Endogenous Mechanisms for the Origins of Spliceosomal Introns

FRANCESCO CATANIA, XIANG GAO, AND DOUGLAS G. SCOFIELD

From the Department of Biology, Indiana University, 1001 East Third Street, Bloomington, IN 47405 (Catania and Gao); and the Department of Ecology and Evolutionary Biology, University of California, 621 Charles E. Young Drive South, Box 951606, Los Angeles, CA 90095-1606 (Scofield).

Address correspondence to Francesco Catania at the address above, or e-mail: fcatania@indiana.edu.

Abstract

Over 30 years since their discovery, the origin of spliceosomal introns remains uncertain. One nearly universally accepted hypothesis maintains that spliceosomal introns originated from self-splicing group-II introns that invaded the uninterrupted genes of the last eukaryotic common ancestor (LECA) and proliferated by “insertion” events. Although this is a possible explanation for the original presence of introns and splicing machinery, the emphasis on a high number of insertion events in the genome of the LECA neglects a considerable body of empirical evidence showing that spliceosomal introns can simply arise from coding or, more generally, nonintronic sequences within genes. After presenting a concise overview of some of the most common hypotheses and mechanisms for intron origin, we propose two further hypotheses that are broadly based on central cellular processes: 1) internal gene duplication and 2) the response to aberrant and fortuitously spliced transcripts. These two nonmutually exclusive hypotheses provide a powerful way to explain the establishment of spliceosomal introns in eukaryotes without invoking an exogenous source.

Key words: group-II introns, internal gene duplication, intronization, spliceosomal introns

Ever since the discovery of spliceosomal introns (Berget et al. 1977; Chow et al. 1977; Evans et al. 1977; Goldberg et al. 1977), an enormous number of multidisciplinary studies have shed light on the evolutionary features of these noncoding sequences and revealed peculiar links between splicing and a number of other cellular processes (Proudfoot et al. 2002; Reed and Hurt 2002). Yet, how (and why) spliceosomal introns first appeared remains an unsolved issue.

A Commonly Accepted View: Spliceosomal Introns Originate from Group-II Introns

Despite the several plausible attempts to explain the origin of spliceosomal introns (Belshaw and Bensasson 2006; Rodriguez-Trelles et al. 2006; Roy and Gilbert 2006), one single hypothesis appears nearly universally accepted, although essentially not verifiable. This hypothesis maintains that spliceosomal introns, together with the spliceosome(s)—a large ribonucleoprotein complex that is actively involved in splicing (Burge et al. 1999; Brow 2002; Kuhn and Kauffer 2003; Nilsen 2003)—originated from a class of self-splicing introns (group-II) (Sharp 1985; Cech 1986; Jacquier 1990; Cavalier-Smith 1991; Palmer and Logsdon 1991). Briefly, group-II introns of α-proteobacterial endosymbionts (presumably primordial mitochondria) of the last eukaryotic common ancestor (LECA) are thought to have first invaded the previously uninterrupted nuclear genes and subsequently lost the ability of self-splicing, eventually coming to be spliced in “trans” by the newly invented spliceosomal machinery (Stoltzfus 1999). In line with this presumptive mode of origin, group-II introns and spliceosomal introns share several structural and functional similarities (Shukla and Padgett 2002; Valadkhan and Manley 2002), and plausible steps have been suggested to have allowed the conversion between the two classes of introns (Lynch and Richardson 2002). Although group-II introns provide a reasonable source for the original presence of spliceosomal introns and of the splicing machinery in the LECA, some aspects of the group-II introns origin hypothesis remain unclear, especially the conjectured widespread invasion of uninterrupted LECA genes. In particular, although group-II introns have been reported in the organelles of plants and lower eukaryotes (Bonen and Vogel 2001) as well as in eubacteria and archea (Dai and Zimmerly 2002, 2003; Rest and Mindell 2003; Toro 2003), neither functional copies nor relics of such sequences have ever been found in animal mitochondria or in eukaryotic
nuclei. A major loss of group-II introns subsequent to spliceosomal intron colonization cannot be entirely ruled out, but the current absence of group-II introns from eukaryotic nuclear genomes conflicts with the described evidence for putative recent intron gains (Coulombe-Huntington and Majewski 2007; Roy and Penny 2007; Omilian et al. 2008).

