eta_g \Delta_t})$. Thus, the gain probability is a product of terms contributed by the branch (ξ_t) and by the gene ($1 - e^{-\eta_g \Delta_t}$). Similarly, the probability to retain an existing intron is $(1 - \phi_t)e^{-\theta_g \Delta_t}$. Thus, for an intron to be retained, it should not be lost along the branch ($1 - \phi_t$) and not be lost by the gene ($e^{-\theta_g \Delta_t}$). To complete the probabilistic model, we denote by π_i the prior probability of the root of the tree to be in state i ($i = 0,1$) in a particular site.

The second major improvement in the model is that we allow for rate variability across the sites of each gene. In phylo-

genetic analysis, rate variability is typically modeled by associating each site with a rate variable, r, which scales the branch lengths of the corresponding phylogenetic tree, $\Delta_t \leftarrow r \cdot \Delta_t$ (Felsenstein 2004). This rate variable is drawn from a probability distribution with non-negative domain and unit mean, typically the unit-mean gamma distribution. This, however, should be modified for intron evolution, where the gain and loss processes are not necessarily correlated. Therefore, we model rate variability using two independent rate variables, r^η and r^θ, such that $\eta_g \leftarrow r^\eta \cdot \eta_g$ and $\theta_g \leftarrow r^\theta \cdot \theta_g$. These rates are independently drawn from the two distributions

$$r^\eta \sim \nu\delta(\eta) + (1-\nu)\Gamma(\eta;\lambda_\eta)$$
$$r^\theta \sim \Gamma(\theta;\lambda_\theta).$$

Here, $\Gamma(x;\lambda)$ is the unit-mean gamma distribution of variable x with shape parameter λ, $\delta(x)$ is the Dirac delta-function, and ν is the fraction of sites that are incapable of gaining introns (hereafter intronless sites). The intronless sites are a direct realization of the proto-splice sites hypothesis that suggests that introns are preferentially inserted into short, specific sequence motifs termed proto-splice sites, whereas sites that deviate significantly from these motifs are extremely unlikely to gain introns (Dibb and Newman 1989; Dibb 1991; Sverdlov et al. 2004b). Although the identity of the intronless sites might vary between lineages, we assume that their density is constant throughout eukaryotic evolution. There is no analog of the intronless sites when it comes to intron loss as it is assumed that, once an intron is gained, it can always be lost. As is the common practice in the field (Yang 1994), we approximate the continuous gamma distributions by discrete versions, using K_η and K_θ categories for $\Gamma(\eta;\lambda_\eta)$ and $\Gamma(\theta;\lambda_\theta)$, respectively.

The two-phase data analysis technique: Homogeneous and heterogeneous phases

The parameters of the model are estimated using an expectation–maximization (EM) algorithm, which is an efficient realization of the maximum-likelihood (ML) approach for parameter estimation (see Methods). If G is the number of genes and S is the number of species, the complete model is characterized by $2G + 4S$ parameters. With a data set in the hundreds of genes, this number becomes prohibitively large, resulting in an intolerably high variance of the parameters' estimates. The plurality of parameters, therefore, hinders straightforward application of the algorithm and forces us to use more elaborate techniques. To this end, we developed a two-phase approach to the data analysis. In the first, "homogeneous evolution," phase, all genes were concatenated and hence all were assumed to have equal rates of intron loss and gain (thus, $\eta_g = \eta_0$ and $\theta_g = \theta_0$ for each gene g). Gene concatenation is effective in reducing the number of parameters ($G = 1$) but obscures differences between genes. In the second, "heterogeneous evolution," phase, all parameters estimated in the homogeneous phase were fixed, and only the gene-specific intron gain and loss rates (η_g and θ_g, respectively) were estimated.

The algorithm not only estimates the model parameters but also provides estimates for ancestral states; i.e., it computes the probability of finding each of the ancestral nodes in any given state, and the probability of gain and loss events along each branch. This information is summarized in a set of three matrices, hereafter denoted *reconstruction*:

1. Intron presence/absence, P: A matrix of size $S - 1$ (number of internal nodes) over G, with $P(t,g)$ estimating the number of introns in gene g at ancestral node t.
2. Intron gain, A: A matrix of size $N - 1$ (number of branches) over G, with $A(t,g)$ estimating the number of gain events in gene g along branch t.
3. Intron loss, L: A matrix of size $N - 1$ (number of branches) over G, with $L(t,g)$ estimating the number of loss events in gene g along branch t.

A similar reconstruction is obtained after the homogeneous phase, but with P, A, and L being vectors ($G = 1$) instead of matrices.

We found that the estimated model parameters are poorly suited to serve as the basis for the analysis of intron gain and loss trends because different sets of parameters yield very similar reconstructions (see propositions 1 and 2 in Nguyen et al. 2005,

Table 1. Intron densities (known or inferred) for each node, as well as inferred density of intron gain and loss events along each branch

Node	Intron density[a]	Intron gain density[a]	Intron loss density[a]
Eukaryota	3.19		
AME	3.39	2.06	1.86
Unikonts	3.10	0.13	0.42
Opisthokonts	3.70	0.69	0.10
Metazoa	5.22	1.97	0.44
Coelomata	5.14	0.00	0.09
Deuterostomia	6.17	1.20	0.17
Diptera	1.91	0.45	3.68
Fungi	2.85	0.34	1.19
Ascomycota	2.47	0.63	1.00
ScAfNc	1.30	0.00	1.17
Magnoliophyta	4.97	3.07	1.48
Chordata	6.22	0.39	0.34
Vertebrata	6.20	0.28	0.30
Apicomplexa	2.18	0.00	1.01
Pezizomycotina	1.66	0.56	0.20
Amniota	6.15	0.00	0.05
Mammals	6.10	0.00	0.05
Dicdi	0.96	0.19	2.33
Caeel	2.54	1.33	4.01
Strpu	5.67	0.51	1.01
Cioin	4.16	1.21	3.27
Danre	6.16	0.24	0.28
Galga	6.00	0.23	0.37
Homsa	5.94	0.08	0.24
Roden	5.28	0.07	0.88
Drome	1.28	0.10	0.74
Anoga	1.23	0.16	0.85
Cryne	3.75	1.86	0.96
Schpo	0.75	0.11	1.83
Sacce	0.03	0.01	1.28
Aspfu	1.62	0.18	0.21
Neucr	1.26	0.40	0.80
Arath	4.99	0.23	0.21
Orysa	5.10	0.32	0.20
Thepa	2.54	1.04	0.68
Plafa	0.71	0.16	1.62

The values are for the tree topology in Supplemental Fig. S3. The data for alternative tree topologies are available in Supplemental Table S3A–C. Only the optimal values are given for each lineage. The complete results, with confidence intervals, are given in Supplemental Table S1. Species and lineage abbreviations are as in Fig. 1.
[a]Density is measured as number per 1000 base pairs.

describing the same phenomenon in a simpler model). In contrast, we showed, using an exhaustive simulation study, that the reconstruction produced by the algorithm employed here is highly accurate (see Methods), and the accuracy improves when progressing from the homogeneous phase to the heterogeneous phase. On average, the relative error of the estimates was ~1% for the number of introns in ancestral forms (Supplemental Fig. S1), ~3% for the number of losses (Supplemental Fig. S2), and ~11% for the number of gains (Supplemental Fig. S2).

