

REVIEW

Open Access

Origin and evolution of spliceosomal introns

Igor B Rogozin¹, Liran Carmel², Miklos Csuros³ and Eugene V Koonin^{1*}

Abstract

Evolution of exon-intron structure of eukaryotic genes has been a matter of long-standing, intensive debate. The introns-early concept, later rebranded 'introns first' held that protein-coding genes were interrupted by numerous introns even at the earliest stages of life's evolution and that introns played a major role in the origin of proteins by facilitating recombination of sequences coding for small protein/peptide modules. The introns-late concept held that introns emerged only in eukaryotes and new introns have been accumulating continuously throughout eukaryotic evolution. Analysis of orthologous genes from completely sequenced eukaryotic genomes revealed numerous shared intron positions in orthologous genes from animals and plants and even between animals, plants and protists, suggesting that many ancestral introns have persisted since the last eukaryotic common ancestor (LECA). Reconstructions of intron gain and loss using the growing collection of genomes of diverse eukaryotes and increasingly advanced probabilistic models convincingly show that the LECA and the ancestors of each eukaryotic supergroup had intron-rich genes, with intron densities comparable to those in the most intron-rich modern genomes such as those of vertebrates. The subsequent evolution in most lineages of eukaryotes involved primarily loss of introns, with only a few episodes of substantial intron gain that might have accompanied major evolutionary innovations such as the origin of metazoa. The original invasion of self-splicing Group II introns, presumably originating from the mitochondrial endosymbiont, into the genome of the emerging eukaryote might have been a key factor of eukaryogenesis that in particular triggered the origin of endomembranes and the nucleus. Conversely, splicing errors gave rise to alternative splicing, a major contribution to the biological complexity of multicellular eukaryotes. There is no indication that any prokaryote has ever possessed a spliceosome or introns in protein-coding genes, other than relatively rare mobile self-splicing introns. Thus, the introns-first scenario is not supported by any evidence but exon-intron structure of protein-coding genes appears to have evolved concomitantly with the eukaryotic cell, and introns were a major factor of evolution throughout the history of eukaryotes. This article was reviewed by I. King Jordan, Manuel Irimia (nominated by Anthony Poole), Tobias Mourier (nominated by Anthony Poole), and Fyodor Kondrashov. For the complete reports, see the Reviewers' Reports section.

Keywords: Intron sliding, Intron gain, Intron loss, Spliceosome, Splicing signals, Evolution of exon/intron structure, Alternative splicing, Phylogenetic trees, Mobile domains, Eukaryotic ancestor

Genes in pieces: exon-intron structure of eukaryotic genes and the two spliceosomes

In a memorable phrase of Walter Gilbert, eukaryotes possess "genes in pieces" in which protein-coding sequences are interrupted by non-coding sequences denoted introns [1]. The introns are excised at the donor and acceptor splice sites such that the flanking coding regions, exons, are spliced by an extremely complex ribonucleoprotein

molecular machine, the spliceosome [2,3]. Multiple introns interrupt the coding sequences in the great majority of genes in animals and plants, whereas intron densities in fungi and unicellular eukaryotes are highly variable: many of the unicellular forms contain only a few introns in the entire genome whereas in others the intron density approaches that in animals and plants [4-6]. Remarkably, however, there is no sequenced genome of a full-fledged eukaryote without introns at all; only one intronless genome of a highly degraded remnant of a eukaryotic organism, a nucleomorph that has also lost the genes for the spliceosome subunits, has been reported [7].

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

¹National Center for Biotechnology Information NLM/NIH, 8600 Rockville Pike, Bldg. 38A, Bethesda, MD 20894, USA

Full list of author information is available at the end of the article

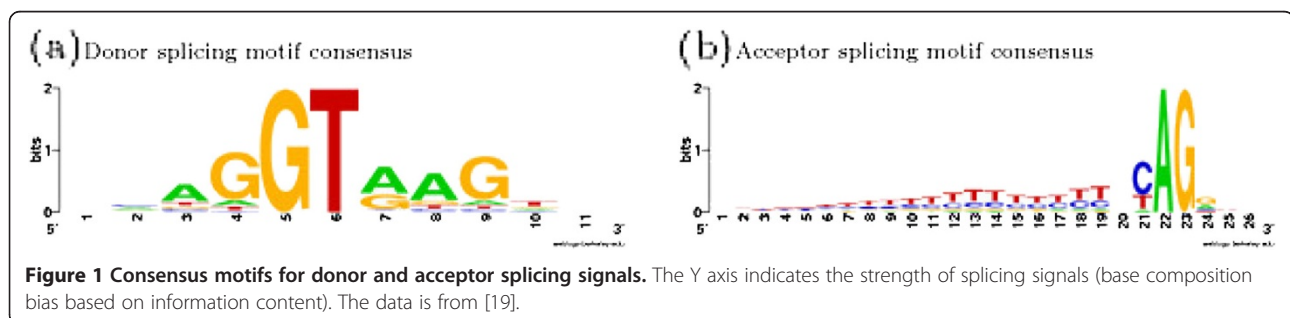
The ubiquity of introns in eukaryotes is complemented by the conservation of the spliceosome. The spliceosome consists of five snRNPs (small nuclear ribonucleoprotein particles), together with numerous less stably associated proteins; the core of the spliceosome is conserved in all well-characterized eukaryotes [2,3,8]. The spliceosome interacts with specific sites in the intron and the flanking exons to ensure accurate and efficient splicing. The nucleotides at the intron termini and the adjacent nucleotides in the exons are involved in these interactions and comprise the splicing signals. The (A/C)AG|GU(A/G)AGU sequence (the splice site is shown by the vertical streak and the first two nucleotides of the intron are underlined) at the donor splice signal is complementary to the 5' end of the U1 snRNA, and this interaction appears to be the major requirement for splicing [9-11]. The (C,U)AG|G sequence (the last two nucleotides of the intron are underlined) preceded by a polypyrimidine tract is typical of the acceptor splice signal (Figure 1) and is recognized by the U5 snRNA [12,13]. A short branch point signal is located in the intron sequence upstream of the acceptor splice signals and contains the reactive adenosine that is involved in the formation of the lariat-like structure in the splicing intermediate [12,13]. The functionally important (A/C)AG||G exon sequences flanking introns have been dubbed protosplice sites with the implication that new introns insert into sites of this structure [14,15]. Some lineage-specific deviations from the canonical variants of splice signals are known to exist. For example, some unicellular eukaryotes lack recognizable polyT tracts between the branch point signal and the 3' splice signal [16,17]. Some extremely intron-poor species such as yeast possess an unusual, strictly constrained donor splice signal |GTA(T,A,C)G(T,A,C) with a substantial excess of T at position +4 [16-18].

The vast majority of spliceosomal introns contain |GT at the donor splice site and AG| at the acceptor splice site. However, a distinct class of rare introns has been recognized on the basis of their unusual terminal dinucleotides: these introns contain |AT at the donor splice site and AC| at the acceptor splice site [20,21]. A closer examination of the sequences of these atypical introns

revealed several properties that distinguish them from the majority of the introns including conservation of unusual signals at the donor splice signal (|ATATCCTT) and immediately upstream of the acceptor splice signal (TCCTTAAC 10-15 bases from the splice junction) [20,21]. Introns of this class are excised by a distinct, so-called minor or U12 spliceosome, which contains several specific, low-abundance snRNPs. It has been subsequently shown that some |GT-AG| introns are also removed by the U12 spliceosome [22]. The U12 introns and the associated minor spliceosome are not universally conserved, like the major U2 spliceosome, but are also widespread in eukaryotes, being represented in vertebrates, insects, plants, and some protists [23-26].

Phylogenomic reconstructions for the small RNA and protein subunits of the U2 and U12 spliceosomes suggest that both spliceosomes were already present in the last common ancestor of the extant eukaryotes (LECA, Last Eukaryotic Common Ancestor) as a result of ancient duplication of the genes for the respective components [24]. Taking into account a potentially important role of U12 introns in regulation of gene expression [27-29], it might be tempting to speculate that the ancestral introns were of the U12 type (for example, see discussion by the reviewer #3 below) but have been subsequently converted to U2 introns. However, comparison of protosplice sites (exonic sequences surrounding introns) of ancient U2 and U12 introns in human and Arabidopsis revealed close similarity of ancestral introns to U2 but not to U12. Thus, the primordial spliceosomal introns were most likely of the U2-type [30].

The two principal mechanisms of splicing signal recognition are known as exon definition and intron definition [31-34]. Evidence of these two mechanisms has come from analyses of interactions between pre-mRNAs and various splicing factors [32,33,35]. The exon definition mechanism involves SR proteins binding to exonic splicing enhancers (ESE) and recruiting U1 to the downstream donor splicing signal and the splicing factor U2AF to the upstream acceptor splicing signal. The U2AF factor then recruits U2 to the branch site. Therefore, when the SR proteins bind the ESEs, they promote formation of a



“cross-exon” recognition complex by placing the basal splicing machinery at the splice sites flanking the same exon. The intron definition mechanism requires binding of U1 to the upstream donor splice site and binding of U2AF/U2 to the downstream acceptor splice signal and branch site, respectively, of the same intron. Therefore, intron definition selects pairs of splice sites located on both ends of the same intron, and SR proteins can also mediate this process [32,36]. The efficiency of splicing under the exon definition depends on the length of exons but is not affected by the length of introns; conversely, under the intron definition, the efficiency of splicing depends on the length of introns, but not that of exons [31-35,37].

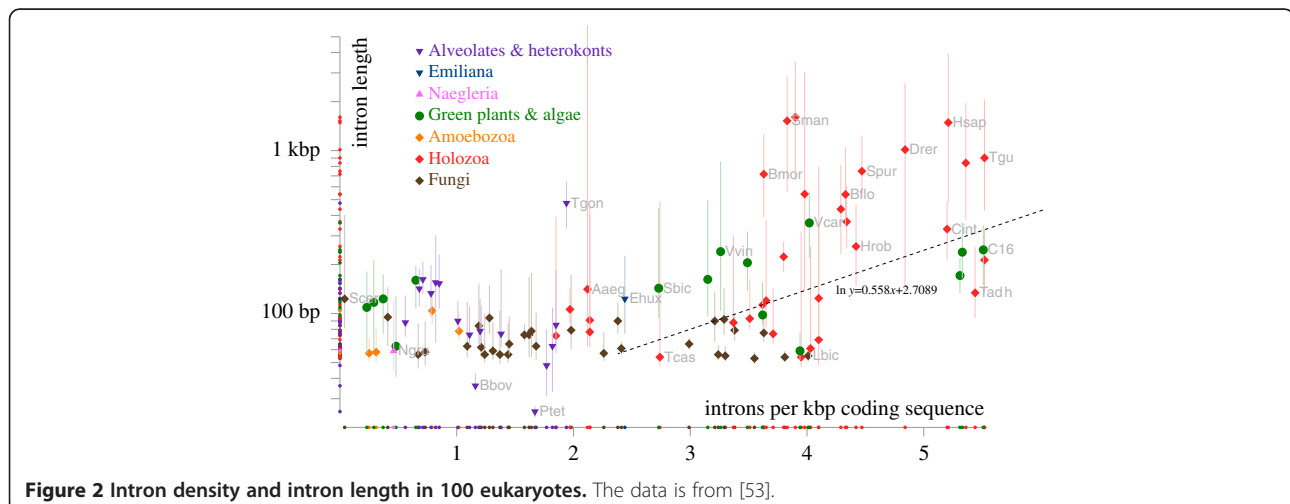
Introns-early, introns-late, introns-first: the competing scenarios of intron origin and evolution

Evolution of exon-intron structure of eukaryotic genes and evolutionary properties of introns had been long considered in the context of the “introns-early” vs. “introns-late” debate [38-42]. The original, “strong” introns-early hypothesis held that eukaryotic genes inherited (nearly) all introns from prokaryotic ancestors and that the differences in gene structure among homologous eukaryotic genes were due mostly to differential intron loss [39]. Under this scenario, the extant prokaryotes have lost all the primordial introns and the spliceosome in the process of ‘genome streamlining’. The later adaptations of the introns-early hypothesis assumed an intermediate position by allowing emergence of some new introns, in addition to the ancient ones [40]. The introns-late concept countered that introns were a eukaryotic novelty and new introns have been emerging continuously throughout eukaryotic evolution; in this scenario, bacteria and archaea never possessed intron or the spliceosome [41,43,44]. These hypotheses have been later merged into a synthetic concept that can be

denoted ‘many introns early in eukaryotic evolution’ [45,46] and that we discuss in greater detail below. In addition, there has been an attempt to revitalize the introns early idea in the ‘introns first’ scenario according to which exons of protein-coding genes emerged from the primordial introns, i.e. non-coding regions that are presumed to have been interspersed between functional RNA sequences in the genes that existed in the RNA world and antedated proteins [47,48].

Intron density, size and distribution in protein-coding genes across the eukaryote domain

Genes of eukaryotes from different groups dramatically differ in intron density and size distribution, from only a few introns in the entire genome (that is, near zero density per gene or per kilobase) in many unicellular organisms to approximately 6 introns per kilobase (kb) of coding sequence in mammals (Figure 2). With respect to intron content, eukaryotic genomes are often crudely classified into intron-poor ones (most unicellular forms) and intron-rich ones including animals, plants, some fungi, and a few unicellular organisms such as *Chlamydomonas* or some diatoms (Figure 2) [42,49-52]. Although this division is appealing in its simplicity and may be convenient for the purpose of various comparative analyses, examination of intron densities in 100 sequenced eukaryotic genomes does not present an obvious bimodal distribution (Figure 2). Actually, it appears that all intron densities between 0 and 6 introns per kilobase are observed in some eukaryote genomes. However, when intron density is plotted against intron length, partitioning of eukaryote genomes into two classes becomes apparent. While up to a density of approximately 3 introns per kilobase, all introns are short, with no significant correlation between the density and length of introns, for more intron-rich genomes, a strong positive correlation is observed (linear correlation



coefficient = 0.16, $P = 0.003$, Figure 2). Even among intron-rich organisms, vertebrates are outstanding in having a substantial fraction of extremely long introns (Figure 2). This strong correlation notwithstanding, there are exceptions to the general trend: intron-rich basidiomycete fungi (3-4 introns/kbp) have only short introns whereas some insects show broad intron length distributions with multiple long introns despite relatively low intron density (2-3 introns/kbp) (Figure 2). We return to the dependencies between intron density, intron length and structure of splice signals later, in the discussion of the selection pressures affecting the evolution of eukaryote gene architecture and the underlying population-genetic factors.

As pointed out above, despite the existence of numerous, diverse intron-poor genomes, eukaryotes do not lose the “last” intron or the spliceosome although degradation of the spliceosome including loss of many components does occur, e.g. in yeast. The only firmly established exception is the tiny genome of a nucleomorph (an extremely degraded intracellular symbiont of algae) that has lost both all the introns and the spliceosome [7]; preliminary genomic data indicate that all introns might have been lost also in a microsporidium, a highly degraded intracellular parasite distantly related to Fungi [54]. In general, it remains unclear whether there are any selective factors or functional constraint underpinning this surprising preservation of at least a few introns in eukaryote genomes [55]. However, in many cases, the few introns that are retained in highly reduced genomes are present in 5'-portions of genes encoding ribosomal proteins [16,56]. The introns in these genes are important for regulation of expression and ribosomal biogenesis, and their deletion leads to significant fitness reduction in yeast [57]. Thus, the extreme rarity of complete loss of introns in eukaryotes at least in part is likely to be due to deleterious effect of the loss of specific, functionally important introns.

Evolutionary conservation of intron positions and routes of gene architecture evolution of eukaryotes

The realization that (nearly) all eukaryotes possess ‘genes in pieces’ but that the intron densities and size widely vary, triggered intense, ongoing discussion of possible evolutionary scenarios behind these patterns. Several mechanisms of intron evolution have been suggested including intron loss, gain, and sliding [44,58-61]. Intron loss and gain are the major phenomena in the evolution of eukaryotic gene architecture. The relative contributions of these two processes have been a matter of considerable debate and controversy. Systematic comparative analyses of exon-intron structures of orthologous genes from animals, fungi and plants have shown

that approximately 25% to 30% of the intron positions are shared (that is, located in the exact same position in orthologous genes) by at least two of these three lineages of complex eukaryotes with intron-rich genomes [45,62]. The prevailing interpretation of these fundamental observations is that most, if not all, introns that occupy the same positions in orthologous genes are conserved, i.e. were already present in the equivalent position of the corresponding ancestral gene. However, the alternative view, i.e., that a substantial fraction or even most of the shared introns have been independently inserted in the same position in orthologous genes in different lines of descent, cannot be automatically dismissed (see discussion below).

The apparent conservation of many intron positions in distant eukaryotes notwithstanding, intron densities in eukaryotic genomes differ widely (see above), and the location of introns in orthologous genes does not always coincide even in closely related species [63-65]. Likely cases of intron insertion and the more common intron loss have been described (e.g., [59,63,66-82], and indications of high intron turnover rate in some eukaryotic lineages have been obtained [42]. Furthermore, parsimony and maximum likelihood (ML) reconstructions of evolution of exon-intron structure of highly conserved eukaryotic genes (see details below) suggest that both loss and gain of introns have been extensive during evolution of eukaryotic genes [45,83-88]. Together the results of these analyses indicate that the rates of intron gain and loss differ widely between eukaryotic lineages. Some eukaryotes, such as yeast *Saccharomyces cerevisiae*, seem to have lost nearly all of the ancestral introns, whereas others, e.g., nematodes, have experienced highly dynamic evolution, with both loss and acquisition of numerous introns [45,83-88]. However, intron gain is not easy to detect: comparative analysis of intron positions in orthologous genes from vertebrates revealed only a few losses but no apparent gain of introns in mammalian genes [89,90], and similar results have been obtained in an analysis of evolution of exon-intron structure of paralogous genes in several eukaryotic lineages [91]. These findings imply that intron loss dominates at short evolutionary distances, whereas bursts of intron insertion might accompany major evolutionary transitions. However, intron gain could be an ongoing process in nematodes: Coghlan and Wolfe [66] identified 81 introns in *Caenorhabditis elegans* and 41 introns in *C. briggsae* that appear to have been recently inserted. However, the validity of these recent intron gains has been questioned as it has been shown that after including additional genomes in the analysis many of the reported intron gains could be parsimoniously explained by intron loss [92]. A high rate of recent intron gain has been reported for paralogous gene pairs in *Arabidopsis thaliana* that were

duplicated simultaneously 20-60 million years via tetraploidization [93]. A low rate of recent intron gains has been identified in plastid-derived genes in plants [94]. Similarly, spliceosomal introns have been detected in some genes horizontally transferred from bacteria to bdelloid rotifers [95]. Probably, the most striking known case of apparent recent intron gains has been found in populations of *Daphnia pulex* endemic to Oregon where two polymorphic introns have been identified [70]. These new introns do not have an obvious source and are not represented in any studied *D. pulex* populations outside Oregon, other species of *Daphnia* or any other organism for which sequence data are available. Furthermore, the new introns are both found in the same gene that encodes a Rab GTPase (*rab4*), and are inserted within one base pair from each other. These findings put into doubt the rarity of intron gain considering that two intron gain events have been discovered in an initial survey of only 6 nuclear loci in 36 *Daphnia* individuals [70]. This result was further supported by the analysis of 24 discordant intron/exon boundaries between the whole-genome sequences of two *Daphnia pulex* isolates. Sequencing of intron presence/absence loci across a collection of *D. pulex* isolates and outgroup *Daphnia* species has shown that most polymorphisms result from recent gains, with parallel gains often occurring at the same location in independent allelic lineages [96].

The great majority of studies aimed at reconstruction of evolution of gene architecture in eukaryotes have focused on introns in conserved portions of protein-coding regions. For example, the conclusion that there was no appreciable intron gain in mammals [89] is based on this type of data. However, evolution of poorly conserved segments of protein-coding sequences, untranslated regions of protein-coding genes, alternatively spliced regions, and genes originated from transposable elements appears to be much faster and more dynamic, with numerous intron gains in mammals [97-101]. A case of such intron acquisition has been reported for the RNF113B retrogene that encodes a RING finger protein (a predicted E3 subunit of ubiquitin ligase of unknown specificity) and is present in the genomes of all primates (Figure 3) [101]. This primate-specific gene underwent rapid evolution that included an intron gain. The presence of the intron is supported by several human mRNA sequences and comparisons with multiple primate genomes (marmoset, macaque, orangutan, and chimpanzee). Sequence alignment analysis shows that the intron of RNF113B is not a *de novo* insertion but rather a derivative of an exonic sequence (a tandem mutation AG > GT generated the donor site). The new intron contains 59 nucleotides from former coding sequence and 46 nucleotides from the 3' UTR. This finding was further supported by sequencing of the human RNF113B cDNAs which

revealed two alternative RNF113B isoforms (Figure 3) [101]. In general, due to the lack of evolutionary conservation in such genes and gene regions, reconstruction of intron gain and loss events in their evolution is difficult and sometimes inaccurate (especially without experimental verification). Accordingly, evolutionary studies tend to concentrate on highly conserved genes. Thus, the conclusions on intron stasis in some groups of eukaryotes, such as mammals, in part appear to stem from sampling biases whereas the overall intron turnover might be much more extensive than is currently appreciated.