Origin of Spliceosomal Introns from Alternative Sources?

As opposed to widespread invasion of LECA genes by group-II introns, other hypotheses have been proposed that do not invoke an exogenous invading source. Transposable elements have been suggested as one possible endogenous source. After inserting into the coding sequence, a transposon would be removed from the pre-mRNA either by means of latent splice sites that may exist in the inserted element or after the acquisition of splicing signals. The very few spliceosomal introns that are known to derive from mobile elements (Giroux et al. 1994; Iwamoto et al. 1999) give limited support for the generality of this hypothesis. Moreover, excision of transposable elements that are inserted into genes is only rarely perfect (e.g., Giroux et al. 1994; Greco et al. 2005) and in most cases generates indels (e.g., Wessler 1989; Fridell et al. 1990; Giroux et al. 1994; Rushforth and Anderson 1996) so that the mutant gene either partially or completely retains the wild-type gene’s phenotype only if the open reading frame (ORF) of the sequence downstream of the insertion site is not disrupted.

Spliceosomal introns have also been suggested to directly arise from random primordial and canonical ancestral gene sequences (the “split-gene model” and the “proto-splice site model,” respectively). In particular, Senapathy (1986) observed that the distribution of reading-frame lengths in a random nucleotide sequence is exponential (but see Hoglund et al. 1990), with a 600-nt upper limit that is in line with a number of observed average exon sizes (Hawkins 1988; Yandell et al. 2006). Under the assumption that in the most primitive unicellular eukaryotes a selective pressure existed to generate longer coding sequences, the split-gene model (Senapathy 1986, 1988) maintains that random sequences that were populated with in-frame nonsense codons, and intervened between short reading frames, started to be excised by an already existing spliceosome. The model proposes not only that the sequences excised contained random clusters of in-frame nonsense codons but also that the splice junction signal sequences and the branchpoint sequence originate from nonsense codons (Senapathy 1988). One implication of this model is the existence of a nuclear scanning mechanism to recognize in-frame termination codons. Although some of the split-gene model’s propositions have been challenged (Stoltzfus et al. 1995), the hypothesis that nonsense codons may have played a role in the origin of introns remains conceivable (see below and Senapathy 1988; Harris and Senapathy 1990).

In the proto-splice site model (Dibb 1991), introns arise at proto-splice sites, which contain the partially conserved exon consensus sequence \([C|A]AG^1R\) adjacent to introns, where “\(^1\)” is the point of intron insertion (Dibb and Newman 1989). In this model, spliceosomal introns are originally coding and began as a result of the evolution of alternatively spliced isoforms of a transcript. In particular, initially inefficient splicing occurring at proto-splice sites would lead to the emergence of alternative transcript isoforms with subsequent mutations that stabilize beneficial isoforms. Splicing would become constitutive as a result of this selective filter and the new intronic sequences would drift so that any reinsertion in the mature transcript would be deleterious. To deal with the possible criticism that ancient proteins must have been very large if all introns originated from coding sequences, the model also maintains that only a limited number of introns would originate from the hypothetical process described above, and intron proliferation would largely proceed by insertion events driven by the spliceosomal machinery through the mechanism of reverse splicing (Tseng and Cheng 2008). Although a number of this model’s observations appear to be consistent with the insertion of introns at proto-splice sites (Dibb and Newman 1989; Lee et al. 1991; Frugoli et al. 1998; Logsdon et al. 1998; Funke et al. 1999; Bhattacharya et al. 2000; Kent and Zahler 2000; Sverdlov et al. 2003; Coghlan and Wolfe 2004; Qiu et al. 2004; Sadusky et al. 2004; Sverdlov et al. 2004; Yoshihama et al. 2006; Omilian et al. 2008), no convincing sequence similarity has been ever observed between (putative) recent intron gains and available genomic sequences.