No significant variability of intron gain and loss rates within genes

The method outlined above was applied to 391 sets of orthologous genes from 19 eukaryotic species (Supplemental Fig. S3), a substantial extension of the eight-species data set developed by Rogozin et al. (2003) and employed in most of the subsequent studies on evolution of the exon–intron structure of eukaryotic genes (Csuros 2005; Nguyen et al. 2005; Roy and Gilbert 2005a,c). Intron positions were mapped on the multiple alignments of the analyzed genes as previously described (see Methods; Rogozin et al. 2003), and the resulting matrices of intron presence–absence were used to reconstruct the history of intron gain and loss during eukaryotic evolution, contingent on the phylogenetic tree topology.

We found that within-gene rate variability played no significant role in the current analysis. Genes were found to have a uniform distribution of intron loss rate throughout their length (the 95% confidence interval of λ_θ spans all permissible values). On average, 86% of the sites in each gene are incapable of gaining introns (ν = 0.86), in agreement with the proto-splice sites hypothesis (Dibb and Newman 1989; Dibb 1991; Sverdlov et al. 2004b) and with the previous estimates of Nguyen et al. (2005). The 14% of the sites where gain is tolerated also show uniform distribution of intron gain rate along the gene's length (the 95% confidence interval of λ_η spans all permissible values).

Reconstruction of intron density in ancestral forms: Intron-rich ancestors

For all nodes, we computed intron densities (Table 1) and their 95% confidence intervals (Fig. 1; Supplemental Table S1). Based on the results of the simulations (see Methods), we found these reconstructions to be highly accurate (Supplemental Fig. S1). Excluding the root of the tree (termed Eukaryota; see Supplemental Fig. S3), the average standard error was as low as ~1.1% (Supplemental Fig. S4). For ancestral forms younger than ~1.3 billion yr (a total of 13 nodes out of 18; see Supplemental Fig. S3), the average standard error was even lower, ~0.8%. The standard error of Eukaryota is considerably larger (18.9%), although the estimates remain highly informative (Fig. 1). It should be noted that ML estimations on the root of the phylogenetic tree are expected to have a higher variance than estimates for any of the internal nodes. For the simpler model of Roy and Gilbert (2005a), it has been shown that the number of introns in the root cannot be estimated by an ML technique (Nguyen et al. 2005). Technically, the present model allows for inference on that number, but its reliability is lower than for the other nodes. As expected, the estimates of intron density in ancestral eukaryotic forms obtained here fall in between those yielded by the Dollo parsimony approach (Rogozin et al. 2003) and several ML approaches (Csuros 2005; Nguyen et al. 2005), and those inferred from the hybrid ML/parsimony analysis of Roy and Gilbert (2005a,c) (Supplemental Table S2).

The present reconstruction indicates relatively high intron densities in ancient eukaryotic ancestors. Even taking a conservative stance and considering the lower bound of the 95% confidence interval, the last common ancestor of the eukaryotes studied here (Eukaryota) was unlikely to have <2.15 introns per kb of coding DNA; hence, its intron density was higher than that in modern insects and in most fungi (Fig. 1). This indicates that numerous introns have been gained prior to the divergence of the extant eukaryotic lineages. The optimal computed estimate is much higher, 3.19 introns per kb, suggesting an ancestor that is even richer in introns than the nematode *Caenorhabditis elegans*. Curiously, the inferred intron density of Eukaryota almost exactly coincides with the median of the distribution for all analyzed nodes. The last common ancestor of multicellular life, AME, is inferred to have been even more intron-rich, with an estimate of 3.39 introns per kb. Notably, among the top six intron-rich spe-

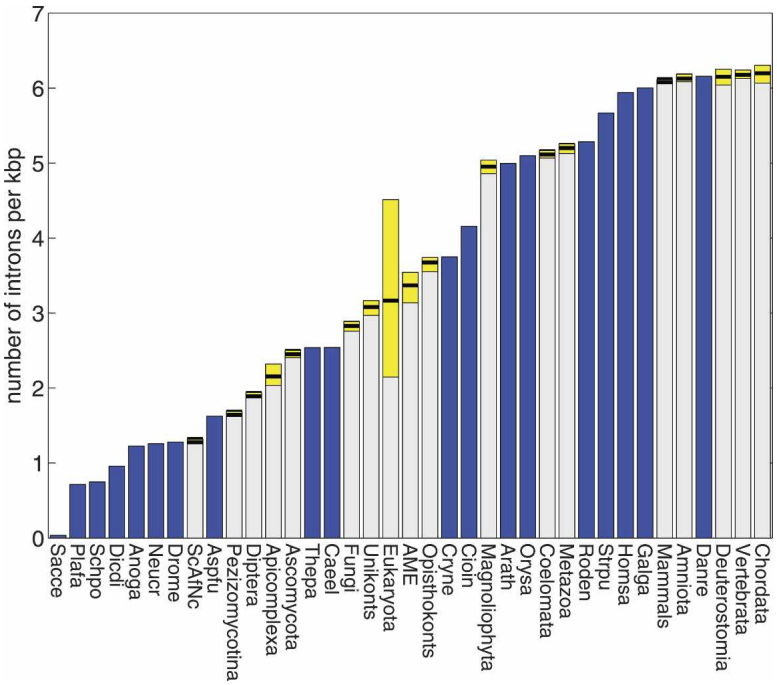

Figure 1. Intron density in extant species and ancestral forms. Densities are measured in introns per 1000 nucleotides. (Blue bars) The observed values for genes from extant species, (yellow bars) the 95% confidence intervals for the densities in ancestral nodes, (internal separator) the optimal value. Species and lineage abbreviations: (Caeel) *Caenorhabditis elegans*, (Strpu) *Strongylocentrotus purpuratus*, (Cioin) *Ciona intestinalis*, (Danre) *Danio rerio*, (Galga) *Gallus gallus*, (Homsa) *Homo sapiens*, (Roden) *Mus musculus* and *Rattus norvegicus* combined, (Drome) *Drosophila melanogaster*, (Anoga) *Anopheles gambiae*, (Cryne) *Cryptococcus neoformans*, (Schpo) *Schizosaccharomyces pombe*, (Sacce) *Saccharomyces cerevisiae*, (Aspfu) *Aspergillus fumigatus*, (Neucr) *Neurospora crassa*, (Arath) *Arabidopsis thaliana*, (Orysa) *Oryza sativa*, (Thepa) *Theileria parva*, (Plafa) *Plasmodium falciparum*, (Dicdi) *Dictyostelium discoideum*, (AME) Ancestor of Multicellular Eukaryotes.