The same problem pertains to non-coding RNA genes. For example, mammalian genomes contain numerous (> 10,000) genes for long non-coding RNAs (lncRNAs) that encompass numerous introns [102]. In a recent detailed study, over 8,000 lncRNA genes have been identified, with a mean intron density of ~1.9 per kilobase, and extensive alternative splicing of these non-coding RNAs has been detected, with ~2.3 RNA isoforms per gene [103]. One of the best studied lncRNAs is Xist which is involved in X-chromosome inactivation in females of eutherian mammals [104]. The Xist RNA appears to have evolved as a result of pseudogenization of the *Lnx3* protein-coding gene in early eutherians followed by integration of mobile elements [105]. Analysis of Xist in several mammalian species revealed an overall conservation of the Xist gene structure (Figure 4). Four of the 10 Xist exons found in eutherians show significant sequence similarity to exons of the *Lnx3* gene (Figure 4) whereas the remaining 6 Xist exons are similar to different transposable elements. Thus, some Xist introns were inherited from the *Lnx3* gene but some appear to have been gained in the course of evolution of the Xist gene [105]. Comparative analysis of >3,000 mouse lncRNA genes suggested that conservation of the exon/intron structure might be a general lncRNA property [106]. It was found that 65% and 40% of mouse lncRNA |GT-AG| splice sites are conserved in human and rat, respectively. These numbers are significantly greater than the number of conserved intronic GT and AG dinucleotides that are not involved in splicing indicating evolutionary conservation of splice signals in lncRNAs [106]. Detailed reconstruction of the origin and evolution of introns in lncRNAs awaits further comparative genomic studies.

The distributions of intron positions over the length of coding regions differ substantially between eukaryotic taxa. In intron-poor genes of single-cell eukaryotes, introns are strongly over-represented in the 5'-portions whereas in intron-rich multicellular organisms, the distribution is closer to uniformity [64,65]. A mechanistic explanation for these patterns suggests that introns are preferentially lost from the 3'-portion of a gene, conceivably due to the over-representation of the respective sequences among the cDNAs that are produced by

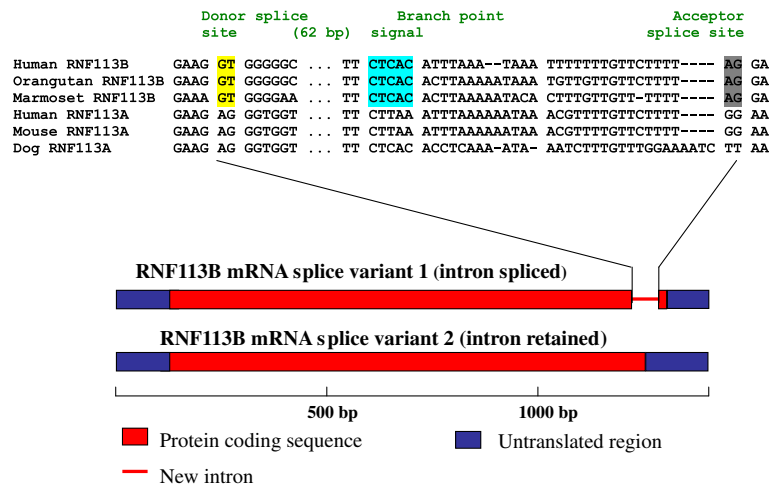


Figure 3 An example of a recent intron acquisition in a retrotransposon-derived gene: structure of two splice variants of RNF113B. The new intron of RNF113B is not a *de novo* insertion but rather a derivative of exonic sequences (this intron contains 59 nucleotides from the former coding sequence and 46 nucleotides from the 3' UTR). A partial alignment of three RNF113B sequences and three RNF113A sequences is shown above the spliced RNF113B isoform. The donor splice site is marked in yellow, the predicted branch point signal is marked in blue, and the acceptor splice site is marked in gray. The data is from [101].

reverse transcription and are thought to mediate intron loss via homologous recombination [65,107-109]. However, a complementary, selectionist interpretation of the observed distributions of introns, to the effect that introns located in the 5'-portion of a gene are more often involved in one or more intron-mediated functions (see below), has been proposed as well [65]. Analysis of distributions of intron positions over the length of the coding region suggested that both loss and insertion of introns occurred preferentially in the 3'-regions of genes, which suggested reverse-transcription-mediated mechanisms for both processes [110]. This hypothesis appears to be compatible with the positive association that has been shown to exist between the rates of intron gain and loss in individual genes [111]. However, a more recent probabilistic analysis of intron gain and loss indicates that the mechanisms of loss and gain are most likely to be different, with reverse transcription involved only in intron loss [112].

Intron sliding (also called slippage or migration; hereinafter IS) can be defined as relocation of intron/exon boundaries over short distances (1-60 bases) in the course of evolution. Intron sliding has been postulated by advocates of the introns-early hypothesis to explain the surprising finding that the positions of apparently orthologous introns sometimes varied among lineages [60]. However, the introns-late camp maintained that IS, if it occurs at all, has contributed little to the diversity of intron positions [44,59]. The reality of IS had been debated for a long time. A Monte Carlo statistical analysis of broadly sampled data on intron positions has strongly suggested that one-base-pair IS, although a relatively rare event occurring in <5% of all introns, was a bona fide evolutionary phenomenon; in contrast, no evidence supporting intron sliding over longer distances was obtained [113]. A recent study has suggested that IS might be an important source of new introns in some lineages and proposed a simple, plausible mechanism for

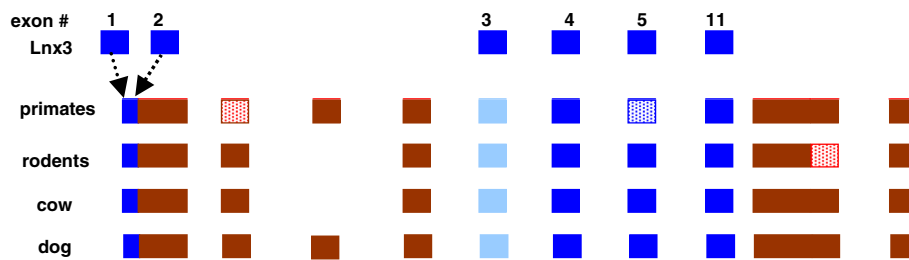


Figure 4 The Xist gene evolved from a protein-coding gene and a set of transposable elements. Blue boxes indicate exons originating from Lnx3; red boxes indicate exons originating from transposable elements; dashed boxes indicate remnants of protein-coding exons. The data is from [105].

the emergence and fixation of IS during evolution [114]. Given the near ubiquity of alternative splicing (AS) in many groups of animals and possibly plants [48], Tarrío et al. proposed that AS could be an intermediate stage in the evolution of IS. Under this scenario, emergence of a new splicing signal adjacent to a pre-existing one results in AS but, if and when the original splicing signal subsequently deteriorates, the net result is IS [114]. The proposed route of IS evolution via AS is likely to be common in poorly conserved regions of protein-coding genes with frequent AS events (e.g. 5'- and 3'-regions of many genes) but rare in conserved portions of protein-coding genes. Comparative analysis of closely located introns among 12 *Drosophila* genomes has suggested that IS is a relatively frequent cause of novel intron positions in *Drosophila* [115]. All things considered, there is currently no doubt that IS is real and can yield new intron positions but the actual impact of IS in the evolution of eukaryotic genes will be accurately determined only when multiple sets of closely related genomes become available and rigorous methods for statistical analysis are developed.

Evolution of splicing signals, protosplice sites, and intron phase distribution

As pointed out above, the densities of spliceosomal introns vary dramatically among eukaryotes (Figure 2), and so does the strength of splicing signals [18,45,51,116]. There is a striking correspondence between low intron density and high information content of donor splice signals across eukaryotic genomes [51]. Intron-poor genes (genomes) with strong donor sites are found in several groups of eukaryotes (e.g. fungi) that also include intron-rich genomes with weaker donor sites. Evolutionary reconstruction suggests that ancestral donor signals had low information content but that many lineages have independently underwent concomitant major intron loss and donor signal strengthening [51]. This evolutionary trend receives a straightforward explanation within the framework of the population-genetic concept of evolution of gene architecture (see below).

However, the acceptor splice signal shows a different trend: it is weak in most fungi, intermediate in plants and some unicellular eukaryotes, and strongest in metazoans where it gradually strengthens from nematodes to mammals [116]. This observation can be interpreted in the light of the results of a large-scale analysis of splicing signals in 61 eukaryotic species which revealed a significant negative correlation between the strength of the branch point signal and the strength of the acceptor splice site (Figure 5; $R = -0.52$, $P = 0.000025$) [117]. Although the correlation between the strength of the donor splice signal and the combined strength of the branch point signal and the acceptor splice signal was not

significant ($R = 0.19$, $P = 0.15$), the positive sign of this correlation still could reflect congruent evolution of splicing signals. In general, a complex interplay exists between intron density, intron size, the strength of splice signals and the strength of splicing enhancers/silencers. For example, splice signals in long and short introns in *Drosophila* show only minor differences [118]. Several weak but statistically significant correlations have been observed between vertebrate intron length, splice site strength, and potential exonic splicing signals that attest to a compensatory relationship between splice sites and exonic splicing signals, depending on vertebrate intron length [119].

It has been proposed that the functionally important (A/C)AG||G exon sequences flanking introns are relics of recognition signals for the insertion of introns that have been dubbed protosplice sites [14,15]. Protosplice sites became an important staple of the introns-late hypothesis of intron evolution because, if intron insertion was limited to strictly defined protosplice sites, parallel gain of introns would be likely and could account for the large number of shared introns among orthologs from distant eukaryotic lineages [41,63,83]. Support for the protosplice site hypothesis has been harnessed from experiments demonstrating that elimination of the regular splice sites in actin genes resulted in activation of cryptic splice sites, most of which coincided with exon junctions in orthologous genes from other species [120]. Nevertheless, it remained unclear whether the consensus nucleotides flanking the splice junctions were remnants of the original protosplice sites or evolved convergently after intron insertion. The existence of protosplice sites was directly addressed by examining the context of introns inserted within codons which encode amino acids conserved in all eukaryotes and, accordingly, are not subject to selection for splicing efficiency. According to the parsimony principle, these codons (e.g., GGN for conserved glycines or CCN for conserved prolines) can be inferred to have been present already in the common ancestor of all extant eukaryotes, so the ancient protosplice sites (if such existed) should have survived and could be examined directly. This analysis has shown that introns, indeed, predominantly insert into and/or are preferentially fixed in specific (protosplice) sites with the consensus sequence (A/C)AG||Gt [121].

Recently, correlation between positions of cryptic splicing signals (sequences that are similar to splicing signals but normally do not function in splicing) and introns has been found: cryptic splicing signals within exons of one species frequently match the exact position of introns in orthologous genes from another species. This observation suggests that in the course of evolution many introns were inserted into cryptic splicing signals that had been in place prior to intron insertion [122]. However, this conclusion is contradicted by another

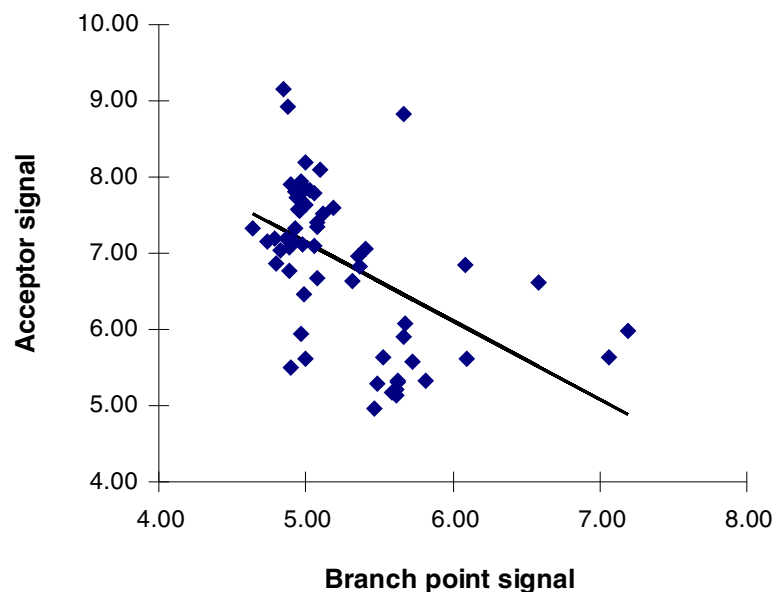


Figure 5 Correlation between the strength of the branch point signal and the strength of the acceptor splice site. The linear correlation coefficient is $R = -0.52$ ($P = 0.000025$) after exclusion of the obvious outlier *Aureococcus anophagefferens* [117]. The information content of splicing signals in 61 eukaryotic species is from [117]. Species names: *B. taurus*, *C. familiaris*, *E. caballus*, *H. sapiens*, *M. domestica*, *M. musculus*, *O. anatinus*, *R. norvegicus*, *S. scrofa*, *B. florida*, *C. intestinalis*, *C. savignyi*, *D. rerio*, *G. gallus*, *O. latipes*, *P. marinus*, *T. guttata*, *X. tropicalis*, *A. gambiae*, *A. mellifera*, *C. elegans*, *D. pulex*, *D. melanogaster*, *H. magnipapillata*, *L. gigantea*, *M. brevicollis*, *N. vectensis*, *S. purpuratus*, *T. castaneum*, *B. dendrobatidis*, *C. heterostrophus*, *C. neoformans*, *M. grisea*, *N. haematococca*, *P. chrysosporium*, *P. blakesleeanus*, *P. infestans*, *P. placenta*, *S. cerevisiae*, *S. commune*, *T. virens*, *A. anophagefferens*, *D. discoideum*, *D. purpureum*, *N. gruberi*, *O. lucimarinus*, *P. tricornutum*, *T. pseudonana*, *T. adhaerens*, *A. thaliana*, *Chlorella NC64A*, *C. reinhardtii*, *M. pusilla*, *Micromonas RCC299*, *O. sativa*, *P. patens*, *P. trichocarpa*, *S. moellendorffii*, *S. bicolor*, *V. vinifera*, *V. carteri*.

observation, that recently gained introns in animal genes of the alpha-amylase were not associated with specific sequence motifs (protosplice sites) [123]. In the same gene family, old introns were embedded within strong protosplice motifs which were found to be much weaker in homologous genes lacking the intron in the given position [123]. These findings are consistent with the hypothesis that sites of *de novo* intron insertion are effectively random and that selection drives the emergence of protosplice-like sequences following intron insertion. The presence of much stronger protosplice sites around old introns compared to young introns [123] seems to suggest that evolution of protosplice sites subsequent to intron insertion is a slow process [123,124].

The hypothesis that selection acts on the exonic parts of splice signals was supported by comparison of the nucleotide sequences around the splice junctions that flank old (shared by two or more major lineages of eukaryotes) compared with new (lineage-specific) introns in eukaryotic genes. The distributions of information content between the intronic and exonic parts of the splices signals have been found to be substantially different in old and new introns [125]. Old introns have lower information content in the exonic regions adjacent to the splice sites than new introns but, conversely, have higher information content in the intron itself. These findings

imply that introns insert into protosplice sites but during the evolution of an intron after insertion, the splice signal shifts from the flanking exonic regions to the ends of the intron itself. Accumulation of information inside the intron during evolution is best compatible with the view that new introns, largely, emerge *de novo* and not via propagation and migration of pre-existing introns [125].

The contradictory findings on the protosplice site involvement versus the evolution of these motifs after intron gain (in which case 'protosplice site' becomes a misnomer) might reflect objectively existing differences in the evolution of the gene architectures among gene families, in particular between highly conserved and more dynamic families. The definitive assessment of the validity of the protosplice site hypothesis requires further, comprehensive comparative genomic studies.

Introns occur in three phases (0, 1, and 2) which are defined as the position of the intron within or between codons: introns of phase 0, 1, and 2 are located, respectively, between two codons, after the first position in a codon, and after the second position. In (nearly) all analyzed genomes, there is a significant excess of phase 0 introns over those in the other two phases [125-130]. The only known remarkable exception is the rapidly evolving tunicate *Oikopleura* that shows a uniform distribution of introns among the three phases [131].

An excess of protosplice sites in phase 0 was noticeable in some species (Figure 6) [132], however the protosplice site model cannot fully explain the observed over-representation of phase 0 introns under the assumption that introns randomly insert into protosplice sites (Figure 6) [125,127,128]. Furthermore, it has been shown that phase 0 introns were, on average, located in more highly conserved portions of genes than phase 1 and 2 introns [45]. This observation suggests that phase 1 and phase 2 introns experience a greater deleterious-mutation-driven loss and could reconcile the observed phase distribution with the protosplice site hypothesis [125,127,128,130].

Conservation versus parallel gains of introns

As discussed above, comparative analyses revealed numerous introns that occupy the same position in orthologous genes from distant species [45,62]. In particular, orthologous genes from humans and the green plant *A. thaliana* share ~25% intron positions [45]. The straightforward interpretation of these observations is that the shared introns were inherited from the common ancestor of the respective species whereas lineage-specific introns were inserted into genes at later stages of evolution [45,62]. Under this premise, parsimonious reconstructions indicate that even early eukaryotes already had a relatively high intron density, perhaps, comparable (at least within an order of magnitude) to that in modern plant and animal genes. However, the inference that shared intron positions reflect evolutionary conservation is challenged by the potential non-randomness of intron insertion: introns are inserted or fixed mostly in distinct protosplice sites as discussed in the preceding section. In principle, if there were few protosplice sites in each gene, the presence of introns in the same positions of

orthologous genes in distantly related species could be completely or at least to a large extent explained by parallel gains. At least two cases of apparent parallel gain of introns in orthologous genes from plants and animals have been reported [133,134]. Moreover, probabilistic modeling of intron evolution discussed above suggested that many, if not most, introns shared by phylogenetically distant species were likely to originate by parallel gain of introns in protosplice sites [83]. The implication is that intron distribution in extant organisms is largely determined by relatively recent insertions and cannot be used to infer exon-intron structure of ancestral genes. However, the dataset of 10 gene superfamilies by Qiu and co-workers [83] contained numerous ancient duplications combined with frequent lineage-specific losses of genes, because of which analysis of intron conservation and intron gains is likely to be confounded by problems of phylogenetic reconstructions.

The extent of independent insertion of introns in the same sites (parallel gain) in orthologous genes from phylogenetically distant eukaryotes was assessed within the framework of the protosplice site model [132]. It was shown that protosplice sites are no more conserved during evolution of eukaryotic gene sequences than random sites. Simulation of intron insertion into protosplice sites with the observed protosplice site frequencies and intron densities has shown that parallel gain could account for only 5 to 10% of shared intron positions in distantly related species [132]. The results of this simulation suggest that the presence of numerous introns in the same positions in orthologous genes from diverse eukaryotes, such as animals, fungi, and plants, reflects mostly *bona fide* evolutionary conservation [132].

Analysis of intron gain and loss rates over branches of the phylogenetic tree for 19 eukaryotic species allowed

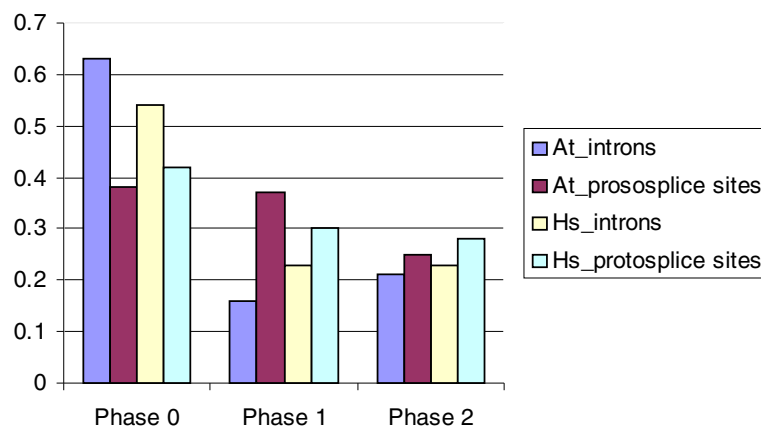


Figure 6 Fractions of protosplice sites and actual introns in the three phases. Species abbreviations: (At) green plant *Arabidopsis thaliana*, (Hs) human *Homo sapiens*. An excess of protosplice sites in phase 0 is noticeable, however the 'protosplice site' hypothesis, which posits that introns are randomly inserted into protosplice sites, is unable to fully explain the observed over-representation of phase 0 introns. The data is from [125,132].

for the estimation of the probability of parallel gain for each intron position that is shared by more than one species [111]. The resulting estimates indicated that parallel gain, on average, accounts for only ~8% of the shared intron positions, in agreement with the simulation results discussed above [111]. However, the distribution of parallel gains over the phylogenetic tree of eukaryotes is highly non-uniform. There are almost no parallel gains in closely related lineages, whereas for distant lineages, such as animals and plants, parallel gains could contribute up to 20% of the shared intron positions. Taken together, the results of these analyses indicate that, although parallel gain of introns is non-negligible, the substantial majority of introns that occupy the same positions in orthologous genes from distantly related eukaryotes are ancestral including many inherited from LECA [111].

Reconstruction of evolution of exon-intron structure of eukaryote genes

The patterns of conservation and variation of intron positions in orthologous and paralogous genes can be employed to reconstruct evolutionary scenarios for the exon-intron structure of eukaryotic genes using evolutionary parsimony or maximum likelihood approaches. Once multiple eukaryotic genomes have been sequenced, such genome-wide evolutionary reconstruction has become realistic. The comparative data on intron positions can be naturally represented as a matrix of intron presence/absence (usually encoded as 1/0), and to these matrices, various reconstruction methods can be applied. The first of such studies employed orthologous gene sets from 8 eukaryotic species and took the most straightforward approach to evolutionary reconstruction by applying the parsimony principle in the specific form of Dollo parsimony [45]. Given a species tree topology and a pattern of intron presence/absence, the Dollo algorithm constructs the most parsimonious (simplest) scenario for the evolution of gene structure, i.e. the distribution of inferred intron gain and loss events over the tree branches. The main underlying assumption is that a character (intron in a given position) once lost cannot be regained whereas as many parallel intron losses in different branches of the tree are allowed as needed to account for the observed pattern of states. By analyzing more than 7,000 intron positions in highly conserved genes of eukaryotes, the Dollo parsimony approach produced an evolutionary scenario under which the common ancestor of modern eukaryotes possessed intron-rich genes, with intron density only a few fold lower than that in most intron-rich modern forms (vertebrates and plants). Massive intron losses were inferred for several groups, especially yeasts, nematodes

and insects, whereas in vertebrates and plants intron gain was inferred to be the main evolutionary trend [45].

The parsimony approach is obviously oversimplified given that all lineage-specific introns are automatically treated as newly gained, with the possibility that some of these introns could be ancestral, having been lost in multiple lines of descent. Furthermore, parsimony does not provide confidence estimates for the estimates of ancestral states. These limitations of parsimony potentially could grossly distort the results of evolutionary reconstruction, especially if the number of analyzed species is small. Probabilistic approaches such as maximum likelihood models can address these problems, at least in principle. However, the first two statistical studies into intron evolution produced opposite results: Qiu *et al.* concluded that ancient introns (if they ever existed) have not survived in extant genes [83] whereas Roy and Gilbert came to the conclusion that the great majority of introns, even lineage-specific ones, were ancient [84]. The first conclusion implies that intron gain was dominant over intron loss in the evolution of eukaryotic genes, whereas the second one suggests that intron loss is the principal evolutionary process. This major discrepancy between the results of the two studies has indicated that optimal parameters for maximum likelihood modeling of intron evolution remained to be determined [135].

The next generation of increasingly sophisticated ML reconstructions of intron gain and loss during eukaryotic evolution suggested that the protein-coding genes of ancient eukaryotic ancestors, including the Last Eukaryotic Common Ancestor (LECA), already possessed intron density comparable to that found in modern, moderately intron-rich genomes [85-88,136]. Accordingly, the history of eukaryotic genes, with respect to the dynamics of introns, appears to be to a large extent dominated by losses, perhaps punctuated by a few episodes of major gain [87,88,91,137]. This conclusion is based on analyses of considerably larger data sets than those used in earlier studies; for example, Carmel and co-workers [87] analyzed 391 sets of orthologous genes from 19 eukaryotic species. This extended data set not only allowed for a more definitive reconstruction of the gene structure evolution, but also permitted zooming in on specific portions of the eukaryotic tree [87]. A comprehensive probabilistic model of intron evolution was developed that incorporated heterogeneity of intron gain and intron loss rates with respect to both lineages and genes as well as variability among sites within a gene [87]. It was demonstrated that ancestral eukaryotic forms were intron-rich and that evolution of eukaryotic genes involved numerous gains and losses of introns, with losses being somewhat more common. Three distinct modalities of intron gain and loss during eukaryotic evolution were identified. The 'balanced' mode appears

to operate in all eukaryotic lineages, and is characterized by proportional intron gain and loss rates [87]. In addition to this, apparently universal process, many lineages exhibit elevated loss rate, whereas a few others exhibit elevated gain rate. Moreover, the rates of intron gain and loss are highly non-uniform over the time course of the evolution of eukaryotes: both rates seem to have been decreasing with time over the last several hundred million years. 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 [87].

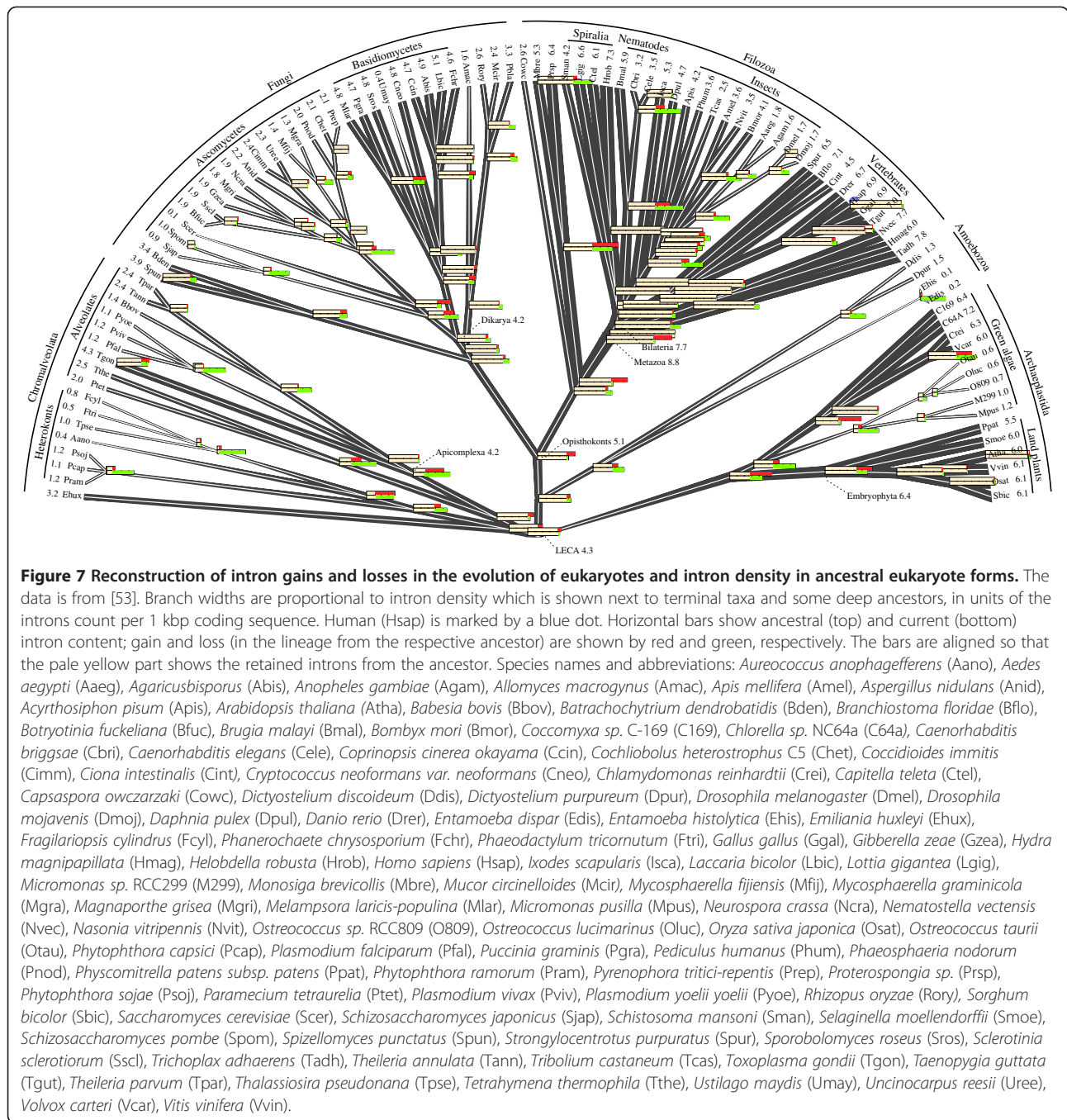
Figure 7 illustrates the latest reconstruction of intron gain and loss for 99 eukaryotic species that was performed using a Markov Chain Monte Carlo (MCMC) approach [53]. In this, so far the most extensive study, the Malin software package [138] was used to infer ancestral states from a matrix of shared introns which comprised 8403 intron presence-absence profiles from 245 sets of orthologous genes. The MCMC method infers ancestral intron density by using a probabilistic intron gain-loss model, taking into account rate heterogeneity across lineages and across sites within genes. This reconstruction provides a thorough view of the evolution of gene structure in three eukaryotic supergroups and reveals several general trends (Figure 7) [53]. Most lineages show net intron loss that can be substantial as in alveolates, some lineages of fungi, green algae or insects, or offset by concomitant intron gains as in land plants, most animal lineages, and some fungi. Massive intron gains were inferred only for several deep branches, most conspicuously the stem of the Metazoa, and to a lesser extent, the stems of Mamiellales (a branch of green algae), Viridiplantae, Opisthokonta, and Metazoa together with Choanoflagellata (Figure 7). These findings vindicate, on a much larger data set and with greater confidence, the previous conclusions that compared to the common and substantial intron loss, extensive intron gain was rare during the evolution of eukaryotes. Episodes of substantial intron gain seem to coincide with the emergence of major new groups of organisms with novel biological characteristics such as the Metazoa (Figure 7) [53]. Several previous studies, performed on much smaller data sets and with less robust reconstruction methods, have suggested that at least some eukaryotic ancestral forms could have possessed intron-rich genes [84,85,136], and observations on gene structures in primitive animals such as the sea anemone *Nematostella* [139] and the flatworm *Platynereis* [140] were compatible with these inferences. A particularly striking conclusion has been reached in the reconstruction of the evolution of gene architecture in Chromalveolata: although all sequenced genomes in this supergroup of eukaryotes are intron-poor, intron-rich last common ancestors have been inferred for Chromalveolata and particularly Alveolata [141]. Clearly, the reconstruction led to this conclusion

only because, although very few intron positions are conserved among the intron-poor orthologous genes of different chromalveolates, many of these genes share a large fraction of intron positions with intron-rich orthologs from plants and/or animals.

The latest MCMC reconstruction reinforced these conclusions by inferring high intron densities for the ancestors of each major group of eukaryotes within each of the three supergroups (Figure 7) [53]. The implication is that, whenever an extant eukaryotic genome shows a low intron density, this intron-poor state is a result of extensive, lineage-specific intron loss. Remarkably, so many intron positions are shared between eukaryotes that, with the large and apparently representative set of compared genomes, Dollo parsimony reconstruction infers similarly intron-rich ancestral genomes as the MCMC and maximum likelihood methods [53]. The results of this reconstruction indicate in particular that the entire line of descent from LECA to mammals was a continuous intron-rich state (Figure 7) that would provide for uninterrupted evolution of the growing repertoire of functional alternative spliced forms (see below). The unprecedented intron gain at the onset of animal evolution (Figure 7) could further contribute to the expansion of alternative forms. This spurt of intron gain conceivably resulted from a combination of a population bottleneck that led to weak purifying selection with increased transposon activity (see below).

Evolution of exon-intron structure in paralogous gene families

The reconstructions of the evolution of gene architecture in eukaryotes were performed on sets of orthologous genes with a single representative (or a single most conserved representative) in each of the analyzed genomes. Obviously, this type of reconstruction reflects only one facet of evolution of gene structure given that all eukaryotic genomes encompass numerous families of paralogous genes with broad distributions of the number of members. Reconstruction of parsimonious scenarios of gene structure evolution in paralogous gene families in animals, plants and malaria parasites revealed numerous apparent gains and losses of introns [91,142]. In all analyzed lineages, the number of acquired new introns was substantially greater than the number of lost ancestral introns. This trend held even for lineages in which vertical evolution of genes involved many more intron losses than gains, suggesting that gene duplication boosts intron insertion. However, dating gene duplications and the associated intron gains and losses based on the molecular clock assumption showed that very few, if any, introns were gained during the last approximately 100 million years of animal and plant evolution, in agreement with previous conclusions reached through



analysis of orthologous gene sets. These results are generally compatible with the emerging notion of intensive insertion and loss of introns during transitional epochs in contrast to the relative quiet (stasis) of the intervening evolutionary spans [91,137,143]. The prevalence of intron gain over intron loss in evolving families of paralogs remains a somewhat controversial issue. It has been suggested that the Dollo parsimony approach used by Babenko and co-workers [91] could significantly underestimate the rate of intron losses [144]. However, even

should that be the case, the independently estimated number of intron gains in the same data set that was used in the original work of Babenko and coworkers [91] still exceeded the number of intron losses [144]. Furthermore, numerous anecdotal observations (e.g., [93,145-147]) have suggested that at least some paralogous gene families have gained more introns than they have lost.

In contrast, comparison of the exon-intron structures of ancient eukaryotic paralogs reveals the absence of

conserved intron positions in these genes (Figure 8) [148]. This is in contrast to the conservation of intron positions in orthologous genes from even the most evolutionarily distant eukaryotes and in more recent paralogs (Figure 8) [91]. The lack of conserved intron positions in ancient eukaryotic paralogs probably reflects the origin of these genes during the earliest phase of eukaryotic evolution that was characterized by concomitant invasion of genes by group II self-splicing elements (which were to become spliceosomal introns subsequently; see below) (Figure 9) and extensive duplication of genes [148,149]. Similar estimates were obtained for parallel intron gains in 'pseudo-paralogous' genes encoding cytosolic and mitochondrial ribosomal proteins that by definition have acquired their intron independently: approximately 2.3% of the intron positions were found in homologous positions [150]. The lack of conserved introns in ancient eukaryotic paralogs [148,150] is consistent with the results of an earlier analysis of intron distribution in 20 most ancient (duplicated before the divergence of bacteria and archaea) paralogous families which appear to have accumulated introns independently [151]. Along with other lines of evidence, these observations do not seem to be compatible with the introns early hypothesis.

Evolution of exon-intron structure in connection with other features of eukaryote genes

The combined advances of comparative genomics and systems biology provide means to characterize genes by many features, for example expression level and connectivity in protein-protein interaction or regulatory networks. Various significant correlations have been demonstrated to exist between these variables; in particular, one of the most prominent, recurrent links is that the sequence of

highly expressed genes tends, on average, to be more conserved [152-154]. Connections between various features of introns and other characteristics of genes also have emerged. Here, we discuss the links between two key features of introns, the rates of gain and loss and intron length, and other aspects of gene evolution, expression and function.

Probabilistic evolutionary reconstruction of gene structure yields gene-specific rates of intron gain and loss and thus provides for analysis of possible relationships between these rates and other characteristics of the respective genes [87]. It has been shown that intron gain rate was negatively and significantly correlated with the sequence evolutionary rate; conversely, intron loss rate was positively and significantly correlated with the rate of sequence evolution. Thus, perhaps somewhat counter-intuitively, highly conserved genes appear to accumulate introns in the course of evolution, even if slowly. Also significant, although of a lesser magnitude, was the positive correlation between gene expression level and intron gain rate and the converse negative of expression with intron loss rate. This finding suggests that introns might contribute to optimal gene expression [155] although this effect is confounded by the stronger connection between expression and evolution rate.

Although expression may be enhanced by the mere presence of multiple introns in a gene, highly expressed gene in human and *Drosophila* have, on average, shorter introns than genes expressed at a lower level [156]. This finding has been subsequently validated and expanded by several independent research groups on other model eukaryotes, for exon lengths as well, and for a variety of methods used to measure expression level [157-165]. Two competing (although not necessarily mutually exclusive) hypotheses have been proposed to explain the apparent

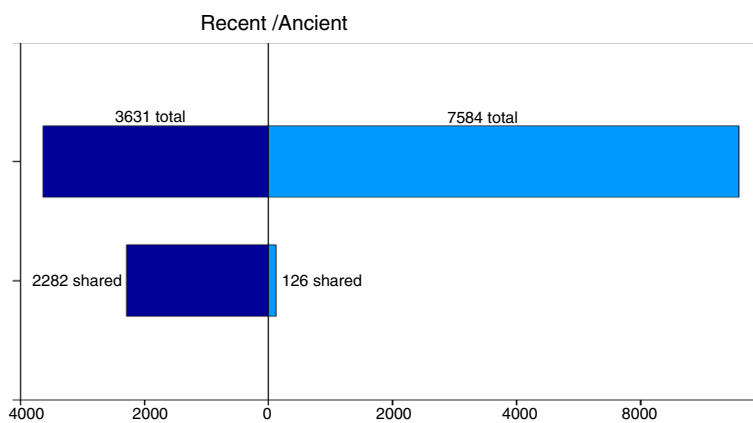
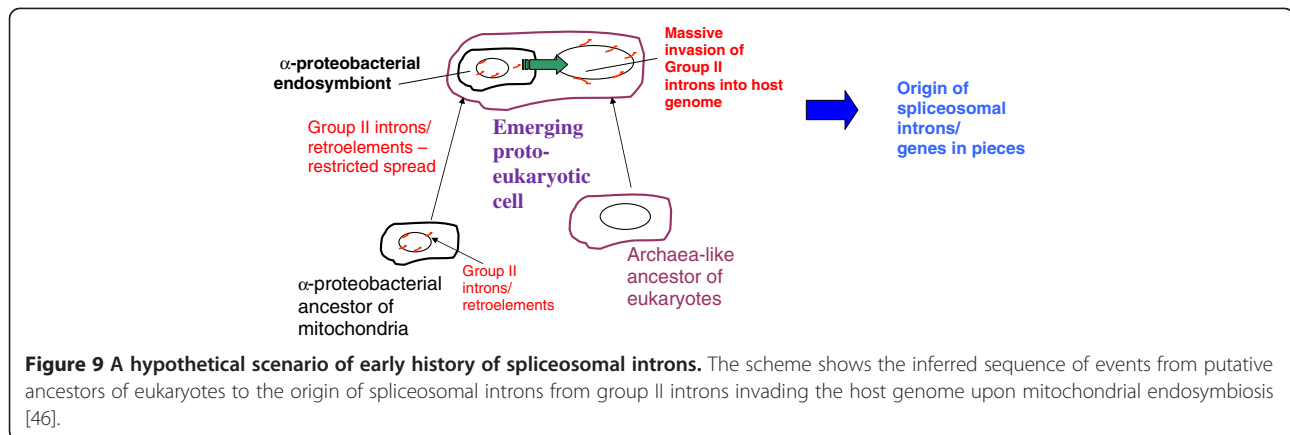


Figure 8 Conservation of intron positions in ancient and recent eukaryotic paralogs. Conservation of introns was assessed by analysis of multiple alignments of paralogous sequences from 6 species (*H. sapiens*, *C. elegans*, *D. melanogaster*, *S. pombe*, *S. cerevisiae*, *A. thaliana*). An intron position was considered to be conserved if it was shared by any pair of paralogs [148].



link between gene compactness and expression. The selection hypothesis holds that evolution of highly expressed genes is driven by selection for minimization of the time of transcription and/or energy expenditure resulting in shrinking of these genes, especially introns [156]. The alternative view, known as the genomic design hypothesis, holds that genes that are expressed under tight tissue- and developmental stage-dependent control require complex regulation and therefore need long introns to accommodate additional regulatory elements. Under the genomic design view, due to the positive association between the breadth and rate of gene expression, genes that are constitutively expressed at a high level and do not require complex regulation possess shorter introns [160].