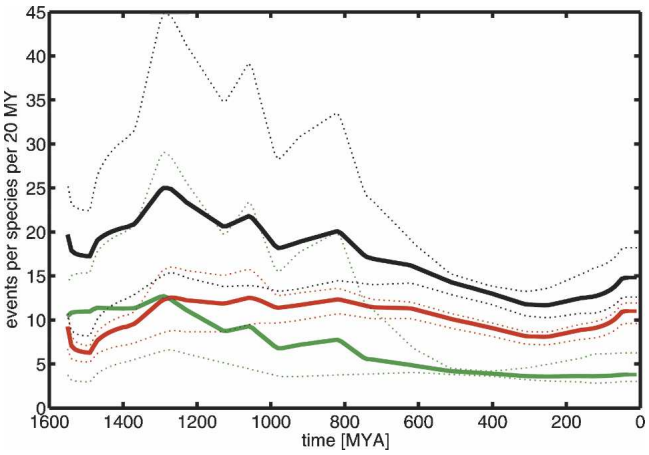

Figure 2. Time dependence of the overall number of intron gain and loss events during eukaryotic evolution. (Green lines) Number of gain events, (red lines) number of loss events, (black lines) total number of events per species per 20 million yr (Myr). Events were counted in a 20-Myr window every 10 Myr. The results were smoothed using the moving average algorithm with a 31-time-points window. (Dashed lines) Highly conservative confidence intervals, obtained by taking the lowest and highest values in the 95% confidence interval of each lineage.

cies, five are ancestral forms (Fig. 1), reflecting modern trends of excessive intron loss. More generally, in the graph of intron densities, the inferred values for ancestral forms intermingled with those for extant species (Fig. 1), emphasizing that a distribution of intron densities resembling that in modern genes was probably reached at an early stage of the evolution of eukaryotes.

Reconstruction of intron gain and loss densities: Ancient gains versus recent losses

Based on the simulation results, we found that the density (i.e., the number of events per 1000 nucleotides) of losses could be reconstructed with an average relative error of ~3%, whereas the density of gains could be reconstructed with an average relative error of ~11% (Supplemental Fig. S2). The density of inferred intron gain and loss events over the phylogenetic tree of eukaryotes reveals a complex pattern (Table 1; Supplemental Table S1). As discussed in the previous section, the number of introns in the root of the tree (Eukaryota) is estimated with a lower confidence than the numbers for the rest of the nodes. Consequently, gain and loss estimates along the branches stemming directly from the root (AME and Apicomplexa; Supplemental Fig. S3) have elevated error levels, too, and were excluded from the analysis. Overall, the present reconstruction suggests that both intron gains and intron losses played important roles in eukaryotic evolution, with some excess of loss. In total, we inferred 9410 losses and 5261 gains, i.e., an ~1.8-fold excess of losses. As with the intron density of the ancestral genes, these estimates fall in between the previously published gain-dominated (Qiu et al. 2004) and loss-dominated (Roy and Gilbert 2005a,c, 2006) scenarios of intron evolution. The current analysis suggests that, during the last ~1.5 billion yr of eukaryotic evolution, there were about twice as many intron losses as intron gains. However, such global counting is not particularly illuminating as lineages vastly differ in their gain and loss patterns, and, furthermore, these patterns are hardly uniform in time. In the following, we analyze the comparative contributions of intron gains and losses in different parts of the eukaryotic tree and, globally, as a function of time.

There is a growing body of evidence that intron loss and, especially, intron gain have been extremely rare in several eukaryotic lineages in the last ~100–200 million yr (Fedorov et al. 2003; Babenko et al. 2004; Roy and Hartl 2006; Roy and Penny 2006, 2007; Coulombe-Huntington and Majewski 2007). Ignoring for the time being the lineage-specific trends, simple averaging lends strong support to this conclusion (Fig. 2). It appears that, on average, introns maintained a high gain rate, presumably a continuation of their original proliferation that antedates the last common ancestor of current eukaryotes, until ~1.3 billion yr ago (Bya). Since then, the intron gain rate has been steadily decreasing down to the low level observed in recent history. While overshadowed by gains in ancient times, intron loss became the dominant process ~1.3 Bya, and since then showed only a mild decrease with time (Fig. 2). Interestingly, ~1.3 Bya, both processes showed high and comparable levels, which appears to approximately coincide with the time when the major eukaryotic lineages, such as metazoa and fungi, were radiating (Hedges et al. 2001).

Clearly, the relative contributions of intron gain and loss vary not only with time, but also among eukaryotic lineages. It is generally accepted that vertebrates have gained very few introns, if any (Fedorov et al. 2003; Babenko et al. 2004; Coulombe-Huntington and Majewski 2007). Nematodes are characterized by a high number of events, with losses being more plentiful than gains (Cho et al. 2004; Coghlan and Wolfe 2004). Fungi also show numerous events, with gains only slightly less numerous than losses (Nielsen et al. 2004). By contrast, few events have

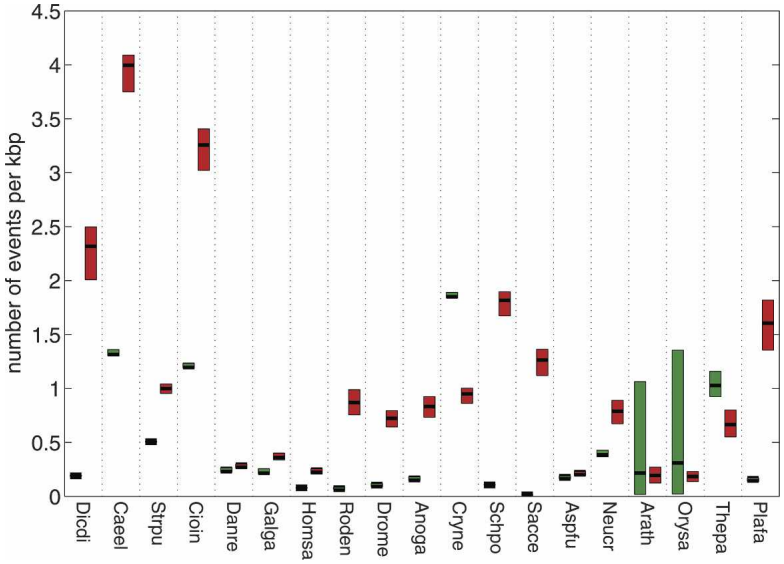

Figure 3. Density of intron gain and loss events in extant species. (Green bars) 95% confidence interval of gains; (red bars) 95% confidence interval of losses; (central black line) the optimal value. Species abbreviations are as in Fig. 1. Density is measured as number of events per 1000 nucleotides.

Table 2. Ratio of the number of intron losses to the number of intron gains in selected clades

Lineage	Intron loss/ gain ratio	*P*-value (compared with mean)
Vertebrata	3.01	1.69×10^{-9}
Metazoa	2.61	0.00
Fungi	1.99	8.31×10^{-3}
Magnoliophyta	0.73	6.34×10^{-14}
Apicomplexa	1.92	3.17×10^{-1}
Mean over the tree	1.79	—

been detected in Apicomplexa, with an excess of loss over gain (Roy and Hartl 2006). The trends of intron gain and loss in plants are less clear, with one study (Knowles and McLysaght 2006) finding gains to be 1.4 times more abundant than losses, and another (Roy and Penny 2007) reporting a dramatic excess of losses over gains (a loss-to-gain ratio of 12.6). In agreement with the trends over time (Fig. 2), the present analysis shows that, for most of the extant species, the total number of losses outnumbers the number of gains, even if the 95% confidence intervals are taken into account (Fig. 3). For 14 species, the number of losses was unequivocally greater (nonoverlapping 95% confidence intervals) than the number of gains (*Dictyostelium discoideum*, *C. elegans*, *Strongylocentrotus purpuratus*, *Ciona intestinalis*, *Gallus gallus*, *Homo sapiens*, Rodents, *Drosophila melanogaster*, *Anopheles gambiae*, *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae*, *Aspergillus fumigatus*, *Neurospora crassa*, and *Plasmodium falciparum*), and for three others, the confidence intervals of gain and loss overlapped (*Danio rerio*, *Arabidopsis thaliana*, and *Oryza sativa*). Only two species, namely a fungus (*Cryptococcus neoformans*) and an apicomplexan (*Theileria parva*), showed significantly more gains than losses.

At the clade level, there was a substantial excess of losses over gains in all clades except for plants, which showed an excess of gains (Table 2). Given the overall dominance of losses and the recent report on a dramatic dominance of intron losses in plants (Roy and Penny 2007), the latter finding was unexpected. However, because we analyzed only two plant genomes that are linked to the rest of the tree through a very long branch (~1.4 billion yr), this result should be interpreted with caution; in particular, it is possible that most of the gains occurred at an early stage of evolution, e.g., prior to the divergence of plants and green algae, whereas plant evolution per se still could be dominated by losses like the evolution of other eukaryotic lineages. Analysis of additional plant and algal genomes is required to definitely determine the trend of intron evolution in this lineage. In some cases, the clade-specific trends hide substantial within-clade heterogeneity. Thus, in the entire fungal clade, there was an approximately twofold excess of losses over gains (Table 2). However, at the species level, while most fungi, indeed, exhibit more losses (*S. pombe*, *S. cerevisiae*, *N. crassa*, and *A. fumigatus*), *C. neoformans* showed a clear excess of gains. These observations are in a good agreement with the findings of Nielsen et al. (2004) despite the fact that there was only one species in common to the two studies (*N. crassa*).

Decomposition of the contributions of the branches and the genes

The analysis in the previous section involved only the total number of events (or, equivalently, density) and disregarded branch lengths. While this is sufficient to allow a comparison of the numbers of gains and losses on the same branch, this approach is less suited for the purposes of conducting comparisons between species or lineages, as the results heavily depend on the specific tree topology and species sampling. The intrinsic tendency of a lineage to gain or lose an intron is captured by the event rates, i.e., the estimated number of gains or losses per unit time per site. In the present model, these rates are given in the form of the branch-specific gain and loss rates, Ξ_t and Φ_t, respectively. As indicated above, we developed an algorithm to estimate these parameters, as well as the gene-specific rates, directly from the highly robust reconstruction matrices P, A, and L. The parameters are estimated up to a multiplicative constant, and therefore no units were assigned to these rates. The detailed description of this algorithm is given in the accompanying paper (Carmel et al. 2007) where the gene-specific rates are analyzed. The finding pertinent to the analysis presented below is that the simulations proved the algorithm to be highly accurate in estimating the branch-specific parameters, having a correlation coefficient of 0.97 with the simulated parameters for loss rates, and 0.90 for gain rate (Supplemental Fig. S5).

Classification of eukaryotic lineages by intron gain and loss rates: A universal positive correlation between gains and losses

Different eukaryotic lineages show a wide range of intron gain and loss rates (Table 3). Using these rates, each branch was tested

Table 3. Intron gain and loss rates of individual branches

	Node	Gain rate	Loss rate
Balanced evolution	Coelomata	0	0
	Magnoliophyta	0.035	0.268
	Vertebrata	0.062	0.015
	Pezizomycotina	0.370	0.698
	Amniota	0	0
	Mammalia	0.000	0
	Strpu	0.013	0.043
	Danre	0.020	0.000
	Galga	0.034	0.083
	Homsa	0.210	0.352
	Cryne	0.033	0.242
	Aspfu	0.010	0.083
	Arath	0.088	0.087
	Orysa	0.132	0.032
	Thepa	0.011	0.214
	Chordata	0.446	0.262
Elevated loss rate	Unikonts	0.191	0.946
	Diptera	0.045	2.028
	Fungi	0.222	1.575
	ScAfNc	0.000	1.985
	Dicdi	0.003	0.653
	Caeel	0.024	1.048
	Cioin	0.044	0.773
	Drome	0.011	1.252
	Anoga	0.018	1.333
	Schpo	0.002	1.099
	Sacce	0.001	4.289
	Neucr	0.035	0.780
	Plafa	0.003	0.949
	Roden	0.222	1.963
Elevated gain rate	Opisthokonts	0.907	0.114
	Metazoa	0.516	0.351
	Deuterostomia	5.919	0.457
Dynamic evolution[a]	Ascomycota	4.944	5.401

The values are for the tree topology in Supplemental Fig. S3. Species and lineage abbreviations are as in Fig. 1.
[a]This term means that this lineage shows elevated rates of both gains and losses.

to detect those that had a statistically significant excess of gains or losses over the respective mean rates across the phylogenetic tree (see Methods). The lineages were partitioned into three clusters: (1) those with predominant intron loss, (2) those with predominant intron gain, and (3) balanced, with both gain and loss rates not significantly greater than the mean (Fig. 4; Table 3). Only one lineage (Ascomycota) had both gain and loss rates significantly above the mean. Technically, it could have been included in the balanced cluster, but we preferred putting it in a cluster of its own. The revealed evolutionary landscape, now based on gain and loss rates rather than absolute numbers of events, is generally consistent with the results presented above as well as previous reports. For instance, the gene structure of vertebrates is remarkably stable, whereas fungi, *D. discoideum*, and insects show high loss rates. However, although *C. neoformans*, *T. parva*, and, notably, plants show overall excess in number of gain events (Fig. 3), they are classified in the balanced cluster because, when the branch lengths are taken into account, their gain rates are not significantly elevated above the mean (Fig. 4; Table 3).

Also in agreement with the results presented earlier (Fig. 2), extensive intron loss seems to have occurred in several lineages relatively recently such that all extant species are classified in either the balanced cluster or in the elevated loss cluster. In a sharp contrast, all episodes of massive intron gain dominating over losses are ancient (Fig. 4; see also Fig. 2). Specifically, lineages leading to animals seem to have experienced a phase of massive intron invasion early in their evolution (Fig. 4; Table 3). The inferred pattern of intron gain and loss did not show a strong dependence on the topology of the phylogenetic tree of eukaryotes, as becomes evident from the comparison of the scenarios for alternative topologies (Supplemental Fig. S6A–C; Supplemental Table S3A–C).