Surprisingly, however, the opposite trend has been reported to exist in plants, with highly expressed genes containing longer introns [166]. This discrepancy was resolved by examining the relationship between gene length and expression level at a finer resolution: the relationship between intron length (as well as other measures of gene compactness such as the length of exons or entire genes) and expression level is universal across all eukaryotes (for which sufficient amount of data on expression was available) but is non-monotonic [167]. Genes that are highly expressed indeed tend to have shorter introns but genes expressed at low to medium levels show a positive correlation between intron length and expression; hence a roughly bell-shaped dependency between expression level and intron length (Figure 10) [167]. The phenomena that underlie this non-monotonic dependency are not quite clear but might involve competition between two opposing trends. Selective pressure to maximize expression rate and minimize energy expenditure could be dominant in highly expressed genes as originally suggested [156]. In contrast, requirement for more complex regulation in moderately expressed genes that gain additional functions with increased expression might result in the positive correlation between intron length and expression [167].

A population-genetic perspective on evolution of introns and eukaryotic gene architecture

The question famously posed by Walter Gilbert in the seminal note on the origin of splicing [1] - *Why Genes in Pieces?* - certainly remains pertinent to this day. To further sharpen the question: Why are some genomes, in particular those of multicellular eukaryotes (plants and animals), intron-rich whereas others, i.e. those of the great majority of unicellular eukaryotes, are intron-poor? In principle, accumulation of introns in genes of multicellular organisms could be considered as an adaptation that ensures evolution of organismal complexity, especially via AS. This is indeed the position taken by the proponents of the genome design hypothesis discussed in the preceding section. However, a simpler explanation that appears to be better compatible with the data has been proposed by Lynch as part of the non-adaptive theory for the evolution of complexity [42,49,50,168,169]. A simple estimate based on the number of nucleotide sites required for accurate intron excision during splicing (that is, the donor and acceptor sites and the branching point motif) shows that the power of purifying selection is sufficient to eliminate the majority of introns only in populations with a large effective population size (N_e) such as found in many unicellular eukaryotes ($N_e \sim 10^7 - 10^8$) [50,170] but not in the relatively small populations of vascular plants and vertebrates ($N_e \sim 10^5 - 10^6$ and $10^4 - 10^5$, respectively) [50,170,171]. Numerical simulations based on this estimate reveal a phase transition-like shift from intron-rich to intron-poor genomes [50,168,169] which roughly matches the observed distribution of intron densities (see Figure 2).

This non-adaptive population genetic perspective on the evolution of introns and eukaryotic gene architecture is compatible with the results of empirical reconstruction according to which the general (perhaps counter-intuitive) trend is evolution of intron-poor genomes in multiple lineages from intron-rich ancestors (see Figure 2). This

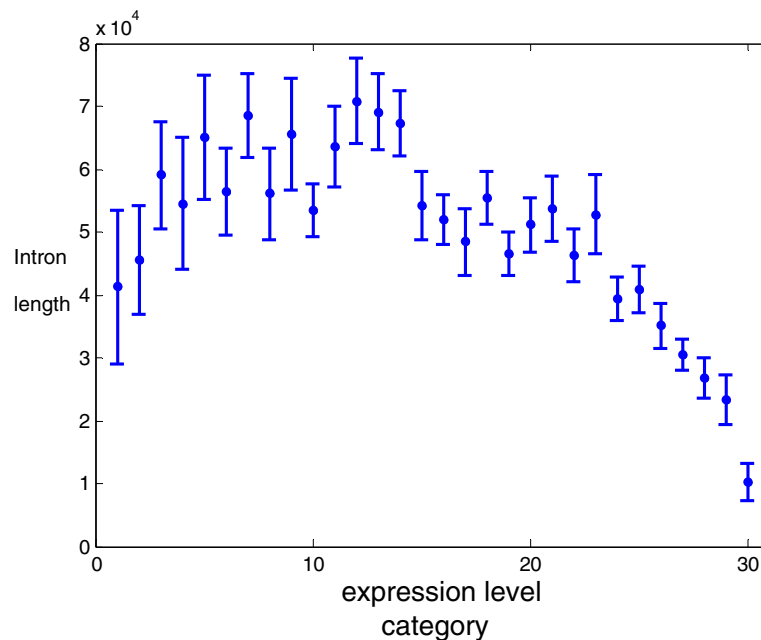


Figure 10 Total intron length as a function of expression level category. Intron length is measured in nucleotides. Expression levels are binned into 30 categories, with higher categories matching higher expression levels, as described previously [167]. Each point is the mean value for all genes in the given expression category, and the error bar indicates the standard deviation of the mean.

evolutionary trend appears to be a form of ‘genomic streamlining’ occurring in evolutionarily successful lineages that reach high effective population sizes which prevent effects of genetic drift and eliminate even slightly deleterious features such as introns. Conversely, the apparent bursts of intron gain linked to the origin of major groups of eukaryotes such as the Metazoa would coincide with population bottlenecks which are typical of such transitional epochs [42,49,50,172]. The non-adaptive population genetic concept is also compatible with the finding that intron-rich organisms possess much weaker donor splice signal than intron-poor organisms: the pressure of purifying selection in intron-rich lineages is insufficient to strictly maintain the consensus nucleotides at the donor sites [51]. A more direct analysis that compared the rates of consensus-to-variant and variant-to-consensus substitutions in the donor sites of three intron-rich lineages supported the existence of purifying selection against variants that, however, is too weak to maintain the consensus in most of the introns [52].

A major consequence of the inability of purifying selection in small populations to eliminate introns or to maintain strong donor splice signals is the accumulation of aberrant splice variants. Such error-prone splicing could eventually give rise to functional alternative splicing. Notably, the latest scenario of intron gain and loss in widespread eukaryotic genes includes only intron-rich intermediates on the path of evolution from the LECA to mammals (see above; Figure 7), with the implication

that this line of descent never went through a stage of strong purifying selection allowing continuous evolution of alternative splice variants [53].

Although the non-adaptive population genetic theory appears to be the best available conceptual framework for the evolution of eukaryotic gene architecture, splicing and introns, at least two notable problems remain outstanding. First, it is unclear why the acceptor splice signal does not follow the same trend as the donor site and is stronger in intron-rich multicellular eukaryotes than it is in intron-poor unicellular forms although the observed positive correlation between the strength of the donor splicing signal and the combined strength of the branch point signal + acceptor splice signal [117] might explain this incongruence. Second, the preservation of at least a few introns even in the most intron-poor organisms remains enigmatic because at face value the non-adaptive scenario would predict complete loss of introns and accordingly the spliceosome in multiple lineages.

Evolution of alternative splicing in coding and non-coding regions of eukaryote genes

In multicellular organisms, particularly animals, AS is a major mechanism for regulating gene expression and function [173-176]. Large-scale studies based on mapping of expressed sequence data on genomic sequences and RNAseq surveys have shown that more than 90% of human and over 40% of *Arabidopsis thaliana* and rice

genes are capable of producing multiple diverse mRNA molecules through alternative splicing [177-183].

Alternative splicing has been identified in many eukaryotic groups; however, it remains unclear whether frequent alternative splicing emerged early in eukaryotic evolution [176,184] because ancestral splice signals were weak and failed to provide for highly accurate splicing, or has evolved more recently and independently in multiple lineages via mutation of strong ancestral splice signal in multi-intron genes [33]. As pointed out in the previous section, the non-adaptive population genetic model that is in excellent agreement with the empirical reconstructions of eukaryote gene architecture evolution implies that AS evolved already at the earliest stages of eukaryote evolution through accumulation of aberrant splice variants under conditions of weak purifying selection. A further implication of this scenario is that initially all alternative transcripts were non-functional whereas functional AS evolved gradually and independently in multiple lineages, primarily those that have never gone through population bottlenecks leading to extensive loss of introns and tightening of splice signals.

The impact of alternative splicing on protein function has been studied in great detail and is generally recognized as a major source of protein diversity that greatly expands the repertoire of protein function [173-175]. A systematic comparison of 9 animal genomes from nematodes to mammals revealed that intron-flanking domains expanded faster than other protein domains [185]. Intriguingly, such mobile domains exhibited a strong preference for phase 1 introns [185-188] in contrast to the general excess of phase 0 among introns (Figure 6). This finding suggests that evolution of introns flanking mobile domains is fundamentally different from the evolution of introns in conserved portions of genes but the nature of these differences remains to be elucidated [185,187,188].

Evolutionary conservation of alternative splicing is a controversial matter. Only limited conservation of alternatively spliced (cassette) exons within mammals and within dipterans has been detected [189-193]. However, a strikingly different pattern has been reported for *Caenorhabditis* nematodes: more than 92% of cassette exons found in *C. elegans* are conserved in *C. briggsae* and/or *C. remanei* [194]. The differences in conservation between lineages might reflect differences in the fractions of functional alternative transcripts but possibly also differences in intron length and the strength of splicing signals [194].

Evolution of alternative splicing has been also analyzed in the context of splicing signals [195]. The GT dinucleotide in the first two intron positions is the most conserved element of the U2 donor splice signal. However, in a small fraction of donor signals (<1%), GT is

replaced by GC. A substantial enrichment of GC in donor signals of alternatively spliced genes has been observed in human, nematode and *Arabidopsis*, suggesting that GC signals are important for regulation of alternative splicing [196-198]. Parsimony analysis was used to reconstruct evolution of donor splice signals resulting in 298 inferred GT to GC conversion events compared to 40 GC to GT conversion events in primate and rodent genomes. Thus, there was substantial accumulation of GC donor splice signals during the evolution of mammals. Accumulation of GC sites might have been driven by selection for AS [195]. Several studies have dealt with the evolution of alternative splicing from the perspective of the evolution of splicing enhancers and silencers, and some signs of negative and positive selection have been detected [199-206].

Alternative splicing is one of the primary sources of 5'UTR transcript diversity, and in several reports the hypothesis has been put forward that these mechanisms might play an important role in orchestrating complex regulatory mechanisms within 5'UTRs [207-209]. Estimates of the number of genes with alternative 5'UTRs vary from 10% to 22% [179,210,211]. A genome-wide comparative study of 5'UTR sequences in primates and rodents revealed a much greater abundance of alternative splicing (and alternative transcription) than detected in the corresponding coding sequences, conceivably because 5'UTRs are not bound by constraints on protein structure that limit alternative splicing in coding regions [209]. Alternative regions of 5'UTRs contain numerous upstream AUG codons and short upstream open reading frames, consistent with the hypothesis that alternative events in 5'UTRs of mammalian genes contribute to the regulation of translation [209].

Functions of introns

The non-adaptive population genetic perspective on the evolution of eukaryotic gene architecture implies that introns are devoid of function, at least originally. This conjecture is compatible with the numerous analyses indicating that, beyond the splice signals, intron sequences are subject to weak purifying selection at best, or evolve in a regime that is indistinguishable from neutral evolution [212,213]. However, this (nearly) neutral background of intron evolution does not rule out the possibilities that, first, the very presence of introns affects the regulation of expression of the respective genes (presumably through the interaction with the splicing machinery) and hence their function, and second, that many introns harbor specific functional elements. Indeed, there is abundant evidence that introns are often functional at both levels.

Potential functions of introns can be separated into three categories: i) functions associated with splicing, ii) generic functions of non-coding DNA, and iii) genes

nested within introns. In addition, the possibility has been discussed that introns might act as 'catalysts' of evolution by facilitating intergenic recombination (this could be considered a variation on the theme of generic non-coding DNA functions). Experimentally demonstrated and potential functions of introns have been reviewed in detail [214,215]. Here we do not attempt a comprehensive coverage of this subject but rather briefly discuss several aspects that appear directly relevant for understanding evolution of introns and eukaryote gene structure.

Functions of introns associated with splicing

Splicing occurs before mature mRNAs are transported from the nucleus to the cytosol by the nuclear export system. Numerous studies indicate that splicing and mRNA export are directly coupled (see reviews [32,35]). Evidence of such coupling initially came from the observation that mRNAs generated by splicing are more efficiently exported than their identical counterparts transcribed from a complementary DNA [216]. This effect of splicing on export was explained by the finding that spliced mRNAs (but not cDNA transcripts) are assembled into a distinct mRNP complex that promotes efficient export [32,35,216]. This complex, or at least some of its components, has been subsequently shown to assemble adjacent to newly formed exon-exon junctions [217]. The increased export efficiency of the spliced mRNP is thought to be due to recruitment of the mRNA export factor ALY to the mRNA during the splicing reaction [218,219]. The splicing factor UAP56, which interacts directly with ALY, plays a role in recruiting ALY to the spliced mRNA [220-222]. In a subsequent step, a hand-off occurs in which the ALY/TAP interaction is established, thus delivering the mRNP to the nuclear pore for export [221]. The numerous eukaryotes that possess only a few introns in the entire genome nevertheless retain a full-fledged or partially degraded spliceosome machinery [8,65,223], suggesting the possibility that the spliceosome might have functions other than splicing as such, perhaps primarily nucleocytoplasmic trafficking. However, the transport mechanisms for numerous intron-less transcripts are not well characterized, and the possibility remains that intron-less RNAs are recruited to the export machinery via a spliceosome-independent route [32,35]. Compatible with this hypothesis, UAP56 is required for export of both spliced and intronless mRNAs [220-222,224]. In metazoan intronless mRNAs, specific mRNA sequence elements are required for export, and some of these elements associate with members of the SR family of splicing factors which are thought to mediate export of the intronless mRNA [225,226]. The SR proteins could either recruit the conserved export machinery or play a

direct role in export [226]. In both yeast and metazoans the export of intronless mRNAs also could be coupled to polyadenylation [32,35,226,227]. It has been shown that in mammalian neurons some retained introns are coupled with targeting of mRNA sequences to dendrites, apparently via so called ID sequences that represent a distinct class of SINE retrotransposons resident in the retained introns [228]. Thus, functionally relevant retention of intronic sequence might be a more general phenomenon than previously suspected.

The speed of splicing could be another important mechanism of gene expression regulation [27,28]. Analysis of minor, U12 introns (see above) suggested that their positions are conserved in orthologous genes from human and Arabidopsis to an even greater extent than the positions of the major, U2 introns [29]. The U12 introns, especially conserved ones, are concentrated in 5'-portions of plant and animal genes, whereas the U12 to U2 conversion occurs preferentially in the 3'-portions of genes. These results are compatible with the hypothesis that the high level of conservation of U12 intron positions and their persistence in genomes, despite the unidirectional U12 to U2 conversion, have to do with the role of the slowly excised U12 introns in down-regulation of gene expression [27-29,229].

As already pointed out above, introns in yeast ribosomal protein genes substantially affect the expression of these genes and contribute to the organismal fitness and stress response via mechanisms that are not yet well understood [57]. These seminal findings indicate that in many cases the regulatory functions of introns could be specific to a class of genes or even an individual gene. This conclusion is compatible with the results of an earlier study which has shown that yeast spliceosome can distinguish between different transcripts including related ones, such as paralogous ribosomal protein genes, thus providing a distinct regulation mode for expression of specific proteins [230].

Introns as functionally important non-coding DNA sequences

Compared to prokaryotes, eukaryotes possess a much greater number of multidomain proteins that substantially contribute to the functional complexity of the eukaryotic cell [187,188,231-234]. Moreover, a striking feature of eukaryotic protein architectures is the wide spread of the so-called promiscuous domains that combine with other domains much more often than expected by chance [234,235]. The 'exon theory' posits that exon shuffling via recombination within introns is an important route of evolution that in particular is responsible for the diversity of the domain architectures of multidomain proteins [39,40,236]. In the specific case of vertebrate membrane receptor proteins, this hypothesis

seems to be compatible with empirical observations: these proteins consist of multiple modules each of which typically is encoded by an individual exon [185,187,188]. However, in other classes of proteins, there is no strong preference for intron location between domains, so exon shuffling is unlikely to be a major, general mechanism of multidomain protein evolution [43,135,185,187,188,234].

Introns have the potential to serve as “enhancers” of meiotic crossing-over occurring within protein-coding genes because the probability of crossing over between segments of a coding sequence (exons) separated by long introns greatly increases compared to the same coding sequences in the absence of an intron [214,237]. This meiotic recombination between exons of two alleles of the same gene is likely to be a major factor of protein evolution through combining mutations from different alleles, “trying out” different combinations and avoiding accumulation of deleterious mutations within the same allele [1,214,237].

Trans-splicing is a special form of RNA processing whereby exons from two different primary RNA transcripts are joined end-to-end and ligated. The most common form of trans-splicing is spliced-leader (SL) trans-splicing where the leader is donated by a short SL RNA. The SL trans-splicing is widespread among some unicellular eukaryotes, in particular trypanosomes [238]. Other than trypanosomes, the only organisms known to heavily rely on SL trans-splicing for gene expression are nematodes. More than half of the pre-mRNAs in the *Caenorhabditis* nematodes are trans-spliced to one of two short leader RNAs, SL1 or SL2. This process occurs at the 5' ends of pre-mRNAs, and it is essential for the efficient processing of polycistronic pre-mRNAs [35,239-242]. The patchy distribution of trans-splicing suggests that SL trans-splicing has evolved repeatedly among eukaryotic lineages and SL precursor RNAs have readily evolved from ubiquitous small nuclear RNAs that are involved in conventional splicing [243]. Several cases of trans-splicing between different pre-mRNAs (no SL RNAs are involved) have been identified in tunicates, mammals, flies and plants (reviewed by [214,242,244,245]).

Functional elements and genes within introns

Some introns contain various regulatory elements as well as sequences involved in chromatin structure formation such as scaffold-matrix attachment regions, although it remains uncertain whether intron sequences show any substantial enrichment for regulatory and structural elements compared to other non-coding DNA [214,246]. Some long introns, especially those in 5'-terminal regions of coding sequences, might be enriched for various regulatory elements, and consequently, could be subject to purifying selection [160,247-253]. Long introns in several genes of *Oikopleura* have been shown

to contain key developmental regulators [131], and similar observations have been reported for genes involved in development of diverse metazoans [254-257] as well as associated “bystander” genes that are not known to be directly involved in development [258-261].

Many introns contain within their sequences various non-coding RNA genes, especially numerous genes for snoRNAs [262,263] and precursors of microRNAs [264,265]. Specifically, some short animal introns with hairpin formation potential, known as mirtrons, can be spliced and debranched into pre-miRNAs [266-268]. These pre-miRNAs are then cleaved by the RNase III enzyme Dicer and incorporated into typical miRNA silencing complexes [268,269].

A small fraction of introns contain nested protein-coding genes [270]. Comparative analysis of these nested genes in vertebrates, fruit flies and nematodes revealed substantially higher rates of gain of intron-embedded genes compared to loss [271]. However, the accumulation of nested gene structures is likely to represent an increase of organizational complexity of animal genomes via a neutral process given that there seem to be no functional links between the intron-contained genes and the ‘host’ genes [271]. Effectively, it seems that introns serve as neutral substrate that can be randomly colonized by various genes.

Molecular mechanisms of intron loss and gain

Mechanisms of intron loss and gain remain poorly understood. A plausible, common mechanism for intron loss could be homologous recombination between cDNAs that are produced by reverse transcription and the genomic copies of the respective genes [65,67,107-110,112]. Intron gain/loss events must be associated with a transient phase of segregating alleles either carrying or lacking the intron within natural populations [49]. Until now, only 25 transient intraspecific intron presence-absence polymorphisms have been reported, one in *Drosophila teissieri* [272] and 24 in *Daphnia pulex* [70,96]. In *Daphnia*, recently gained intron sequences were frequently associated with short repeats, suggesting a role for double-strand break repair in intron gain [96]. Analysis of several closely-related fungi revealed 74 presence-absence polymorphisms of introns [273]. Examination of the positions of these introns has suggested that extensive intron transposition among unrelated genes is the major mechanism of intron gain in the analyzed fungal genomes [273]. The existence of large families of highly similar intron sequences in these genomes suggests that certain intron sequences are much more likely to be transposed than others and that specific sequence patterns might promote intron transposition [273].