Having developed this classification of eukaryotic lineages, we can directly address the issue of the sign of the correlation between lineage-specific intron gain and loss rates. Population-genetic reasoning suggests that these rates should be inversely related (Lynch 2002; Lynch and Conery 2003), a prediction that appears to have been supported by at least two independent analyses (Nguyen et al. 2005; Roy and Gilbert 2005c). The present results reveal a more complex pattern of dependencies and effectively refute the prediction. Taking all the lineages together or selected subsets of interest, no significant correlation was observed between lineage-specific intron gain and loss rates (Supplemental Table S4). However, when all the lineages are plotted on a two-dimensional plane spanned by the intron gain and loss rates, a striking pattern becomes apparent: The classification of lineages into balanced ones, those with an elevated loss rate, and those with an elevated gain rate divides the plane into three well-separated regions (Fig. 5). The large cluster with balanced evolution includes almost half of the lineages (16/34), and its gain and loss rates are significantly and positively correlated (Spearman correlation coefficient of 0.69; $P = 0.003$; Fig. 5). Thus, this balanced mode of evolution is characterized by roughly proportional gain and loss rates. It should be emphasized that, in this case, balance does not mean equilibrium; i.e., the rates of intron gain and loss are approximately proportional, but, taken together with the number of sites available for gain or loss, this does not translate into a prediction of stasis with respect to the number of introns. Indeed, some of the balanced lineages, e.g., plants, have gained many more introns than they have lost, whereas others, e.g., sea urchin (Strpu), appear to have lost considerably more introns than they have gained (Table 1).

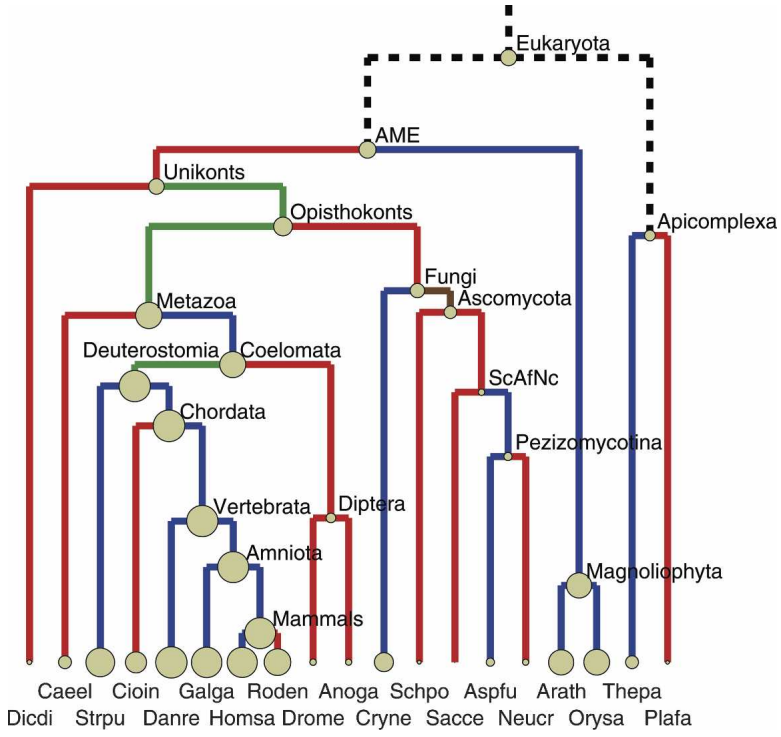

Figure 4. Distribution of intron gain and loss rates over the phylogenetic tree of eukaryotes. Node sizes are proportional to their (known or inferred) intron density, and the branches are color-coded: (green) predominant intron gain; (red) predominant intron loss; (blue) balanced gain and loss. The sole brown branch (Ascomycota) designates extensive (significantly greater than the mean over the tree) gains and losses. Species and lineage abbreviations are as in Fig. 1.

The other two clusters encompass lineages where either gain or loss became dominant. The gain rates of the loss-dominated lineages were statistically indistinguishable from the gain rates of balanced lineages (*t*-test; $P = 0.44$), and, similarly, the loss rates of the gain-dominated lineages were indistinguishable from the loss rates of balanced lineages (*t*-test; $P = 0.19$). This strongly suggests that the balanced mode of intron evolution is in operation in all eukaryotic lineages and forms the universal basis of intron dynamics. In this mode, gain and loss are tightly linked, implying the existence of common mechanistic components in these processes. Such a commonality has been proposed previously in the form of reverse-transcription-mediated mechanisms for both intron loss and intron gain (Sverdlov et al. 2004a).

The extensive intron loss in some lineages and, especially, the less common bursts of intron gain might involve additional mechanisms or, alternatively,

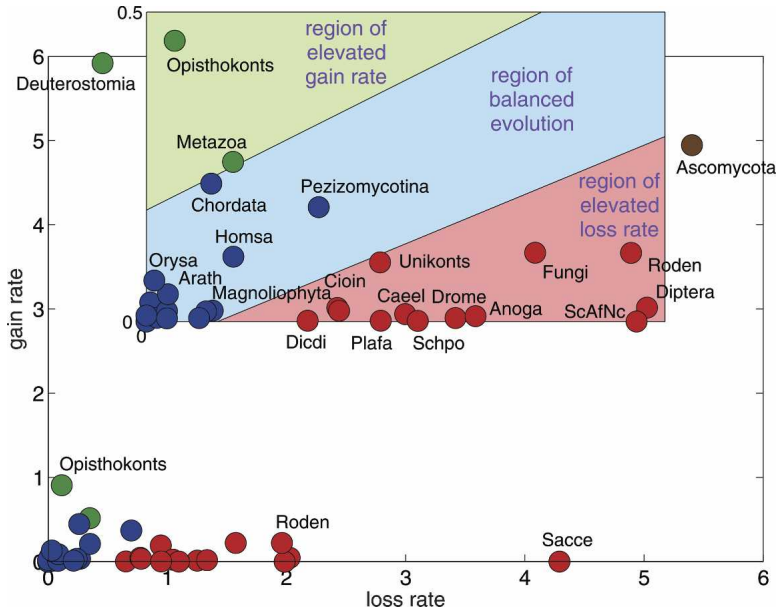

Figure 5. Intron gain and loss rates of eukaryotic lineages. (Blue) Lineages exhibiting the balanced mode of evolution; (red) lineages with elevated loss rate; (green) lineages with elevated gain rate. The brown Ascomycota indicates the only lineage with both the gain rate and loss rate elevated. (*Inset*) An expanded view of the low-rate area that was obtained by excluding three lineages: Deuterostomia, Ascomycota, and *S. cerevisiae*. Selected lineages are labeled. Species and lineage abbreviations are as in Fig. 1.

could be explained by differences in the strength of purifying selection that affects the evolution of the respective lineages. Indeed, the apparent association of massive intron gain with the emergence of major lineages of eukaryotes appears compatible with the population-genetic perspective on evolution of eukaryotic gene structure whereby new introns can be fixed by drift during population bottlenecks (Lynch and Conery 2003; Lynch 2006). This supposedly predominant, neutral mode of intron gain does not rule out the possibility that some of the new introns assume functions that contribute to the increasing organizational complexity in the respective lineages. Indeed, introns can affect gene expression at several levels (Mattick 1994; Long and De Souza 1998; Maniatis and Reed 2002; see also the accompanying paper in this issue, Carmel et al. 2007, and references therein).

Conclusions

The combination of an expanded data set and a comprehensive model of evolution employed here yielded a more nuanced picture of intron evolution in eukaryotes than was previously suspected. The results suggest that relatively high intron density was reached early in the history of eukaryotes; specifically, the root of the tree is inferred to have >2.15 introns per kb, and the last common ancestor of multicellular life is deduced to have contained ~3.39 introns per kb, a greater intron density than is seen in most of the extant fungi and some animals. Both intron gain and intron loss occurred extensively during the subsequent evolution, with some excess of losses (the ratio of losses to gains is ~1.8). The same excess of losses is observed in most individual clades, except for plants, which show more intron gains than losses. On the evolutionary time scale, the rates of both intron gains and intron loss seem to have been decreasing during the last ~1.3 By, with the drop in the gain rate being much steeper. The few inferred episodes of excessive intron gain are ancient, and seem to be associated with major events in eukaryotic evolution, such as the origin of animals. It is conceivable that such major evolutionary events were associated with severe population bottlenecks, resulting in weakened purifying selection and permitting intron proliferation (Lynch and Conery 2003; Lynch 2006). What the contribution, if any, of the new introns was to the increasing organizational complexity at these evolutionary crossroads remains an intriguing question (also see the accompanying paper, Carmel et al. 2007). Aside from the episodes of extensive intron loss and gain, evolution of eukaryotic genes seems to be dominated by the balanced dynamics of introns, where the rates of gain and loss are roughly proportional. This implies mechanistic similarities between these processes and is compatible with reverse transcription as a common underlying mechanism (Sverdlov et al. 2004a). The present results suggest that this mode of evolution operates in all eukaryotic lineages, with additional, perhaps mechanistically distinct loss or gain components in some of the lineages.

Methods

The data set

Using the KOG database and the KOGNITOR program (Tatusov et al. 2003), we identified 400 sets of orthologous genes from 19 eukaryotic species: nine metazoans (*Caenorhabditis elegans*, *Strongylocentrotus purpuratus*, *Ciona intestinalis*, *Danio rerio*, *Gallus gallus*, *Homo sapiens*, rodents [*Mus musculus* and *Rattus norvegicus* combined], *Drosophila melanogaster*, *Anopheles gambiae*); five fungi (*Cryptococcus neoformans*, *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae*, *Aspergillus fumigatus*, *Neurospora crassa*); two plants (*Arabidopsis thaliana*, *Oryza sativa*); two apicomplexans (*Theileria parva*, *Plasmodium falciparum*); and the protist *Dictyostelium discoideum*. For each KOG, we used the MUSCLE program (Edgar 2004) to compute a multiple alignment, upon which the intron positions were projected to form a binary presence/absence map (Rogozin et al. 2003). The raw data file raw_data.zip is available from ftp://ftp.ncbi.nih.gov/pub/koonin/carmel_introns/.

These maps were scanned, both automatically (see log.cdata.txt at ftp://ftp. ncbi.nih.gov/pub/koonin/carmel_introns/) and manually (see log.mcdata.txt at ftp://ftp.ncbi.nih.gov/pub/koonin/carmel_introns/), to fix annotation errors in the intron–exon boundaries. Intron positions shifted by 1 bp were regarded as cases of intron sliding (Rogozin et al. 2000), and were merged (see log.isdata.txt at ftp://ftp.ncbi.nih.gov/pub/koonin/carmel_introns/). Not only is the mere existence of intron sliding questionable, but even if it is happening, it does not necessarily explain every shift in one nucleotide, as some are simply due to chance. Therefore, we have generated another version of the data, where the positions that are 1 bp apart were not merged. The results have not changed in any significant way (see log.

fdata_alt.txt at ftp://ftp.ncbi.nih.gov/pub/koonin/carmel_introns/; Supplemental Fig. S7). Six KOGs with particularly poor annotations were removed (KOG0337, KOG1302, KOG2280, KOG1985, KOG1234, and KOG1122).

Stringent filtering was applied to ensure that only highly reliable portions of the alignments were used for further analysis (Rogozin et al. 2003; see log.fdata.txt at ftp://ftp.ncbi.nih.gov/pub/koonin/carmel_introns/). Three KOGs were removed due to poor alignment (KOG2005, KOG2180, and KOG2851). The final data set used for inferring intron loss and gain consisted of the reliable portions of 19-species alignments for 391 KOGs, which included 289,902 sites in total (the data is in the file final_data.zip, available from ftp://ftp.ncbi.nih.gov/pub/koonin/carmel_introns/); of these 5755 (2%) are intron-bearing sites.

Phylogenetic tree of eukaryotes

The deep branching order of the eukaryotic phylogenetic tree remains uncertain (Baldauf 2003). The model was applied to four alternative tree topologies (Supplemental Figs. S3, S6A–C). In the more traditional topology (Hedges 2002), the root position is between Apicomplexa and the common ancestor of multicellular eukaryotes (plants and animals), whereas in the unikont–bikont topology, the root is between the unikont and bikont clades; i.e., the last common ancestor of plants and animals is the same as the last common ancestor of eukaryotes (Stechmann and Cavalier-Smith 2002). Each of these two topologies was used in two versions, Ecdysozoa and Coelomata, in order to account for a major unresolved issue in animal phylogeny (Aguinaldo et al. 1997; Blair et al. 2002; Wolf et al. 2004). The divergence time estimates for the main eukaryotic clades are given in Supplemental Table S5 (see the caption to Table S5 for the sources and methods).

The expectation–maximization algorithm

Phylogenetic trees can be interpreted as Bayesian networks that depict an underlying evolutionary probabilistic model. Accordingly, the terminal nodes are the observed random variables of the model, and the internal nodes are the hidden random variables. We then estimate the parameters of this model using ML with an expectation–maximization (EM) algorithm (Dempster et al. 1977). Several EM algorithms have been applied to phylogenetic tree analysis with various purposes (Friedman et al. 2002; Siepel and Haussler 2004; Holmes 2005). However, the present model does not fit into any of the existing EM schemes as it includes unique properties, such as the branch-specific coefficients, the intronless sites, and the different treatment of rate variability across sites. Thus, we developed an EM algorithm that allows for estimating the entire set of parameters, the number of introns in internal nodes, and the number of loss and gain events along each branch. A slightly simplified version of this algorithm has been described previously (Carmel et al. 2005). There, we forced equality between the shape parameters of the loss and gain rate distributions, $\lambda_\eta = \lambda_\theta$. In this work, this restriction was removed by a trivial modification of the original algorithm.

Simulation analysis

We performed a series of 100 simulations that served both to validate the algorithm and to derive confidence intervals for the inferences. In each simulation, we used the same 19-species phylogenetic tree that was used in the analysis of the real data (Supplemental Fig. S3; Supplemental Table S5), the same number of genes (391) as in the real data, and the same gene lengths (i.e., number of sites) as in our real data. Then, random model parameters were drawn from distributions chosen such that the characteristics of the simulated data resemble those of the real data (Supplemental Table S6). Specifically, we counted the total number of introns in extant species, and the total number of unique presence/absence patterns (Supplemental Fig. S8). For all files generated during the simulation phase, see simulations.zip at ftp://ftp.ncbi.nih.gov/pub/koonin/carmel_introns/.