Although transposition of introns could be an important factor of intron gain for some fungi, it appears to be

a negligible route of intron evolution in nematodes, green plants, and *Daphnia* [66,94,96]. It is likely that intronization of (parts of) exons is an important source of new introns [110,114,274,275]. However, possible other routes of intron acquisition let alone their quantitative contributions in different groups of eukaryotes have not been characterized in any detail [72,275,276]. A striking case of massive intron gain has been discovered in the course of genome analysis of the marine picoeukaryotic alga *Micromonas pusilla* [277]. The introns of numerous *Micromonas* genes contain repeat sequences that are absent from orthologous genes in closely-related genomes. These abundant 'introner' elements (9904 introners total) were located within introns, extended nearly to donor and acceptor sites, and lacked known characteristics of transposable elements [277]. The high abundance of introner elements suggests that these elements are either functionally important or resistant to purging, or both. It should be noted that mechanisms of massive intron gain events (for example, in the earliest eukaryotes [148,275] could well be different from mechanisms of relatively slow intron acquisition process in various extant eukaryotes [66,94,96,273] which makes delineation of mechanisms of intron gains an even more difficult problem.

Origin and evolution of spliceosomal introns: a synthetic concept

The evidence presented here and elsewhere [42,45,53,62,84] supports a 'numerous introns early in eukaryote evolution' view. The discovery of introns in jacobids [278] and other excavates [4,5] is compatible with this concept. Even more strikingly, approximately 60% of the introns in the parabasalid *Trichomonas vaginalis* occupy the exact position of an intron in an orthologous gene from at least one other eukaryotic lineage [279], and similar observations have been made for the free-living excavate *Naegleria gruberi* [6]. Most importantly, probabilistic reconstructions of intron gain and loss provide consistent and by now compelling evidence that ancestral eukaryotic forms including the LECA possessed intron-rich genes, with intron densities comparable to those in the most intron-rich modern organisms such as mammals [53,141]. These findings have fundamental consequences for our understanding of the evolution of eukaryotes and possibly of the ultimate origin of the eukaryotic cellular organization [46,280,281].

It appears likely that the emergence of the eukaryotic cell or the initial stages of its evolution involved, among other radical innovations, a catastrophic intron invasion (Figure 9) [46]. Structural similarities between the terminal regions of spliceosomal introns and those of self-splicing Group II introns (retro-transcribing elements) leave essentially no doubt in the existence of a direct

evolutionary connection between the two classes of introns [282]. Moreover, the elements of Group II introns involved in the autocatalytic splicing reaction apparently also gave rise to the spliceosomal small RNAs [282,283]. Thus, at an early stage in the evolution of eukaryotes, an irreversible transition apparently took place from autocatalytic splicing to splicing mediated by a universal trans-acting catalyst (the spliceosome). This transition involved the split of the ancestral Group II intron structure into the catalytically inert spliceosomal introns and the catalytically active RNA moiety of the spliceosome that was also accompanied by the degradation of the reverse transcriptase open reading frame within introns [280].

It appears most likely that the Group II intron invasion was triggered by the establishment of the endosymbiosis between an α -proteobacterium and an archaeal host (Figure 9). Notably, α -proteobacteria typically contain in their genomes a relatively large number of Group II elements compared to other bacteria [284]. Upon the endosymbiont invasion of the archaeal host, the symbiont's Group II introns might have been 'unleashed', in part due to repeated lysis of the symbiotic cells (the evolving mitochondria) [280]. At the fundamental evolutionary-theoretical level, the tolerance of the emerging eukaryotic cell to such an invasion could be potentially explained by a population bottleneck which severely limited the efficacy of purifying selection [50,280,285].

Indeed, it has to be emphasized that Group II introns are typical mobile elements that actively spread around the host genome when given a chance by weakness of purifying selection pressure.

However, at the mechanistic level, the adaptation of the early eukaryotes to the swarms of genomic parasites (if this is what introns are, Figure 9), which severely compromised the integrity of their genomes, an adaptation that apparently involved rapid evolution of the dauntingly complex spliceosome, remains an intriguing enigma. The intron invasion, probably spawned by the mitochondrial endosymbiont (Figure 9), could have led to a peculiar, intron-dominated genome architecture of the early eukaryotic, with up to 80% of the genomic DNA comprised of introns [286]. This genome structure could be sustainable only under a severe population bottleneck and might have critically contributed to the emergence of the principal features of the eukaryotic cell [46,286]. The evolution of the signature features of eukaryotic cell organization, such as the endomembrane apparatus including the nucleus, the nonsense-mediate decay system and the ubiquitin system, can all be conceptualized as multiple levels of defense against the deleterious effects of the intron invasion [46,172]. Furthermore, the early, mobile introns could have triggered the

proliferation of multidomain proteins via homologous recombination between introns in different genes. Obviously, most of such events would be strongly deleterious but some might have created potentially useful domain combinations without losing much important information, and thus would be picked by selection. Introns also created the potential for controlled alternative splicing (see above), a mechanism that came to prominence at a later stage of eukaryotic evolution and made a crucial contribution to the evolution of complexity in multicellular organisms. To summarize, the intron invasion that was probably concomitant with the emergence of the first eukaryotic cells can and probably should be envisaged as one of the key factors of eukaryogenesis.

Evolution of exon-intron structure of eukaryotic genes had been long considered in the context of the “introns-early” vs. “introns-late” debate [39-42]. Although the original introns-early idea is hard to reconcile with the absence of spliceosomal introns (and the spliceosome itself) in prokaryotes and the absence of conserved intron positions in ancient eukaryotic paralogs (Figure 8) [148], this concept can be easily restated in more realistic (even if less dramatic) terms. Specifically, the entirety of the observations discussed above, strongly suggests that the spliceosomal introns originated from self-splicing Group II introns which invaded eukaryotic genes (or perhaps more precisely, genes of the archaeal host of the proto-mitochondrial endosymbiont) concomitantly with or at the latest shortly after the origin of the eukaryotic cell. As indicated by evolutionary reconstructions, subsequent evolution involved mostly lineage-specific loss of introns punctuated with a few episodes of new gains. Under this scenario, although there is no evidence of existence of modern-type spliceosomal introns (or spliceosomes) prior to the origin of eukaryotes, their ancestors were ancient mobile elements that probably co-existed with cellular life forms throughout their evolution or possibly even antedated modern cells [287]. Thus, although the ‘exon hypothesis’ and the original idea that the first genes contained multiple introns do not seem to be supported by any evidence, the evolutionary lineage leading to spliceosomal introns indeed could be as old as some of the first replicating genetic entities.

Conclusions

The incentive to write this review was the conviction of the authors that, after 30 years of turmoil, a degree of clarity has been reached in the study of the evolution of eukaryotic gene architecture. This progress has been achieved through the combination of comparative analysis of numerous, diverse genomes of eukaryotes, probabilistic reconstructions of intron gains and losses, and the non-adaptive population genetic theory of evolution of genomic complexity. It now appears well established

that evolution of eukaryotes as a whole as well as evolution of each of the eukaryotic supergroups started with intron-rich genomes with relatively weak, error-prone splice signals. The evolution of these ‘cumbersome’ ancestral genomes was predicated by population bottlenecks that accompany evolutionary transitions and entail weak purifying selection that is incapable of purging introns or evolving efficient splicing. Subsequent evolution of eukaryotes followed the divergent paths of genome streamlining which led to elimination of the majority of introns and tightening of the splice signals, or genome complexification which involved evolution of functional alternative splicing and other intronic functions. The streamlining route is characteristic of many lines of descent that enjoyed evolutionary success and reached large effective population size (primarily unicellular eukaryotes and some fungi), whereas the forms that never achieved high efficiency of purifying selection (primarily multicellular animals and plants) followed the path to complexity.

The elucidation of the general scenario of evolution of eukaryote gene architecture by no account implies that the main problems in the study of intron evolution and function have been solved. Quite the contrary, fundamental questions remains wide open.

What are the mechanisms of intron loss and gain? There is very little direct evidence of any. A consensus seems to exist regarding the role of reverse transcription in intron loss although even this mechanism badly needs experimental corroboration. As for the mechanisms of intron gain, indications of the involvement of double-strand break repair notwithstanding, the study of this key problem has not even started in earnest.

What are the sources of new introns? It is clear that duplication of pre-existing introns is a negligible route of intron evolution in many eukaryotic lineages although it seems to be important in some, whereas intronization of (parts of) exons appears to be a significant contribution throughout the evolution of eukaryotes. However, possible other routes of intron acquisition let alone their relative quantitative contributions remain unknown.

What is the general role of introns in gene expression and function (if any) and why is it the case that new genes (such as those acquired from chloroplasts in plants) are saturated by introns at an apparent high rate? And a related question: why do a handful of introns (and with it the spliceosome, sometimes partially degraded) survive in the great majority of even the most streamlined eukaryotes? At best, only most general and largely speculative answers to these key questions are currently available. These are hard questions, and the only hope to obtain satisfactory answers is to combine comprehensive phylogenomic analysis with population genetic models and extensive experimentation.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

IBR, LC, MC and EVK wrote the manuscript which was read, edited, and approved by all authors.

Reviewers' comments

Reviewer #1: Dr. I. King Jordan, Georgia Institute of Technology

Igor Rogozin and colleagues have written a comprehensive, synthetic and compelling review mainly covering the 'evolution' of spliceosomal introns. As in Darwin's famous tome, the notion of 'origin' is actually given short shrift, but I will come to that point later. In any case, this ambitious work benefits from the broad perspective that the authors have gained over years of investigating the subject. Along with this perspective comes some inevitable bias, or perhaps it is more fair to say a favored world-view, as described below with respect to the authors position on 'introns-kind-of-early'. But this does not represent a liability of the work in my opinion; the authors are clearly entitled to their views, and the conclusions they draw appear to be both nuanced and well-supported by the data. They cover a lot of ground herein and strike a nice balance between thoroughly reviewing the relevant literature and elucidating the most salient points from the large body of work on the subject. One of the major conclusions of this review relates to a resolution, or compromise really, of the 'introns-early' versus 'introns-late' debate that consumed the field for many years. The authors champion a merger of these two hypotheses into the 'many introns early in eukaryotic evolution' view, whereby the earliest eukaryotic lineages contained genomes that were already loaded with many introns and subsequent evolution was dominated by intron loss.

The parts of the review that cover the origin of spliceosomal introns are the most speculative and least supported. This is not a critique per se; it may simply be the case that the study of origins must always be more speculative than the study of evolution. According to the 'many introns early in eukaryotic evolution' hypothesis, the earliest eukaryotic genomes were formed via massive intron invasion that resulted in genomes consisting of up to 80% intronic DNA. Crucially, the authors hold that this invasion was probably facilitated by low effective population sizes and the corollary weak purifying selection, following the influential Michael Lynch model for the non-adaptive evolution of eukaryotic genome complexity. This model accounts for population level dynamics but neglects the internal dynamics of the genome. If spliceosomal introns indeed evolved from Group II introns, as the authors maintain, then the initial intron invasion of eukaryotic genomes would have been driven, to some extent, by a kind of selfish genetic element with its own internal drive mechanism to replicate within the genome. In theory, such selfish replicators can efficiently increase in copy number even in the face of a selective cost to the host. Therefore, the early origin of introns may be attributed to an active internally driven process, rather than a solely passive drift related process, i.e. a mechanism akin to the molecular-drive concept of Gabriel Dover or the mutation bias emphasized by Arlin Stoltzfus. Such an active replicative process inside the genome could have actually outpaced selection's ability to contain it. The authors actually touch on this notion, when they speculate as to whether introns are genomic parasites and how the host may have evolved the spliceosome as an adaptive response to intron invasion, but an explicit connection between their selfish drive to replicate and the origin of introns is not made.

Authors' response: We agree on all accounts. Yes, it comes with the territory: discussion of origins is inevitably more speculative than the analysis of subsequent evolution. More importantly, the role of the active mobility of Group II introns certainly must not be under-appreciated, and we explicitly point this out in the revised manuscript: 'Indeed, it has to be emphasized that Group II introns are typical mobile elements that actively spread around the host genome when given a chance by weakness of purifying selection pressure.'

One specific suggestion as to how the work can be improved relates to the abstract. Currently, the abstract is very short and concise, whereas the manuscript is rather long and presents a lot of material. I think it would be helpful to provide a more detailed abstract that specifically enumerates the authors' most important points, something more like of a summary of the last two sections of the manuscript.

Authors' response: We fully agree, the original short abstract resulted from a misunderstanding regarding the limits on abstract length in review articles. In the revised article, the abstract was substantially expanded.

Reviewer #1: Dr. I. King Jordan, Georgia Institute of Technology (additional comment on the revised version of the manuscript)

I have re-reviewed the manuscript of Rogozin et al. I am satisfied with the changes made, for the most part, and I recommend that the paper be accepted for publication in *Biology Direct* after the following point is addressed.

I would like the authors to elaborate just a bit on their response the first comment that I made, in particular with respect to the connection between Group II intron dynamics and the evolution (emergence) of introns. I think I may have rambled a bit in my original comment and was not explicit enough. I would urge the authors to have a look at the manuscript of Donal Hickey from *Genetics* (Hickey 1982 101; 519), which makes the point much better than I did in my comment. The population genetics models in the manuscript may be a bit simplistic by this time, but I think the ideas contained therein are highly relevant to their own work. In particular, Hickey makes an explicit connection between the genome dynamics of mobile elements, host selection pressure and the evolution of introns. The basic idea is that mobile genetic elements can spread in a population even in the face of a fitness cost to the host, and this kind of process could have resulted in the emergence and spread of introns. Below, I provide a comment in response to the authors' response to my first comment in an attempt to facilitate further discussion and consideration of this issue.

Authors' response: We agree on all accounts. Yes, it comes with the territory: discussion of origins is inevitably more speculative than analysis of subsequent evolution. More importantly, the role of the active mobility of Group II introns certainly must not be under-appreciated, and we explicitly note in the revised manuscript: 'Indeed, it has to be emphasized that Group II introns are typical mobile elements that actively spread around the host genome when given a chance by weakness of purifying selection pressure.'
*Response: I would like the authors to further consider the possibility that mobile elements (such as Group II introns) can increase in frequency in a population, even when they impose a fitness cost on their host organisms, owing to the fact that a replicative transposition process results in a biased transmission rate relative to host genes. This idea was introduced by Donal Hickey 30 years ago, and he also connected this point to the evolution of introns (Hickey 1982 *Genetics* 101: 519). In other words, it is not simply a matter of weak purifying selection allowing active spread of the elements, but an effect of the element mutational dynamics introducing directional bias in the evolutionary process. This idea is very much analogous to the notion that mutation bias in the broader sense can be a cause of direction in evolution (e.g. see Yampolsky and Stoltzfus 2001 *Evol Dev* 3: 73).*

Authors' response: We agree that mutational dynamics of selfish element could be an important driver of their spread. We think that once this additional exchange with the reviewer is published, the emphasis on this issue will be adequate.

Reviewer #2. Dr. Tobias Mourier, University of Copenhagen (nominated by Dr Anthony Poole)

This review provides a comprehensive overview of the current knowledge of intron evolution in eukaryotic genomes.

The advent of numerous eukaryotic genomic sequences has consistently supported the 'many introns early in eukaryotic evolution' concept, as evident from the manuscript. But surely this hypothesis is not a merger from the introns early/late/first ideas (as the authors write in the "Intron-early, introns-late, introns-first . . ." section). All recording of spliceosomal intron features comes from eukaryotic genomes, and regardless of how many eukaryotic genomes are sequenced, extant spliceosomal intron features will never allow one to synthesize past LECA.

In the end of the manuscript, the authors present a scenario proclaiming that an intron-rich LECA is not inconsistent with the introns-late hypothesis. This is not a problem, but the structure of the manuscript may give the impression that this is a conclusion (or synthesis) directly from the current knowledge of eukaryotic gene architecture (that is nicely reviewed in the preceding text).

Authors' response: Actually, we do believe that the synthesis we present in the section of the review preceding the Conclusions follows from the comparative genomic results reviewed in the preceding sections. Certainly, not all parts of the article directly contribute to this synthesis: for instance, the discussion of the functional roles of introns is only tangentially relevant here albeit important in other respects. Nevertheless, we do maintain that in this section we present major implications of the comparative genomic study of eukaryotic gene structure.

The review presents an overview of the comparative approaches taken to delineate intron-exon structures during evolution. The basis for such comparative analyses is well-aligned sequences around splice sites. If intron-exon structures to some extent evolve via mechanisms such as alternative splicing and intronization of exonic sequence, should this not result in sequences that are unlikely to meet the criteria for being included in the above analyses? I think it would be relevant to discuss the implications of this.

Authors' response: This issue is discussed in the section 'Evolutionary conservation of intron positions and routes of gene architecture evolution of eukaryotes'.

Section "Functional elements and genes within introns"

When discussing intronic RNA genes, I'm surprised there is no mentioning of the classical connection between vertebrate snoRNAs and introns (and perhaps even the existence of genes with non-coding exons and introns encoding snoRNAs, (e.g. Tycowski et al., *Nature* 1996).

Authors' response: Yes, this certainly is an important theme, and we added it to the section 'Functional elements and genes within introns'.

Very minor points:

Page 11: "whereas the remaining 6 Xist" rather than "whereas remaining 6 Xist"

Page 14: "and so does the strength" rather than "and so does and the strength"

Page 18: should "introns are inserts or fixed" read "introns are inserted or fixed"?

Authors' response: All corrected, we appreciate the reviewer's attention to these points.

Reviewer #3. Dr. Manuel Irimia, University of Toronto (nominated by Dr Anthony Poole)

Rogozin et al. have put together an impressively comprehensive review on the origin and evolution of splicesomal introns that will certainly become a major reference in the field. Overall, I found it easy and entertaining to read, as well as informative. I have only a few comments and suggestions, often regarding further literature, that I hope can help to improve the piece (listed according to their appearance in the main text):

Authors' response: We appreciate Dr. Irimia's close attention to the details of this article. As detailed below, we found most of the suggestions fully pertinent and modified the manuscript accordingly.

1) P3: The paragraph on splice site consensus sequences could provide a more detailed portrait of canonical intron signals across eukaryotes. For example, not all eukaryotes have polyT tracts between the branch point (BP) and the 3' AG, and some fungal species even have polyT tracts upstream the BP (see Bon et al., *Nucleic Acids Res* 2003; Irimia and Roy, *PLoS Genetics* 2008). Also, some extremely intron-poor species intriguingly have strict GTATGT as consensus 5' sequence (including yeast), which may be worth pointing out. Finally, the 3' consensus is closer to YAG than to CAG, at least in most species.

Authors' response: We added discussion of this issue to the revised text.

2) P5: I found the (exciting) discussion on the ancestry of U2 vs. U12 too short and a bit imbalanced. Personally, I think it is a good idea that the authors give their authoritative opinion/preference on this kind of discussions, but the opposite arguments should also be presented extensively. In this case, I think the arguments supporting an ancestral origin of U12 (i.e. lack of evidence for conversion from U2 to U12, argued higher similarity of U12 to type II introns, etc.) should be fully developed.

Authors' response: In our view, the questionable greater similarity of U12 introns to Group II introns does not immediately imply ancestral status of U12 introns. We added to the text 'it might be tempting to speculate that the ancestral introns were of the U12 type (for example, see discussion by the reviewer #3 below) but have been subsequently converted to U2 introns.'

3) P8: Pleiss et al. (*PLoS Biol* 2007) may be added supporting a global regulatory function of introns in yeast.

Authors' response: We added discussion of this important work to the section 'Functions of introns associated with splicing'.

4) P9: I missed a more comprehensive and complete review of the literature on the genome-wide dynamics of intron gain and loss in this section. For example, on the general slow paucity of intron gain, I missed references on vertebrates (Loh et al., *MBE* 2008; and actually ref 72 is incorrect: Coulombe-Huntington and Majewski, *Genome Res* 2007), flies (actual ref 72), plants (Roy and Penny, *MBE* 2007), apicomplexa (Roy and Hartl, *Genome Res* 2006; Roy and Penny, *Genome Res* 2006), Entamoeba (Roy et al., *MBE* 2006), Fungi (Nielsen et al., *Plos Biol* 2004; Stajich et al., *Genome Biol* 2007; in *Aspergillus* (Zhang et al., *JME* 2010)). On the opposite side: tunicates (Seo et al., *Science* 2001; Edvardsen et al., *JME* 2004), diatoms (Roy and Penny, *MBE* 2007), mitochondrial transfers (Ahmadinejad et al., *BMC Evol Biol* 2010). Given the overall level of comprehensiveness and detail of this review and that, as I said above, it is very likely to become a major reference in the field, I think it would be important to cite all relevant references in the main text, in particular from such an important and prolific subtopic.