The computation stops when the likelihood convergence rate reaches some predefined tolerance. Each simulation was run in four different convergence tolerances, 10^{-7}, 10^{-8}, 10^{-9}, and 10^{-10}. To estimate parameters for the real data, we used a tolerance of 10^{-11}, but, to save time, such tight tolerance was not applied to the simulations. It was found that high accuracy is achieved already at the tolerance of 10^{-7}, with slight improvement, if any, in tighter tolerances (Supplemental Figs. S1, S2, S5). The average running time for a single simulation (all four tolerances) was 2 h (on a Pentium 3-GHz machine).

Estimation of the number of introns in ancestral nodes

For reasons that remain unclear, the accuracy of the reconstructions after the homogeneous phase drops with tighter tolerances. Although the homogeneous phase suffices to obtain reliable reconstructions, the heterogeneous phase improves the accuracy by roughly a factor of two (Supplemental Fig. S1). Overall, the relative error of the reconstruction is ~1%. Taking the average, over the simulations, of the relative error for each node, the (not necessarily symmetric) 95% confidence interval of the estimates was determined (Supplemental Table S1).

Estimation of the number of intron gain and loss events

Similarly to estimating the number of introns, the heterogeneous phase improves the accuracy of the reconstructions (Supplemental Fig. S2). The relative error stays at approximately the same level for all tolerance levels. The errors in estimating gains are higher (~11%) than in estimating losses (~3%), probably due to the smaller number of gain events. Again, taking the average (over simulations) relative error for each branch allows us to find the (not necessarily symmetric) 95% confidence interval of the estimates (Supplemental Table S1).

Estimation of the branch-specific intron gain and loss rates

For the purpose of estimating gain rates, the heterogeneous phase adds little accuracy, but for the loss rates, the improvement is substantial (Supplemental Fig. S5). Overall, the estimated loss rates have a mean correlation coefficient of 0.97 with the simulated ones, and the estimated gain rates have a mean correlation coefficient of 0.90 with the simulated one.

Lineage classification

The eukaryotic lineages were classified into the three modalities of intron evolution: balanced, elevated loss, and elevated gain. Let $L(g)$ be the length (number of sites) of the multiple alignment of gene g, and let R be the set of all nodes, excluding the root and its two direct descendants. For a gene g along a branch t, the number of sites capable of gaining introns is $S_G(t,g) = L(g) - P(t^P,g) + \frac{1}{2}L(t,p)$. The last term (which is negligibly small in most cases) accounts for sites that hosted an intron at the beginning of the branch, but later lost it, and are therefore capable of regaining an intron. We can measure the "average," or typical, branch-specific gain rate as

$$\overline{\overline{\Xi}} = -\frac{1}{\Delta}\log(1 - P_G),$$

where $\bar{\Delta}$ is the average branch length, and

$$P_G = \frac{\sum_{t\in R}\sum_{g} A(t,g)}{\sum_{t\in R}\sum_{g} S_G(t,g)}$$

is a measure of an "average" gain probability per site in an "average" branch.

Similarly, the number of sites capable of losing introns is $S_L(t,g) = P(t^P,g)$. We can measure the "average," or typical, branch-specific loss rate as

$$\bar{\Phi} = -\frac{1}{\bar{\Delta}}\log(1 - P_L),$$

where

$$P_L = \frac{\sum_{t\in R}\sum_{g} L(t,g)}{\sum_{t\in R}\sum_{g} S_L(t,g)}.$$

Next, for each branch, we iterate through all genes, and compute for each the expected number of events (based on the rates $\bar{\Xi}$ and $\bar{\Phi}$). We sum these numbers and get a total expected number of events per branch, say $E_G(t)$ and $E_L(t)$ for gain and loss, respectively. Then, we compare the fraction of observed events,

$$\frac{\sum_{g} G(t,g)}{\sum_{g} S_G(t,g)} \quad \text{and} \quad \frac{\sum_{g} L(t,g)}{\sum_{g} S_L(t,g)}$$

with the expectations

$$\frac{E_G(t)}{\sum_{g} S_G(t,g)} \quad \text{and} \quad \frac{E_L(t)}{\sum_{g} S_L(t,g)},$$

and pick those lineages for which we can confidently (here, Bonferroni corrected P-value of 0.01) reject equality.

Acknowledgments

This work was supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health/DHHS.

References

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.

Babenko, V.N., Rogozin, I.B., Mekhedov, S.L., and Koonin, E.V. 2004. Prevalence of intron gain over intron loss in the evolution of paralogous gene families. *Nucleic Acids Res.* **32:** 3724–3733.

Baldauf, S.L. 2003. The deep roots of eukaryotes. *Science* **300:** 1703–1706.

Blair, J.E., Ikeo, K., Gojobori, T., and Hedges, S.B. 2002. The evolutionary position of nematodes. *BMC Evol. Biol.* **2:** 7.

Carmel, L., Rogozin, I.B., Wolf, Y.I., and Koonin, E.V. 2005. An expectation-maximization algorithm for analysis of evolution of exon–intron structure of eukaryotic genes. *Comparative Genomics Lect. Notes Comput. Sci.* **3678:** 35–46.

Carmel, L., Rogozin, I.B., Wolf, Y.I., and Koonin, E.V. 2007. Evolutionarily conserved genes preferentially accumulate introns. *Genome Res.*, (this issue) doi: 10.1101/gr.5978207.

Cho, S., Jin, S.W., Cohen, A., and Ellis, R.E. 2004. A phylogeny of *Caenorhabditis* reveals frequent loss of introns during nematode evolution. *Genome Res.* **14:** 1207–1220.

Coghlan, A. and Wolfe, K.H. 2004. Origins of recently gained introns in *Caenorhabditis*. *Proc. Natl. Acad. Sci.* **101:** 11362–11367.

Collins, L. and Penny, D. 2005. Complex spliceosomal organization ancestral to extant eukaryotes. *Mol. Biol. Evol.* **22:** 1053–1066.

Coulombe-Huntington, J. and Majewski, J. 2007. Characterization of intron loss events in mammals. *Genome Res.* **17:** 23–32.

Csuros, M. 2005. Likely scenarios of intron evolution. *Comparative Genomics. Lect. Notes Comput. Sci.* **3678:** 47–60.

Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Methodol.* **39:** 1–38.

Deutsch, M. and Long, M. 1999. Intron–exon structures of eukaryotic model organisms. *Nucleic Acids Res.* **27:** 3219–3228.

Dibb, N.J. 1991. Proto-splice site model of intron origin. *J. Theor. Biol.* **151:** 405–416.

Dibb, N.J. and Newman, A.J. 1989. Evidence that introns arose at proto-splice sites. *EMBO J.* **8:** 2015–2021.

Doolittle, W.F. 1978. Genes in pieces: Were they ever together? *Nature* **272:** 581–582.

Edgar, R.C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32:** 1792–1797.

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.* **99:** 16128–16133.

Fedorov, A., Roy, S., Fedorova, L., and Gilbert, W. 2003. Mystery of intron gain. *Genome Res.* **13:** 2236–2241.

Felsenstein, J. 2004. *Inferring phylogenies*. Sinauer Associates, Sunderland, MA.

Friedman, N., Ninio, M., Pe'er, I., and Pupko, T. 2002. A structural EM algorithm for phylogenetic inference. *J. Comput. Biol.* **9:** 331–353.