Authors' response: There is indeed a lot of evidence on specific events in individual lineage. We appreciate their importance but it is hardly possible to discuss 'everything' in detail. That said, the revised version of the review cites all the references pointed out by the reviewer.

5) P9: when commenting on ref 61, the use of the word "dispute" may give the impression that there is an ongoing controversy or a difference in opinions between the authors, which I guess is really not the case. Ref. 74 showed that most reported gains in ref. 61 were indeed losses by adding more species to the analysis that were not available by the time of the original study. This may not be clear to general readers that have not followed the specialized literature.

Authors' response: We added this explanation to the text.

6) P11: the authors may wish to mention here the recent work by Cabili et al. (*Genes Dev* 2011), which describe >8,000 lincRNA genes, with an average of ~1.9 introns per Kbp and that are extensively alternatively spliced, with 2.3 isoforms per gene.

Authors' response: We added a brief description to the text.

7) P14: I was quite surprised to read that the sequences at the 3' of the intron behave completely different from those at the 5'. Many of the extremely intron-poor species (although not all, in this case) that show strict 5' splice site consensus also have very strict BPs, and sometimes even very constrained branch-point-to-AG distances (Irimia and Roy, *Plos Genetics* 2008). I guess this apparent contradiction is due to the fact that these species are all missing from the analysis by Iwata and Gotoh (represented in Figure 5), which is strongly biased towards multicellular organisms, and I suspect that the inclusion of the intron-poor species would fully disrupt the observed negative correlation. In my opinion, this section should be modified to give a more complete view of the evolution of the 3' intronic signals (more like 3-4 qualitatively different behaviors related to, but not fully determined by, intron densities). Also, I recommend removing Figure 5 or making a new one using a more complete eukaryotic taxon sampling.

Authors' response: We added a list of species to the legend. Robust estimation of the information content require hundreds of splice signals, so it is impossible for the extremely intron-poor species. This is why these species are missing from the analysis of Iwata and Gotoh, and accordingly, from our Figure 5. We believe that it is fully legitimate to present only the data for those organisms that possess enough introns for meaningful statistical analysis. Furthermore, there is no contradiction at all between the observation that some extremely intron-poor species possess strict 5' splice site and also have very strict BPs and the positive correlation between the strength of the donor splice signal and the combined strength of the branch point signal + the acceptor splice signal emphasized in the present article.

8) P14: also related to splicing signals, it would be interesting to include a comment on the effect of intron size on splicing signals (long introns have stronger boundaries, species with extremely short introns often have very weak signals (e.g. paramecium, *B. natans* nucleomorph), etc.).

Authors' response: The effect of intron size is complicated. We added discussion of this issue to the section on 'Evolution of splicing signals, protosplice sites, and intron phase distribution.'

9) Figure 6 may be a bit unclear and "too raw" for non-specialists.

Authors' response: We included an additional explanation in the legend: "An excess of protosplice sites in phase 0 is noticeable, however the 'protosplice site' hypothesis, which posits that introns are randomly inserted into protosplice sites, is unable to fully explain the observed over-representation of phase 0 introns."

10) P26: the authors may want to point out from the beginning that the "two competing hypothesis" they present are not necessarily mutually exclusive.

Authors' response: Added to the text as suggested.

11) P28: ref. 158 also concludes that alternative splicing has emerged early in eukaryotic evolution, so it should be cited along with ref. 166 and not with 28.

Authors' response: Modified as suggested.

12) P31: more references may be added supporting the low conservation of alternative splicing in mammals (currently only one, from 2003, is given, but several studies have reached similar conclusions). Similarly, many other studies have dealt with the evolution of alternative splicing from the perspective of the splicing signals, not only regarding GC splicing donor sequences (e.g. evolution of ESEs and ESSs (Parnley et al., *MBE* 2006; Ke et al., *Genome Res* 2008; Irimia et al., *PLoS One* 2009) and their polymorphism in human populations (Stallings-Mann et al., *PNAS* 1996;

Stanton et al., PNAS 2003; Fairbrother et al., PLoS Biol 2006; Carlini and Genut, JME 2006; Coulombe-Huntington et al., Plos Genetics 2009).
Authors' response: A brief discussion and references added as suggested.
13) P32: perhaps the section "Functions of introns" would fit better before the section on alternative splicing (since the latter is one of those functions).
Authors' response: Alternative splicing is not exactly a function of introns, rather a mechanism of modulation of protein and RNA function. In the functional section we addressed specific functions of intron sequences. This might be debatable but we consider the original order of the sections acceptable.
14) P33: the authors may want to add that some spectacular, functional exceptions are known to the general case that splicing occurs before mRNA is exported to the cytoplasm. For example, Buckley et al. (Neuron 2011) describe the case of some transcripts with retained introns, which drive subcellular location of the transcripts to the dendrites due to the presence of a particular transposable element within their sequence.
Authors' response: We appreciate the reviewer bringing our attention to this exciting work. Cited and briefly discussed.
15) P34: the catalog of U12 introns by Alioto (Nucleic Acids Res 2007) could be referenced here.
Authors' response: Cited as suggested.
16) P35: I think it could be useful to make a clearer distinction between Splice Leader (SL) trans-splicing and trans-splicing between two different genes from the beginning of the paragraph (I found it a bit confusing now). Also, the authors may wish to cite a very elegant analysis searching for trans-splicing in *Drosophila* using RNAseq on hybrids (McManus et al., PNAS 2010).
Authors' response: We agree and have included a brief discussion and references as suggested.
17) P36: in this subsection I missed a paragraph on the (predictable and predictive) association between long introns and the presence of functional elements. For example, Denoeud et al. (Science 2011) found that the few genes with long introns in *Oikopleura* are enriched for key developmental regulators, and that those introns likely contain regulatory information. This has also been observed for many other developmental genes across metazoans [e.g. Shh (Muller et al., Development 1999), FoxP1 and Dach (Sandelin et al., BMC Genomics 2004); Gli3 (Abbasi et al., PLoS One 2007), Meis genes (Irimia et al., GBE 2011), etc.] and for associated non-developmental genes ("bystander" genes) (e.g. Woolfe et al., PLoS Biol 2005; McEwen et al., Genome Res 2006; Kikuta et al., Genome Res 2007; Engstrom et al., Genome Res 2007), with exciting implications for the evolution of genome architecture. Also, supporting the presence of regulatory elements, higher sequence conservation is often found in longer introns (Bergman and Kreitman, Genome Res 2001; Parsch, Genetics 2003; Hadrill et al., Genome Biol 2005; Marais et al., Genetics 2005; Halligan and Keightley, Genome Res 2006; Parsch et al., MBE 2010).
Authors' response: Brief discussion and references included as suggested.
18) P38: the authors may add the report by Curtis and Archibald (Curr Biol 2010) to the list of different sources of spliceosomal introns.
Authors' response: Cited as suggested.

Reviewer #4. Dr. Fyodor Kondrashov, Center for Genome Regulation, Barcelona

This is a straightforward and extensive review of everything that is known about the evolution of introns and then some more. I do not have much to add in addition to what the authors have already said. The only thing that I am left wondering about after reading this review is whether or not the authors think that Group II introns in LECA were involved in the transport of mitochondrial precursor genes into what is now the cytoplasm across the novel intracellular membrane. In light of the previous reviews I would leave it up to the authors to space and moderate the level and format of speculation, even though I believe that the nice synthesis the authors have produced make the review more interesting and useful.
Authors' response: We appreciate this comment. We are not entirely clear about the exact meaning of the reviewer's idea regarding mitochondrial genes. Is this about transfer of genes from the mitochondrial to the nuclear genome? If so, the possibility of involvement of the reverse transcriptase activity of Group II introns is intriguing but in the absence of specific evidence, one would think the main route was DNA recombination.

Acknowledgements

We thank Ben Busby, Nicholas Dibb, Cedric Feschotte, Michael Lynch, Masatoshi Nei, Scott Roy, Arlin Stoltzfus, and Yuri Wolf for useful discussions. This work was supported in part by the Intramural Research Program of the

National Library of Medicine at National Institutes of Health/DHHS (the NIH grant Z01LM000073-12), by a research grant from the National Sciences and Engineering Research Council of Canada, and by the European Union Marie Curie International Reintegration Grant (PIRG05-GA-2009-248639).

Author details

¹National Center for Biotechnology Information NLM/NIH, 8600 Rockville Pike, Bldg. 38A, Bethesda, MD 20894, USA. ²Department of Genetics, The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Edmund J. Safra Campus, Givat Ram, Jerusalem 91904, Israel. ³Department of Computer Science and Operations Research, Université de Montréal, Montréal, Québec, Canada.

Received: 8 December 2011 Accepted: 15 March 2012

Published: 16 April 2012

References

1. Gilbert W: **Why genes in pieces?** *Nature* 1978, **271**(5645):501.
2. Jurica MS, Moore MJ: **Pre-mRNA splicing: awash in a sea of proteins.** *Mol Cell* 2003, **12**(1):5-14.
3. Nilsen TW: **The spliceosome: the most complex macromolecular machine in the cell?** *Bioessays* 2003, **25**(12):1147-1149.
4. Nixon JE, Wang A, Morrison HG, McArthur AG, Sogin ML, Loftus BJ, Samuelson J: **A spliceosomal intron in *Giardia lamblia*.** *Proc Natl Acad Sci U S A* 2002, **99**(6):3701-3705.
5. Simpson AG, MacQuarrie EK, Roger AJ: **Eukaryotic evolution: early origin of canonical introns.** *Nature* 2002, **419**(6904):270.
6. Fritz-Laylin LK, Prochnik SE, Ginger ML, Dacks JB, Carpenter ML, Field MC, Kuo A, Paredes A, Chapman J, Pham J, et al: **The genome of *Naegleria gruberi* illuminates early eukaryotic versatility.** *Cell* 2010, **140**(5):631-642.
7. Lane CE, van den Heuvel K, Kozera C, Curtis BA, Parsons BJ, Bowman S, Archibald JM: **Nucleomorph genome of *Hemiselmis andersenii* reveals complete intron loss and compaction as a driver of protein structure and function.** *Proc Natl Acad Sci U S A* 2007, **104**(50):19908-19913.
8. Collins L, Penny D: **Complex spliceosomal organization ancestral to extant eukaryotes.** *Mol Biol Evol* 2005, **22**(4):1053-1066.
9. Rosbash M, Seraphin B: **Who's on first? The U1 snRNP-5' splice site interaction and splicing.** *Trends Biochem Sci* 1991, **16**(5):187-190.
10. Du H, Rosbash M: **The U1 snRNP protein U1C recognizes the 5' splice site in the absence of base pairing.** *Nature* 2002, **419**(6902):86-90.
11. Carmel I, Tal S, Vig I, Ast G: **Comparative analysis detects dependencies among the 5' splice-site positions.** *RNA* 2004, **10**(5):828-840.
12. Umen JG, Guthrie C: **A novel role for a U5 snRNP protein in 3' splice site selection.** *Genes Dev* 1995, **9**(7):855-868.
13. Chiara MD, Palandjian L, Feld Kramer R, Reed R: **Evidence that U5 snRNP recognizes the 3' splice site for catalytic step II in mammals.** *EMBO J* 1997, **16**(15):4746-4759.
14. Dibb NJ, Newman AJ: **Evidence that introns arose at proto-splice sites.** *EMBO J* 1989, **8**(7):2015-2021.
15. Dibb NJ: **Proto-splice site model of intron origin.** *J Theor Biol* 1991, **151**(7):405-416.
16. Bon E, Casaregola S, Blandin G, Llorente B, Neugeglise C, Munsterkotter M, Guldener U, Mewes HW, Van Helden J, Dujon B, et al: **Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns.** *Nucleic Acids Res* 2003, **31**(4):1121-1135.
17. Irimia M, Roy SW: **Evolutionary convergence on highly-conserved 3' intron structures in intron-poor eukaryotes and insights into the ancestral eukaryotic genome.** *PLoS Genet* 2008, **4**(8):e1000148.
18. Rogozin IB, Milanesi L: **Analysis of donor splice sites in different eukaryotic organisms.** *J Mol Evol* 1997, **45**(1):50-59.
19. Churbanov A, Rogozin IB, Deogun JS, Ali H: **Method of predicting splice sites based on signal interactions.** *Biol Direct* 2006, **1**:10.
20. Jackson U: **A reappraisal of non-consensus mRNA splice sites.** *Nucleic Acids Res* 1991, **19**(14):3795-3798.
21. Hall SL, Padgett RA: **Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites.** *J Mol Biol* 1994, **239**(3):357-365.
22. Dietrich RC, Inconvaia R, Padgett RA: **Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns.** *Mol Cell* 1997, **1**(1):151-160.

23. Burge CB, Padgett RA, Sharp PA: **Evolutionary fates and origins of U12-type introns.** *Mol Cell* 1998, **2**(6):773–785.
24. Russell AG, Charette JM, Spencer DF, Gray MW: **An early evolutionary origin for the minor spliceosome.** *Nature* 2006, **443**(7113):863–866.
25. Davila Lopez M, Rosenblad MA, Samuelsson T: **Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components.** *Nucleic Acids Res* 2008, **36**(9):3001–3010.
26. Lin CF, Mount SM, Jarmolowski A, Makalowski W: **Evolutionary dynamics of U12-type spliceosomal introns.** *BMC Evol Biol* 2010, **10**:47.
27. Patel AA, McCarthy M, Steitz JA: **The splicing of U12-type introns can be a rate-limiting step in gene expression.** *EMBO J* 2002, **21**(14):3804–3815.
28. Patel AA, Steitz JA: **Splicing double: insights from the second spliceosome.** *Nat Rev Mol Cell Biol* 2003, **4**(12):960–970.
29. Basu MK, Makalowski W, Rogozin IB, Koonin EV: **U12 intron positions are more strongly conserved between animals and plants than U2 intron positions.** *Biol Direct* 2008, **3**:19.
30. Basu MK, Rogozin IB, Koonin EV: **Primordial spliceosomal introns were probably U2-type.** *Trends Genet* 2008, **24**(11):525–528.
31. Bergert SM: **Exon recognition in vertebrate splicing.** *J Biol Chem* 1995, **270**(6):2411–2414.
32. Maniatis T, Reed R: **An extensive network of coupling among gene expression machines.** *Nature* 2002, **416**(6880):499–506.
33. Ast G: **How did alternative splicing evolve?** *Nat Rev Genet* 2004, **5**(10):773–782.
34. Schellenberg MJ, Ritchie DB, MacMillan AM: **Pre-mRNA splicing: a complex picture in higher definition.** *Trends Biochem Sci* 2008, **33**(6):243–246.
35. Soller M: **Pre-messenger RNA processing and its regulation: a genomic perspective.** *Cell Mol Life Sci* 2006, **63**(7–8):796–819.
36. Reed R: **Initial splice-site recognition and pairing during pre-mRNA splicing.** *Curr Opin Genet Dev* 1996, **6**(2):215–220.
37. Niu DK: **Exon definition as a potential negative force against intron losses in evolution.** *Biol Direct* 2008, **3**:46.
38. Doolittle WF: **Introns-early.** *Nature* 1978, **272**:581–581.
39. Gilbert W: **The exon theory of genes.** *Cold Spring Harb Symp Quant Biol* 1987, **52**:901–905.
40. Gilbert W, Glynias M: **On the ancient nature of introns.** *Gene* 1993, **135**(1–2):137–144.
41. Logsdon JM Jr: **The recent origins of spliceosomal introns revisited.** *Curr Opin Genet Dev* 1998, **8**(6):637–648.
42. Lynch M, Richardson AO: **The evolution of spliceosomal introns.** *Curr Opin Genet Dev* 2002, **12**(6):701–710.
43. Stoltzfus A, Spencer DF, Zuker M, Logsdon JM Jr, Doolittle WF: **Testing the exon theory of genes: the evidence from protein structure [see comments].** *Science* 1994, **265**(5169):202–207.
44. Stoltzfus A, Logsdon JM Jr, Palmer JD, Doolittle WF: **Intron "sliding" and the diversity of intron positions.** *Proc Natl Acad Sci U S A* 1997, **94**(20):10739–10744.
45. Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV: **Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution.** *Curr Biol* 2003, **13**(17):1512–1517.
46. Koonin EV: **The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate?** *Biol Direct* 2006, **1**:22.
47. Poole AM, Jeffares DC, Penny D: **The path from the RNA world.** *J Mol Evol* 1998, **46**(1):1–17.
48. Penny D, Hoepfner MP, Poole AM, Jeffares DC: **An overview of the introns-first theory.** *J Mol Evol* 2009, **69**(5):527–540.
49. Lynch M: **Intron evolution as a population-genetic process.** *Proc Natl Acad Sci U S A* 2002, **99**(9):6118–6123.
50. Lynch M, Conery JS: **The origins of genome complexity.** *Science* 2003, **302**(5649):1401–1404.
51. Irimia M, Penny D, Roy SW: **Coevolution of genomic intron number and splice sites.** *Trends Genet* 2007, **23**(7):321–325.
52. Irimia M, Roy SW, Neafsey DE, Abril JF, Garcia-Fernandez J, Koonin EV: **Complex selection on 5' splice sites in intron-rich organisms.** *Genome Res* 2009, **19**(11):2021–2027.
53. Csuros M, Rogozin IB, Koonin EV: **A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes.** *PLoS Comput Biol* 2011, **7**(9):e1002150.
54. Akiyoshi DE, Morrison HG, Lei S, Feng X, Zhang Q, Corradi N, Mayanja H, Tumwine JK, Keeling PJ, Weiss LM, et al: **Genomic survey of the non-cultivable opportunistic human pathogen, *Enterocytozoon bieneusi*.** *PLoS Pathog* 2009, **5**(1):e1000261.
55. Lee RC, Gill EE, Roy SW, Fast NM: **Constrained intron structures in a microsporidian.** *Mol Biol Evol* 2010, **27**(9):1979–1982.
56. Russell AG, Shutt TE, Watkins RF, Gray MW: **An ancient spliceosomal intron in the ribosomal protein L7a gene (Rpl7a) of *Giardia lamblia*.** *BMC Evol Biol* 2005, **5**:45.
57. Parenteau J, Durand M, Morin G, Gagnon J, Lucier JF, Wellinger RJ, Chabot B, Elela SA: **Introns within ribosomal protein genes regulate the production and function of yeast ribosomes.** *Cell* 2011, **147**(2):320–331.
58. Logsdon JM Jr, Tyshenko MG, Dixon C, DJ J, Walker VK, Palmer JD: **Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory.** *Proc Natl Acad Sci U S A* 1995, **92**(18):8507–8511.
59. Rzhetsky A, Ayala FJ, Hsu LC, Chang C, Yoshida A: **Exon/intron structure of aldehyde dehydrogenase genes supports the "introns-late" theory.** *Proc Natl Acad Sci U S A* 1997, **94**(13):6820–6825.
60. de Souza SJ, Long M, Klein RJ, Roy S, Lin S, Gilbert W: **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 U S A* 1998, **95**(9):5094–5099.
61. Catania F, Lynch M: **Where do introns come from?** *PLoS Biol* 2008, **6**(11):e283.
62. Fedorov A, Merican AF, Gilbert W: **Large-scale comparison of intron positions among animal, plant, and fungal genes.** *Proc Natl Acad Sci U S A* 2002, **99**(25):16128–16133.
63. Logsdon JM Jr, Stoltzfus A, Doolittle WF: **Molecular evolution: recent cases of spliceosomal intron gain?** *Curr Biol* 1998, **8**(16):R560–R563.
64. Sakurai A, Fujimori S, Kochiwa H, Kitamura-Abe S, Washio T, Saito R, Carninci P, Hayashizaki Y, Tomita M: **On biased distribution of introns in various eukaryotes.** *Gene* 2002, **300**(1–2):89–95.
65. Mourier T, Jeffares DC: **Eukaryotic intron loss.** *Science* 2003, **300**(5624):1393.
66. Coghlan A, Wolfe KH: **Origins of recently gained introns in *Caenorhabditis*.** *Proc Natl Acad Sci U S A* 2004, **101**(31):11362–11367.
67. Stajich JE, Dietrich FS: **Evidence of mRNA-mediated intron loss in the human-pathogenic fungus *Cryptococcus neoformans*.** *Eukaryot Cell* 2006, **5**(5):789–793.
68. Coulombe-Huntington J, Majewski J: **Intron loss and gain in *Drosophila*.** *Mol Biol Evol* 2007, **24**(12):2842–2850.
69. Roy SW, Penny D: **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* 2007, **24**(1):171–181.
70. Omilian AR, Scofield DG, Lynch M: **Intron presence-absence polymorphisms in *Daphnia*.** *Mol Biol Evol* 2008, **25**(10):2129–2139.
71. Irimia M, Rukov JL, Penny D, Vinther J, Garcia-Fernandez J, Roy SW: **Origin of introns by 'intronization' of exonic sequences.** *Trends Genet* 2008, **24**(8):378–381.
72. Roy SW, Irimia M: **Mystery of intron gain: new data and new models.** *Trends Genet* 2009, **25**(2):67–73.
73. Seo HC, Kube M, Edvardsen RB, Jensen MF, Beck A, Spriet E, Gorsky G, Thompson EM, Lehrach H, Reinhardt R, et al: **Miniature genome in the marine chordate *Oikopleura dioica*.** *Science* 2001, **294**(5551):2506.
74. Nielsen CB, Friedman B, Birren B, Burge CB, Galagan JE: **Patterns of intron gain and loss in fungi.** *PLoS Biol* 2004, **2**(12):e422.
75. Edvardsen RB, Lerat E, Maeland AD, Flat M, Tewari R, Jensen MF, Lehrach H, Reinhardt R, Seo HC, Chourrout D: **Hypervariable and highly divergent intron-exon organizations in the chordate *Oikopleura dioica*.** *J Mol Evol* 2004, **59**(4):448–457.
76. Roy SW, Hartl DL: **Very little intron loss/gain in *Plasmodium*: intron loss/gain mutation rates and intron number.** *Genome Res* 2006, **16**(6):750–756.
77. Roy SW, Irimia M, Penny D: **Very little intron gain in *Entamoeba histolytica* genes laterally transferred from prokaryotes.** *Mol Biol Evol* 2006, **23**(10):1824–1827.
78. Roy SW, Penny D: **Large-scale intron conservation and order-of-magnitude variation in intron loss/gain rates in apicomplexan evolution.** *Genome Res* 2006, **16**(10):1270–1275.
79. Roy SW, Penny D: **A very high fraction of unique intron positions in the intron-rich diatom *Thalassiosira pseudonana* indicates widespread intron gain.** *Mol Biol Evol* 2007, **24**(7):1447–1457.