Gilbert, W. 1978. Why genes in pieces? *Nature* **271:** 501.

Hedges, S.B. 2002. The origin and evolution of model organisms. *Nat. Rev. Genet.* **3:** 838–849.

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.

Holmes, I. 2005. Using evolutionary expectation maximization to estimate indel rates. *Bioinformatics* **21:** 2294–2300.

Jeffares, D.C., Mourier, T., and Penny, D. 2006. The biology of intron gain and loss. *Trends Genet.* **22:** 16–22.

Knowles, D.G. and McLysaght, A. 2006. High rate of recent intron gain and loss in simultaneously duplicated *Arabidopsis* genes. *Mol. Biol. Evol.* **23:** 1548–1557.

Llopart, A., Comeron, J.M., Brunet, F.G., Lachaise, D., and Long, M. 2002. Intron presence–absence polymorphism in *Drosophila* driven by positive Darwinian selection. *Proc. Natl. Acad. Sci.* **99:** 8121–8126.

Logsdon Jr., J.M. 1998. The recent origins of spliceosomal introns revisited. *Curr. Opin. Genet. Dev.* **8:** 637–648.

Long, M. and De Souza, S.J. 1998. Intron–exon structures: From molecular to population biology. In *Advances in genome biology: Genes and genomes* (ed. R.S. Verma), Vol. 5A, pp. 143–178. JIA Press, Greenwich, CT.

Lynch, M. 2002. Intron evolution as a population-genetic process. *Proc. Natl. Acad. Sci.* **99:** 6118–6123.

Lynch, M. 2006. The origins of eukaryotic gene structure. *Mol. Biol. Evol.* **23:** 450–468.

Lynch, M. and Conery, J.S. 2003. The origins of genome complexity. *Science* **302:** 1401–1404.

Lynch, M. and Richardson, A.O. 2002. The evolution of spliceosomal introns. *Curr. Opin. Genet. Dev.* **12:** 701–710.

Maniatis, T. and Reed, R. 2002. An extensive network of coupling among gene expression machines. *Nature* **416:** 499–506.

Martin, W. and Koonin, E.V. 2006. Introns and the origin of nucleus–cytosol compartmentalization. *Nature* **440:** 41–45.

Mattick, J.S. 1994. Introns: Evolution and function. *Curr. Opin. Genet. Dev.* **4:** 823–831.

Mourier, T. and Jeffares, D.C. 2003. Eukaryotic intron loss. *Science* **300:** 1393.

Nguyen, H.D., Yoshihama, M., and Kenmochi, N. 2005. New maximum likelihood estimators for eukaryotic intron evolution. *PLoS Comput. Biol.* doi: 10.1371/journal.pcbi.0010079.

Nielsen, C.B., Friedman, B., Birren, B., Burge, C.B., and Galagan, J.E. 2004. Patterns of intron gain and loss in fungi. *PLoS Biol.* doi: 10.1371/journal.pbio.0020422.

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.* **99:** 3701–3705.

Qiu, W.G., Schisler, N., and Stoltzfus, A. 2004. The evolutionary gain of spliceosomal introns: Sequence and phase preferences. *Mol. Biol. Evol.* **21:** 1252–1263.

Raible, F., Tessmar-Raible, K., Osoegawa, K., Wincker, P., Jubin, C.,

Balavoine, G., Ferrier, D., Benes, V., de Jong, P., Weissenbach, J., et al. 2005. Vertebrate-type intron-rich genes in the marine annelid *Platynereis dumerilii*. *Science* **310:** 1325–1326.
Rodriguez-Trelles, F., Tarro, R., and Ayala, F.J. 2006. Origins and evolution of spliceosomal introns. *Annu. Rev. Genet.* **40:** 47–76.
Rogozin, I.B., Lyons-Weiler, J., and Koonin, E.V. 2000. Intron sliding in conserved gene families. *Trends Genet.* **16:** 430–432.
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.
Rogozin, I.B., Babenko, V.N., Wolf, Y.I., and Koonin, E.V. 2005a. Dollo parsimony and reconstruction of genome evolution. In *Parsimony, phylogeny, and genomics* (ed. V.A. Albert), pp. 190–200. Oxford University Press, Oxford.
Rogozin, I.B., Sverdlov, A.V., Babenko, V.N., and Koonin, E.V. 2005b. Analysis of evolution of exon–intron structure of eukaryotic genes. *Brief. Bioinform.* **6:** 118–134.
Roy, S.W. and Gilbert, W. 2005a. Complex early genes. *Proc. Natl. Acad. Sci.* **102:** 1986–1991.
Roy, S.W. and Gilbert, W. 2005b. The pattern of intron loss. *Proc. Natl. Acad. Sci.* **102:** 713–718.
Roy, S.W. and Gilbert, W. 2005c. Rates of intron loss and gain: Implications for early eukaryotic evolution. *Proc. Natl. Acad. Sci.* **102:** 5773–5778.
Roy, S.W. and Gilbert, W. 2006. The evolution of spliceosomal introns: Patterns, puzzles and progress. *Nat. Rev. Genet.* **7:** 211–221.
Roy, S.W. and Hartl, D.L. 2006. Very little intron loss/gain in *Plasmodium*: Intron loss/gain mutation rates and intron number. *Genome Res.* **16:** 750–756.
Roy, S.W. and Penny, D. 2006. Smoke without fire: Most reported cases of intron gain in nematodes instead reflect intron losses. *Mol. Biol. Evol.* **23:** 2259–2262.
Roy, S.W. and Penny, D. 2007. Patterns of intron loss and gain in plants: Intron loss-dominated evolution and genome-wide comparison of *O. sativa* and *A. thaliana*. *Mol. Biol. Evol.* **24:** 171–181.
Siepel, A. and Haussler, D. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* **21:** 468–488.
Simpson, A.G., MacQuarrie, E.K., and Roger, A.J. 2002. Eukaryotic evolution: Early origin of canonical introns. *Nature* **419:** 270.
Stechmann, A. and Cavalier-Smith, T. 2002. Rooting the eukaryote tree by using a derived gene fusion. *Science* **297:** 89–91.
Sverdlov, A.V., Babenko, V.N., Rogozin, I.B., and Koonin, E.V. 2004a. Preferential loss and gain of introns in 3′ portions of genes suggests a reverse-transcription mechanism of intron insertion. *Gene* **338:** 85–91.
Sverdlov, A.V., Rogozin, I.B., Babenko, V.N., and Koonin, E.V. 2004b. Reconstruction of ancestral protosplice sites. *Curr. Biol.* **14:** 1505–1508.
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.
Vanacova, S., Yan, W., Carlton, J.M., and Johnson, P.J. 2005. Spliceosomal introns in the deep-branching eukaryote *Trichomonas vaginalis*. *Proc. Natl. Acad. Sci.* **102:** 4430–4435.
Wolf, Y.I., Rogozin, I.B., and Koonin, E.V. 2004. Coelomata and not Ecdysozoa: Evidence from genome-wide phylogenetic analysis. *Genome Res.* **14:** 29–36.
Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.* **39:** 306–314.

Received February 26, 2007; accepted in revised form March 28, 2007.