80. Loh YH, Brenner S, Venkatesh B: Investigation of loss and gain of introns in the compact genomes of pufferfishes (Fugu and Tetraodon). *Mol Biol Evol* 2008, **25**(3):526–535.
81. Ahmadijeh N, Dagan T, Gruenheit N, Martin W, Gabaldon T: Evolution of spliceosomal introns following endosymbiotic gene transfer. *BMC Evol Biol* 2010, **10**:57.
82. Zhang LY, Yang YF, Niu DK: Evaluation of models of the mechanisms underlying intron loss and gain in *Aspergillus* fungi. *J Mol Evol* 2010, **71**(5–6):364–373.
83. Qiu WG, Schisler N, Stoltzfus A: The evolutionary gain of spliceosomal introns: sequence and phase preferences. *Mol Biol Evol* 2004, **21**(7):1252–1263.
84. Roy SW, Gilbert W: Complex early genes. *Proc Natl Acad Sci U S A* 2005, **102**(6):1986–1991.
85. Csuros M: Likely scenarios of intron evolution. *Comparative Genomics Lecture Notes in Computer Science* 2005, **3678**:47–60.
86. Nguyen HD, Yoshihama M, Kenmochi N: New maximum likelihood estimators for eukaryotic intron evolution. *PLoS Comput Biol* 2005, **1**(7):e79.
87. Carmel L, Wolf YI, Rogozin IB, Koonin EV: Three distinct modes of intron dynamics in the evolution of eukaryotes. *Genome Res* 2007, **17**(7):1034–1044.
88. Csuros M, Holey JA, Rogozin IB: In search of lost introns. *Bioinformatics* 2007, **23**(13):i87–i96.
89. Roy SW, Fedorov A, Gilbert W: Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc Natl Acad Sci U S A* 2003, **100**(12):7158–7162.
90. Coulombe-Huntington J, Majewski J: Characterization of intron loss events in mammals. *Genome Res* 2007, **17**(1):23–32.
91. Babenko VN, Rogozin IB, Mekhedov SL, Koonin EV: Prevalence of intron gain over intron loss in the evolution of paralogous gene families. *Nucleic Acids Res* 2004, **32**(12):3724–3733.
92. Roy SW, Penny D: Smoke without fire: most reported cases of intron gain in nematodes instead reflect intron losses. *Mol Biol Evol* 2006, **23**(12):2259–2262.
93. Knowles DG, McLysaght A: High rate of recent intron gain and loss in simultaneously duplicated *Arabidopsis* genes. *Mol Biol Evol* 2006, **23**(8):1548–1557.
94. Basu MK, Rogozin IB, Deusch O, Dagan T, Martin W, Koonin EV: Evolutionary dynamics of introns in plastid-derived genes in plants: saturation nearly reached but slow intron gain continues. *Mol Biol Evol* 2008, **25**(1):111–119.
95. Gladyshev EA, Meselson M, Arhipova IR: Massive horizontal gene transfer in bdelloid rotifers. *Science* 2008, **320**(5880):1210–1213.
96. Li W, Tucker AE, Sung W, Thomas WK, Lynch M: Extensive, recent intron gains in *Daphnia* populations. *Science* 2009, **326**(5957):1260–1262.
97. Hong X, Scofield DG, Lynch M: Intron size, abundance, and distribution within untranslated regions of genes. *Mol Biol Evol* 2006, **23**(12):2392–2404.
98. Cordaux R, Udit S, Batzer MA, Feschotte C: Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc Natl Acad Sci U S A* 2006, **103**(21):8101–8106.
99. Zhang XH, Chasin LA: Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. *Proc Natl Acad Sci U S A* 2006, **103**(36):13427–13432.
100. Zhuo D, Madden R, Elela SA, Chabot B: Modern origin of numerous alternatively spliced human introns from tandem arrays. *Proc Natl Acad Sci U S A* 2007, **104**(3):882–886.
101. Szczesniak MW, Ciomborowska J, Nowak W, Rogozin IB, Makalowska I: Primate and rodent specific intron gains and the origin of retrogenes with splice variants. *Mol Biol Evol* 2011, **28**(1):33–37.
102. Ponting CP, Oliver PL, Reik W: Evolution and functions of long noncoding RNAs. *Cell* 2009, **136**(4):629–641.
103. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL: Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* 2011, **25**(18):1915–1927.
104. Chang SC, Tucker T, Thorogood NP, Brown CJ: Mechanisms of X-chromosome inactivation. *Front Biosci* 2006, **11**:852–866.
105. Elisaphenko EA, Kolesnikov NN, Shevchenko AI, Rogozin IB, Nesterova TB, Brockdorff N, Zakian SM: A dual origin of the *Xist* gene from a protein-coding gene and a set of transposable elements. *PLoS One* 2008, **3**(6):e2521.
106. Ponjavic J, Ponting CP, Lunter G: Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* 2007, **17**(5):556–565.
107. Fink GR: Pseudogenes in yeast? *Cell* 1987, **49**(1):5–6.
108. Derr LK: The involvement of cellular recombination and repair genes in RNA-mediated recombination in *Saccharomyces cerevisiae*. *Genetics* 1998, **148**(3):937–945.
109. Feiber AL, Rangarajan J, Vaughn JC: The evolution of single-copy *Drosophila* nuclear 4f-rnp genes: spliceosomal intron losses create polymorphic alleles. *J Mol Evol* 2002, **55**(4):401–413.
110. Sverdlov AV, Babenko VN, Rogozin IB, Koonin EV: Preferential loss and gain of introns in 3' portions of genes suggests a reverse-transcription mechanism of intron insertion. *Gene* 2004, **338**(1):85–91.
111. Carmel L, Rogozin IB, Wolf YI, Koonin EV: Patterns of intron gain and conservation in eukaryotic genes. *BMC Evol Biol* 2007, **7**:192.
112. Cohen NE, Shen R, Carmel L: The Role of Reverse Transcriptase in Intron Gain and Loss Mechanisms. *Mol Biol Evol* 2011, **29**:179–186.
113. Rogozin IB, Lyons-Weiler J, Koonin EV: Intron sliding in conserved gene families. *Trends Genet* 2000, **16**(10):430–432.
114. Tarrío R, Ayala FJ, Rodríguez-Trelles F: Alternative splicing: a missing piece in the puzzle of intron gain. *Proc Natl Acad Sci U S A* 2008, **105**(20):7223–7228.
115. Lehmann J, Eisenhardt C, Stadler PF, Krauss V: Some novel intron positions in conserved *Drosophila* genes are caused by intron sliding or tandem duplication. *BMC Evol Biol* 2010, **10**:156.
116. Schwartz SH, Silva J, Burstein D, Pupko T, Eyras E, Ast G: Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res* 2008, **18**(1):88–103.
117. Iwata H, Gotoh O: Comparative analysis of information contents relevant to recognition of introns in many species. *BMC Genomics* 2011, **12**:45.
118. Mount SM, Burks C, Hertz G, Stormo GD, White O, Fields C: Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucleic Acids Res* 1992, **20**(16):4255–4262.
119. Dewey CN, Rogozin IB, Koonin EV: Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. *BMC Genomics* 2006, **7**:311.
120. Sadosky T, Newman AJ, Dibb NJ: Exon junction sequences as cryptic splice sites: implications for intron origin. *Curr Biol* 2004, **14**(6):505–509.
121. Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV: Reconstruction of ancestral protosplice sites. *Curr Biol* 2004, **14**(16):1505–1508.
122. Kapustin Y, Chan E, Sarkar R, Wong F, Vorechovsky I, Winston RM, Tatusova T, Dibb NJ: Cryptic splice sites and split genes. *Nucleic Acids Res* 2011, **39**(14):5837–5844.
123. Da Lage JL, Maczkowiak F, Cariou ML: Phylogenetic distribution of intron positions in alpha-amylase genes of bacteria suggests numerous gains and losses. *PLoS One* 2011, **6**(5):e19673.
124. Babenko V, Ward W, Ruvinsky A: Does drive toward canonic exonic splicing sites exist in mammals? *J Mol Evol* 2010, **70**(4):387–394.
125. Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV: Evidence of splice signal migration from exon to intron during intron evolution. *Curr Biol* 2003, **13**(24):2170–2174.
126. Fedorov A, Suboch G, Bujakov M, Fedorova L: Analysis of nonuniformity in intron phase distribution. *Nucleic Acids Res* 1992, **20**(10):2553–2557.
127. Long M, Deutsch M: Association of intron phases with conservation at splice site sequences and evolution of spliceosomal introns. *Mol Biol Evol* 1999, **16**(11):1528–1534.
128. Long M, Rosenberg C: Testing the "proto-splice sites" model of intron origin: evidence from analysis of intron phase correlations. *Mol Biol Evol* 2000, **17**(12):1789–1796.
129. Ruvinsky A, Eskesen ST, Eskesen FN, Hurst LD: Can codon usage bias explain intron phase distributions and exon symmetry? *J Mol Evol* 2005, **60**(1):99–104.
130. Nguyen HD, Yoshihama M, Kenmochi N: Phase distribution of spliceosomal introns: implications for intron origin. *BMC Evol Biol* 2006, **6**:69.
131. DENOUD F, HENRIET S, MUNGPAKDEE S, AURY JM, DA SILVA C, BRINKMANN H, MIKHALEVA J, OLSEN LC, JUBIN C, CANESTRO C, et al: Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science* 2010, **330**(6009):1381–1385.
132. Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV: Conservation versus parallel gains in intron evolution. *Nucleic Acids Res* 2005, **33**(6):1741–1748.

133. Hankeln T, Friedl H, Ebersberger I, Martin J, Schmidt ER: **A variable intron distribution in globin genes of Chironomus: evidence for recent intron gain.** *Gene* 1997, **205**(1-2):151-160.
134. Tarrío R, Rodríguez-Trelles F, Ayala FJ: **A new Drosophila spliceosomal intron position is common in plants.** *Proc Natl Acad Sci U S A* 2003, **100**(11):6580-6583.
135. Rogozin IB, Sverdlov AV, Babenko VN, Koonin EV: **Analysis of evolution of exon-intron structure of eukaryotic genes.** *Brief Bioinform* 2005, **6**(2):118-134.
136. Carmel L, Rogozin IB, Wolf YI, Koonin EV: **An expectation-maximization algorithm for analysis of evolution of exon-intron structure of eukaryotic genes.** *Comparative Genomics Lecture Notes in Computer Science* 2005, **3678**:35-46.
137. Roy SW: **Intron-rich ancestors.** *Trends Genet* 2006, **22**(9):468-471.
138. Csuros M: **Malin: maximum likelihood analysis of intron evolution in eukaryotes.** *Bioinformatics* 2008, **24**(13):1538-1539.
139. Sullivan JC, Reitzel AM, Finnerty JR: **A high percentage of introns in human genes were present early in animal evolution: evidence from the basal metazoan Nematostella vectensis.** *Genome Inform* 2006, **17**(1):219-229.
140. Raible F, Tessmar-Raible K, Osoegawa K, Wincker P, Jubin C, Balavoine G, Ferrier D, Benes V, de Jong P, Weissenbach J, et al: **Vertebrate-type intron-rich genes in the marine annelid Platynereis dumerilii.** *Science* 2005, **310**(5752):1325-1326.
141. Csuros M, Rogozin IB, Koonin EV: **Extremely intron-rich genes in the alveolate ancestors inferred with a flexible maximum-likelihood approach.** *Mol Biol Evol* 2008, **25**(5):903-911.
142. Castillo-Davis CI, Bedford TB, Hartl DL: **Accelerated rates of intron gain/loss and protein evolution in duplicate genes in human and mouse malaria parasites.** *Mol Biol Evol* 2004, **21**(7):1422-1427.
143. Nguyen HD, Yoshihama M, Kenmochi N: **The evolution of spliceosomal introns in alveolates.** *Mol Biol Evol* 2007, **24**(5):1093-1096.
144. Roy SW, Penny D: **On the incidence of intron loss and gain in paralogous gene families.** *Mol Biol Evol* 2007, **24**(8):1579-1581.
145. Flakowski J, Bolivar I, Fahrni J, Pawlowski J: **Tempo and mode of spliceosomal intron evolution in actin of foraminifera.** *J Mol Evol* 2006, **63**(1):30-41.
146. Kumar A, Ragg H: **Ancestry and evolution of a secretory pathway serpin.** *BMC Evol Biol* 2008, **8**:250.
147. García-España A, Mares R, Sun TT, Desalle R: **Intron evolution: testing hypotheses of intron evolution using the phylogenomics of tetraspanins.** *PLoS One* 2009, **4**(3):e4680.
148. Sverdlov AV, Csuros M, Rogozin IB, Koonin EV: **A glimpse of a putative pre-intron phase of eukaryotic evolution.** *Trends Genet* 2007, **23**(3):105-108.
149. Makarova KS, Wolf YI, Mekhedov SL, Mirkin BG, Koonin EV: **Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell.** *Nucleic Acids Res* 2005, **33**(14):4626-4638.
150. Yoshihama M, Nguyen HD, Kenmochi N: **Intron dynamics in ribosomal protein genes.** *PLoS One* 2007, **2**(1):e141.
151. Cho G, Doolittle RF: **Intron distribution in ancient paralogs supports random insertion and not random loss [published erratum appears in J Mol Evol 1997 Aug;45(2):206].** *J Mol Evol* 1997, **44**(6):573-584.
152. Pal C, Papp B, Hurst LD: **Highly expressed genes in yeast evolve slowly.** *Genetics* 2001, **158**(2):927-931.
153. Krylov DM, Wolf YI, Rogozin IB, Koonin EV: **Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution.** *Genome Res* 2003, **13**(10):2229-2235.
154. Drummond DA, Wilke CO: **The evolutionary consequences of erroneous protein synthesis.** *Nat Rev Genet* 2009, **10**(10):715-724.
155. Carmel L, Rogozin IB, Wolf YI, Koonin EV: **Evolutionarily conserved genes preferentially accumulate introns.** *Genome Res* 2007, **17**(7):1045-1050.
156. Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA: **Selection for short introns in highly expressed genes.** *Nat Genet* 2002, **31**(4):415-418.
157. Eisenberg E, Levanon EY: **Human housekeeping genes are compact.** *Trends Genet* 2003, **19**(7):362-365.
158. Urrutia AO, Hurst LD: **The signature of selection mediated by expression on human genes.** *Genome Res* 2003, **13**(10):2260-2264.
159. Comerón JM: **Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence.** *Genetics* 2004, **167**(3):1293-1304.
160. Vinogradov AE: **Compactness of human housekeeping genes: selection for economy or genomic design?** *Trends Genet* 2004, **20**(5):248-253.
161. Chen J, Sun M, Hurst LD, Carmichael GG, Rowley JD: **Human antisense genes have unusually short introns: evidence for selection for rapid transcription.** *Trends Genet* 2005, **21**(4):203-207.
162. Li SW, Feng L, Niu DK: **Selection for the miniaturization of highly expressed genes.** *Biochem Biophys Res Commun* 2007, **360**(3):586-592.
163. Fahey ME, Higgins DG: **Gene expression, intron density, and splice site strength in Drosophila and Caenorhabditis.** *J Mol Evol* 2007, **65**(3):349-357.
164. Seoighe C, Gehring C, Hurst LD: **Gametophytic selection in Arabidopsis thaliana supports the selective model of intron length reduction.** *PLoS Genet* 2005, **1**(2):e13.
165. Stenoién HK: **Compact genes are highly expressed in the moss Physcomitrella patens.** *J Evol Biol* 2007, **20**(3):1223-1229.
166. Ren XY, Vorst O, Fiers MW, Stiekema WJ, Nap JP: **In plants, highly expressed genes are the least compact.** *Trends Genet* 2006, **22**(10):528-532.
167. Carmel L, Koonin EV: **A universal nonmonotonic relationship between gene compactness and expression levels in multicellular eukaryotes.** *Genome Biol Evol* 2009, **1**:282-290.
168. Lynch M: **The origins of eukaryotic gene structure.** *Mol Biol Evol* 2006, **23**(2):450-468.
169. Lynch M: **The frailty of adaptive hypotheses for the origins of organismal complexity.** *Proc Natl Acad Sci U S A* 2007, **104**(Suppl 1):8597-8604.
170. Charlesworth B: **Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation.** *Nat Rev Genet* 2009, **10**(3):195-205.
171. Strasburg JL, Kane NC, Raduski AR, Bonin A, Michelmore R, Rieseberg LH: **Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers.** *Mol Biol Evol* 2011, **28**(5):1569-1580.
172. Koonin EV: *The Logic of Chance: The Nature and Origin of Biological Evolution.* FT Press; 2011.
173. Black DL: **Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology.** *Cell* 2000, **103**(3):367-370.
174. Kriventseva EV, Koch I, Apweiler R, Vingron M, Bork P, Gelfand MS, Sunyaev S: **Increase of functional diversity by alternative splicing.** *Trends Genet* 2003, **19**(3):124-128.
175. Resch A, Xing Y, Modrek B, Gorlick M, Riley R, Lee C: **Assessing the impact of alternative splicing on domain interactions in the human proteome.** *J Proteome Res* 2004, **3**(1):76-83.
176. Keren H, Lev-Maor G, Ast G: **Alternative splicing and evolution: diversification, exon definition and function.** *Nat Rev Genet* 2010, **11**(5):345-355.
177. Brett D, Hanke J, Lehmann G, Haase S, Delbruck S, Krueger S, Reich J, Bork P: **EST comparison indicates 38% of human mRNAs contain possible alternative splice forms.** *FEBS Lett* 2000, **474**(1):83-86.
178. Croft L, Schandorff S, Clark F, Burrage K, Arctander P, Mattick JS: **ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome.** *Nat Genet* 2000, **24**(4):340-341.
179. Modrek B, Resch A, Grasso C, Lee C: **Genome-wide detection of alternative splicing in expressed sequences of human genes.** *Nucleic Acids Res* 2001, **29**(13):2850-2859.
180. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.** *Nat Genet* 2008, **40**(13):1413-1415.
181. Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong WK, Mockler TC: **Genome-wide mapping of alternative splicing in Arabidopsis thaliana.** *Genome Res* 2010, **20**(1):45-58.
182. Lu T, Lu G, Fan D, Zhu C, Li W, Zhao Q, Feng Q, Zhao Y, Guo Y, Huang X, et al: **Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq.** *Genome Res* 2010, **20**(9):1238-1249.
183. Severing EI, van Dijk AD, van Ham RC: **Assessing the contribution of alternative splicing to proteome diversity in Arabidopsis thaliana using proteomics data.** *BMC Plant Biol* 2011, **11**(1):82.
184. Irimia M, Rukov JL, Penny D, Roy SW: **Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing.** *BMC Evol Biol* 2007, **7**:188.
185. Liu M, Walch H, Wu S, Grigoriev A: **Significant expansion of exon-bordering protein domains during animal proteome evolution.** *Nucleic Acids Res* 2005, **33**(1):95-105.

186. Patthy L: **Intron-dependent evolution: preferred types of exons and introns.** *FEBS Lett* 1987, **214**(1):1–7.
187. Patthy L: **Genome evolution and the evolution of exon-shuffling—a review.** *Gene* 1999, **238**(1):103–114.
188. Patthy L: **Modular assembly of genes and the evolution of new functions.** *Genetica* 2003, **118**(2–3):217–231.
189. Modrek B, Lee CJ: **Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss.** *Nat Genet* 2003, **34**(2):177–180.
190. Malko DB, Makeev VJ, Mironov AA, Gelfand MS: **Evolution of exon-intron structure and alternative splicing in fruit flies and malarial mosquito genomes.** *Genome Res* 2006, **16**(4):505–509.
191. Nurtudinov RN, Neverov AD, Favorov AV, Mironov AA, Gelfand MS: **Conserved and species-specific alternative splicing in mammalian genomes.** *BMC Evol Biol* 2007, **7**:249.
192. Takeda J, Suzuki Y, Sakate R, Sato Y, Seki M, Irie T, Takeuchi N, Ueda T, Nakao M, Sugano S, et al: **Low conservation and species-specific evolution of alternative splicing in humans and mice: comparative genomics analysis using well-annotated full-length cDNAs.** *Nucleic Acids Res* 2008, **36**(20):6386–6395.
193. Mudge JM, Frankish A, Fernandez-Banet J, Alioto T, Derrien T, Howald C, Reymond A, Guigo R, Hubbard T, Harrow J: **The origins, evolution, and functional potential of alternative splicing in vertebrates.** *Mol Biol Evol* 2011, **28**(10):2949–2959.
194. Irimia M, Rukov JL, Penny D, Garcia-Fernandez J, Vinther J, Roy SW: **Widespread evolutionary conservation of alternatively spliced exons in *Caenorhabditis*.** *Mol Biol Evol* 2008, **25**(2):375–382.
195. Churbanov A, Winters-Hilt S, Koonin EV, Rogozin IB: **Accumulation of GC donor splice signals in mammals.** *Biol Direct* 2008, **3**:30.
196. Thanaraj TA, Clark F: **Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions.** *Nucleic Acids Res* 2001, **29**(12):2581–2593.
197. Farrer T, Roller AB, Kent WJ, Zahler AM: **Analysis of the role of *Caenorhabditis elegans* GC-AG introns in regulated splicing.** *Nucleic Acids Res* 2002, **30**(15):3360–3367.
198. Campbell MA, Haas BJ, Hamilton JP, Mount SM, Buell CR: **Comprehensive analysis of alternative splicing in rice and comparative analyses with *Arabidopsis*.** *BMC Genomics* 2006, **7**:327.
199. Stallings-Mann ML, Ludwiczak RL, Klinger KW, Rottman F: **Alternative splicing of exon 3 of the human growth hormone receptor is the result of an unusual genetic polymorphism.** *Proc Natl Acad Sci U S A* 1996, **93**(22):12394–12399.
200. Fairbrother WG, Holste D, Burge CB, Sharp PA: **Single nucleotide polymorphism-based validation of exonic splicing enhancers.** *PLoS Biol* 2004, **2**(9):E268.
201. Stanton T, Boxall S, Hirai K, Dawes R, Tonks S, Yasui T, Kanaoka Y, Yuldasheva N, Ishiko O, Bodmer W, et al: **A high-frequency polymorphism in exon 6 of the CD45 tyrosine phosphatase gene (PTPRC) resulting in altered isoform expression.** *Proc Natl Acad Sci U S A* 2003, **100**(10):5997–6002.
202. Parnley JL, Chamary JV, Hurst LD: **Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers.** *Mol Biol Evol* 2006, **23**(2):301–309.
203. Carlini DB, Genut JE: **Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers.** *J Mol Evol* 2006, **62**(1):89–98.
204. Ke S, Zhang XH, Chasin LA: **Positive selection acting on splicing motifs reflects compensatory evolution.** *Genome Res* 2008, **18**(4):533–543.
205. Irimia M, Rukov JL, Roy SW: **Evolution of alternative splicing regulation: changes in predicted exonic splicing regulators are not associated with changes in alternative splicing levels in primates.** *PLoS One* 2009, **4**(6):e5800.
206. Coulombe-Huntington J, Lam KC, Dias C, Majewski J: **Fine-scale variation and genetic determinants of alternative splicing across individuals.** *PLoS Genet* 2009, **5**(12):e1000766.
207. Ji H, Zhang Y, Zheng W, Wu Z, Lee S, Sandberg K: **Translational regulation of angiotensin type 1a receptor expression and signaling by upstream AUGs in the 5' leader sequence.** *J Biol Chem* 2004, **279**(44):45322–45328.
208. Irvin-Wilson CV, Chaudhuri G: **Alternative initiation and splicing in dicer gene expression in human breast cells.** *Breast Cancer Res* 2005, **7**(4):R563–R569.
209. Resch AM, Ogurtsov AY, Rogozin IB, Shabalina SA, Koonin EV: **Evolution of alternative and constitutive regions of mammalian 5'UTRs.** *BMC Genomics* 2009, **10**:162.
210. Zhang T, Haws P, Wu Q: **Multiple variable first exons: a mechanism for cell- and tissue-specific gene regulation.** *Genome Res* 2004, **14**(1):79–89.
211. Nagasaki H, Arita M, Nishizawa T, Suwa M, Gotoh O: **Automated classification of alternative splicing and transcriptional initiation and construction of visual database of classified patterns.** *Bioinformatics* 2006, **22**(10):1211–1216.
212. Hoffman MM, Birney E: **Estimating the neutral rate of nucleotide substitution using introns.** *Mol Biol Evol* 2007, **24**(2):522–531.
213. Resch AM, Carmel L, Marino-Ramirez L, Ogurtsov AY, Shabalina SA, Rogozin IB, Koonin EV: **Widespread positive selection in synonymous sites of mammalian genes.** *Mol Biol Evol* 2007, **24**(8):1821–1831.
214. Fedorova L, Fedorov A: **Introns in gene evolution.** *Genetica* 2003, **118**(2–3):123–131.
215. Rodriguez-Trelles F, Tarrío R, Ayala FJ: **Origins and evolution of spliceosomal introns.** *Annu Rev Genet* 2006, **40**:47–76.
216. Luo MJ, Reed R: **Splicing is required for rapid and efficient mRNA export in metazoans.** *Proc Natl Acad Sci U S A* 1999, **96**(26):14937–14942.
217. Le Hir H, Moore MJ, Maquat LE: **Pre-mRNA splicing alters mRNP composition: evidence for stable association of proteins at exon-exon junctions.** *Genes Dev* 2000, **14**(9):1098–1108.
218. Zhou Z, Luo MJ, Straesser K, Katahira J, Hurt E, Reed R: **The protein Aly links pre-messenger-RNA splicing to nuclear export in metazoans.** *Nature* 2000, **407**(6802):401–405.
219. Le Hir H, Izaurralde E, Maquat LE, Moore MJ: **The spliceosome deposits multiple proteins 20–24 nucleotides upstream of mRNA exon-exon junctions.** *EMBO J* 2000, **19**(24):6860–6869.
220. Luo ML, Zhou Z, Magni K, Christoforides C, Rappsilber J, Mann M, Reed R: **Pre-mRNA splicing and mRNA export linked by direct interactions between UAP56 and Aly.** *Nature* 2001, **413**(6856):644–647.
221. Strasser K, Hurt E: **Splicing factor Sub2p is required for nuclear mRNA export through its interaction with Yra1p.** *Nature* 2001, **413**(6856):648–652.
222. Gatfield D, Le Hir H, Schmitt C, Braun IC, Kocher T, Wilm M, Izaurralde E: **The DEXH/D box protein HEL/UAP56 is essential for mRNA nuclear export in *Drosophila*.** *Curr Biol* 2001, **11**(21):1716–1721.
223. Aravind L, Watanabe H, Lipman DJ, Koonin EV: **Lineage-specific loss and divergence of functionally linked genes in eukaryotes.** *Proc Natl Acad Sci U S A* 2000, **97**(21):11319–11324.
224. Jensen TH, Boulay J, Rosbash M, Libri D: **The DECD box putative ATPase Sub2p is an early mRNA export factor.** *Curr Biol* 2001, **11**(21):1711–1715.
225. Huang Y, Gattoni R, Stevenin J, Steitz JA: **SR splicing factors serve as adapter proteins for TAP-dependent mRNA export.** *Mol Cell* 2003, **11**(3):837–843.
226. Huang Y, Steitz JA: **SRprises along a messenger's journey.** *Mol Cell* 2005, **17**(5):613–615.
227. Hilleren P, McCarthy T, Rosbash M, Parker R, Jensen TH: **Quality control of mRNA 3'-end processing is linked to the nuclear exosome.** *Nature* 2001, **413**(6855):538–542.
228. Buckley PT, Lee MT, Sul JY, Miyashiro KY, Bell TJ, Fisher SA, Kim J, Eberwine J: **Cytoplasmic intron sequence-retaining transcripts can be dendritically targeted via ID element retrotransposons.** *Neuron* 2011, **69**(5):877–884.
229. Alioto TS: **U12DB: a database of orthologous U12-type spliceosomal introns.** *Nucleic Acids Res* 2007, **35**(Database issue):D110–D115.
230. Pleiss JA, Whitworth GB, Bergkessel M, Guthrie C: **Transcript specificity in yeast pre-mRNA splicing revealed by mutations in core spliceosomal components.** *PLoS Biol* 2007, **5**(4):e90.
231. Doolittle RF: **The multiplicity of domains in proteins.** *Annu Rev Biochem* 1995, **64**:287–314.
232. Koonin EV, Aravind L, Kondrashov AS: **The impact of comparative genomics on our understanding of evolution.** *Cell* 2000, **101**:573–576.
233. Tordai H, Nagy A, Farkas K, Banyai L, Patthy L: **Modules, multidomain proteins and organismic complexity.** *FEBS J* 2005, **272**(19):5064–5078.
234. Basu MK, Poliakov E, Rogozin IB: **Domain mobility in proteins: functional and evolutionary implications.** *Brief Bioinform* 2009, **10**(3):205–216.
235. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences.** *Science* 1999, **285**(5428):751–753.
236. de Souza SJ: **The emergence of a synthetic theory of intron evolution.** *Genetica* 2003, **118**(2–3):117–121.
237. Carvalho AB, Clark AG: **Intron size and natural selection.** *Nature* 1999, **401**(6751):344.

238. Nilsen TW: **Evolutionary origin of SL-addition trans-splicing: still an enigma.** *Trends Genet* 2001, **17**(12):678–680.
239. Evans D, Blumenthal T: **trans splicing of polycistronic *Caenorhabditis elegans* pre-mRNAs: analysis of the SL2 RNA.** *Mol Cell Biol* 2000, **20**(18):6659–6667.
240. Blumenthal T, Evans D, Link CD, Guffanti A, Lawson D, Thierry-Mieg J, Thierry-Mieg D, Chiu WL, Duke K, Kiraly M, et al: **A global analysis of *Caenorhabditis elegans* operons.** *Nature* 2002, **417**(6891):851–854.
241. Allen MA, Hillier LW, Waterston RH, Blumenthal T: **A global analysis of *C. elegans* trans-splicing.** *Genome Res* 2011, **21**(2):255–264.
242. Lasda EL, Blumenthal T: **Trans-splicing.** *Wiley Interdiscip Rev RNA* 2011, **2**(3):417–434.
243. Douris V, Telford MJ, Averof M: **Evidence for multiple independent origins of trans-splicing in Metazoa.** *Mol Biol Evol* 2010, **27**(3):684–693.
244. Herai RH, Yamagishi ME: **Detection of human interchromosomal trans-splicing in sequence databanks.** *Brief Bioinform* 2010, **11**(2):198–209.
245. McManus CJ, Duff MO, Eipper-Mains J, Graveley BR: **Global analysis of trans-splicing in *Drosophila*.** *Proc Natl Acad Sci U S A* 2010, **107**(29):12975–12979.
246. Glazko GV, Koonin EV, Rogozin IB, Shabalina SA: **A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions.** *Trends Genet* 2003, **19**(3):119–124.
247. Bergman CM, Kreitman M: **Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences.** *Genome Res* 2001, **11**(8):1335–1345.
248. Majewski J, Ott J: **Distribution and characterization of regulatory elements in the human genome.** *Genome Res* 2002, **12**(12):1827–1836.
249. Parsch J: **Selective constraints on intron evolution in *Drosophila*.** *Genetics* 2003, **165**(4):1843–1851.
250. Marais G, Nouvellet P, Keightley PD, Charlesworth B: **Intron size and exon evolution in *Drosophila*.** *Genetics* 2005, **170**(1):481–485.
251. Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P: **Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content.** *Genome Biol* 2005, **6**(8):R67.
252. Halligan DL, Keightley PD: **Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison.** *Genome Res* 2006, **16**(7):875–884.
253. Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P: **On the utility of short intron sequences as a reference for the detection of positive and negative selection in *Drosophila*.** *Mol Biol Evol* 2010, **27**(6):1226–1234.
254. Muller F, Chang B, Albert S, Fischer N, Tora L, Strahle U: **Intronic enhancers control expression of zebrafish sonic hedgehog in floor plate and notochord.** *Development* 1999, **126**(10):2103–2116.
255. Sandelin A, Bailey P, Bruce S, Engstrom PG, Klos JM, Wasserman WW, Ericson J, Lenhard B: **Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes.** *BMC Genomics* 2004, **5**(1):99.
256. Abbasi AA, Papatidis Z, Malik S, Goode DK, Callaway H, Elgar G, Grzeschik KH: **Human GLI3 intragenic conserved non-coding sequences are tissue-specific enhancers.** *PLoS One* 2007, **2**(4):e366.
257. Irimia M, Maeso I, Burguera D, Hidalgo-Sanchez M, Puelles L, Roy SW, Garcia-Fernandez J, Ferran JL: **Contrasting 5' and 3' evolutionary histories and frequent evolutionary convergence in Meis/hth gene structures.** *Genome Biol Evol* 2011, **3**:551–564.
258. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, et al: **Highly conserved non-coding sequences are associated with vertebrate development.** *PLoS Biol* 2005, **3**(1):e7.
259. McEwen GK, Woolfe A, Goode D, Vavouri T, Callaway H, Elgar G: **Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis.** *Genome Res* 2006, **16**(4):451–465.
260. Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engstrom PG, Fredman D, Akalin A, Caccamo N, Sealy I, Howe K, et al: **Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates.** *Genome Res* 2007, **17**(5):545–555.
261. Engstrom PG, Ho Sui SJ, Drivenes O, Becker TS, Lenhard B: **Genomic regulatory blocks underlie extensive microsynteny conservation in insects.** *Genome Res* 2007, **17**(12):1898–1908.
262. Tycowski KT, Shu MD, Steitz JA: **A small nucleolar RNA is processed from an intron of the human gene encoding ribosomal protein S3.** *Genes Dev* 1993, **7**(7A):1176–1190.
263. Tycowski KT, Shu MD, Steitz JA: **A mammalian gene with introns instead of exons generating stable RNA products.** *Nature* 1996, **379**(6564):464–466.
264. Shabalina SA, Koonin EV: **Origins and evolution of eukaryotic RNA interference.** *Trends Ecol Evol* 2008, **23**(10):578–587.
265. Rearick D, Prakash A, McSweeney A, Shepard SS, Fedorova L, Fedorov A: **Critical association of ncRNA with introns.** *Nucleic Acids Res* 2011, **39**(6):2357–2366.
266. Berezikov E, Chung WJ, Willis J, Cuppen E, Lai EC: **Mammalian mirtron genes.** *Mol Cell* 2007, **28**(2):328–336.
267. Golan D, Levy C, Friedman B, Shomron N: **Biased hosting of intronic microRNA genes.** *Bioinformatics* 2010, **26**(8):992–995.
268. Westholm JO, Lai EC: **Mirtrons: microRNA biogenesis via splicing.** *Biochimie* 2011, **93**(11):1897–1904.
269. Kim VN, Han J, Siomi MC: **Biogenesis of small RNAs in animals.** *Nat Rev Mol Cell Biol* 2009, **10**(2):126–139.
270. Kumar A: **An overview of nested genes in eukaryotic genomes.** *Eukaryot Cell* 2009, **8**(9):1321–1329.
271. Assis R, Kondrashov AS, Koonin EV, Kondrashov FA: **Nested genes and increasing organizational complexity of metazoan genomes.** *Trends Genet* 2008, **24**(10):475–478.
272. Llopert A, Cameron JM, Brunet FG, Lachaise D, Long M: **Intron presence-absence polymorphism in *Drosophila* driven by positive Darwinian selection.** *Proc Natl Acad Sci U S A* 2002, **99**(12):8121–8126.
273. Torriani SF, Stukenbrock EH, Brunner PC, McDonald BA, Croll D: **Evidence for extensive recent intron transposition in closely related fungi.** *Curr Biol* 2011, in press.
274. Gao X, Lynch M: **Ubiquitous internal gene duplication and intron creation in eukaryotes.** *Proc Natl Acad Sci U S A* 2009, **106**(49):20818–20823.
275. Farlow A, Meduri E, Dolezal M, Hua L, Schlotterer C: **Nonsense-mediated decay enables intron gain in *Drosophila*.** *PLoS Genet* 2010, **6**(1):e1000819.
276. Curtis BA, Archibald JM: **A spliceosomal intron of mitochondrial DNA origin.** *Curr Biol* 2010, **20**(21):R919–R920.
277. Worden AZ, Lee JH, Mock T, Rouze P, Simmons MP, Aerts AL, Allen AE, Cuvelier ML, Derelle E, Everett MV, et al: **Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*.** *Science* 2009, **324**(5924):268–272.
278. Archibald JM, O'Kelly CJ, Doolittle WF: **The chaperonin genes of jakobid and jakobid-like flagellates: implications for eukaryotic evolution.** *Mol Biol Evol* 2002, **19**(4):422–431.
279. Vanacova S, Yan W, Carlton JM, Johnson PJ: **Spliceosomal introns in the deep-branching eukaryote *Trichomonas vaginalis*.** *Proc Natl Acad Sci U S A* 2005, **102**(12):4430–4435.
280. Martin W, Koonin EV: **Introns and the origin of nucleus-cytosol compartmentalization.** *Nature* 2006, **440**(7080):41–45.
281. Lopez-Garcia P, Moreira D: **Selective forces for the origin of the eukaryotic nucleus.** *Bioessays* 2006, **28**(5):525–533.
282. Lambowitz AM, Zimmerly S: **Group II introns: mobile ribozymes that invade DNA.** *Cold Spring Harb Perspect Biol* 2010, **3**(8):a003616.
283. Keating KS, Toor N, Perlman PS, Pyle AM: **A structural analysis of the group II intron active site and implications for the spliceosome.** *RNA* 2010, **16**(1):1–9.
284. Simon DM, Kelchner SA, Zimmerly S: **A broadscale phylogenetic analysis of group II intron RNAs and intron-encoded reverse transcriptases.** *Mol Biol Evol* 2009, **26**(12):2795–2808.
285. Tautz D, Lässig M: **Of statistics and genomes.** *Trends Genet* 2004, **20**(8):344–346.
286. Koonin EV: **Intron-dominated genomes of early ancestors of eukaryotes.** *J Hered* 2009, **100**(5):618–623.
287. Koonin EV, Senkevich TG, Dolja W: **The ancient Virus World and evolution of cells.** *Biol Direct* 2006, **1**:29.

doi:10.1186/1745-6150-7-11

Cite this article as: Rogozin et al.: Origin and evolution of spliceosomal introns. *Biology Direct* 2012 **7**:11.