Finally, a related mechanism for intron origin that does not invoke exogenous sequences proposes that tandem duplication of a coding region containing an AGGT sequence could elicit splicing by providing both a 5’ donor and a 3’ acceptor sites (Rogers 1989). The validity of this model is supported by a recent study in humans where the authors found an excess of introns that contain perfect matching sequences at their boundaries (Zhuo et al. 2007) (see also Chabot et al. 2008; Roy and Irimia 2008), with most of these introns being young and undergoing alternative splicing. Interestingly, repeats and alternative splicing have been also shown to be associated with the origin of new exons in vertebrates (Zhang and Chasin 2006).

Empirical Observations in Favor of an Endogenous Origin of Spliceosomal Introns

Hypotheses for the endogenous origin of introns have been challenged, and sometimes discarded, on the basis both of current data and perhaps our biased perception of how gene structure has evolved over time, for example, the frequent assumption that parallel evolution occurs at very low frequency. Recent evidence for the parallel origin of unrelated introns at nearly the same insertion site in \(D\)aphnia \(pulex\) (Omilian et al. 2008) directly challenges this assumption.
Although we may not know for certain how the first introns originated, it is worth emphasizing that we do know, by empirical observations, how individual spliceosomal introns can originate. In plants, efficient splicing typically requires an elevated intron AT content (Goodall and Filipowicz 1989, 1991), probably because the latter favors the association of splicing-enhancing RNA-binding proteins (McCullough et al. 1993; Simpson and Filipowicz 1996). Interestingly, elevated AT content is sufficient for splicing of a nonintronic sequence via activation of flanking cryptic splice sites in a legumin J gene construct (Simpson and Brown 1993). Also, an AT-rich region in the 5′ end of the transferred DNA rolA gene of Agrobacterium rhizogenes, unsurprisingly unspliced in the bacterium, undergoes efficient splicing in Arabidopsis plants that are transgenic for this gene (Magrelli et al. 1994). In Caenorhabditis elegans, de novo introns can be created by recruitment of internal exonic sequences (Irimia et al. 2008), whereas in humans, de novo introns have been suggested to result from 1) fortuitous creation of splice sites (Courseaux and Nahon 2001); 2) activation of cryptic splice sites after the insertion of a transposable element (Sela et al. 2007); 3) tandem duplications (Zhuo et al. 2007); and 4) nucleotide substitutions that do not disrupt splice sites (Gromoll et al. 2007).

**Origin of Spliceosomal Introns by Internal Gene Duplication and Intronization of Translatable Sequences**

Along the same lines of the previously mentioned endogeneous models, we propose here two nonmutually exclusive hypotheses that do not invoke an ancient exogenous source for intron origin but where introns constantly emerge 1) following events of internal gene duplication (we call this the “internal gene duplication” model) (Gao X, unpublished data); and 2) by “intronization” of translatable sequences (the intronization model) (Catania and Lynch 2008).

The former hypothesis is observation driven. Briefly, by surveying multiple eukaryotic genomes, we find that 8–17% of genes display internally duplicated sequences and that internal gene duplication is a steady-state birth-and-death process whose rates are comparable to those reported for events of whole-gene duplication (Lynch and Conery 2003). Surprisingly, we find that 7–30% of internally duplicated genes show new intron gains. These new introns derive only in small part from the direct duplication of prior introns within the same gene; the majority are represented by de novo introns that likely emerged from the duplicated sequence after both the spatial change and the acquisition of new/activation of latent splice sites (Gao X, unpublished data).

The second model attempts to integrate the diverse body of observations that have been reported to date on splicing and interrelated mRNA-associated processes. In the intronization model (Catania and Lynch 2008), the systems for mRNA surveillance, capping, and cleavage/polyadenylation interact to play pivotal roles in the physical establishment and distribution of spliceosomal introns along a transcript. In brief, the model is based on the cell’s ability to filter out aberrant transcripts, whereas sparing those mRNAs where the imperfect splice-site recognition of the spliceosome machinery has fortuitously removed ORF-disrupting mutations such as premature termination codons (PTCs). In the model, we propose also that cleavage/polyadenylation factors (CPFs) regularly access U-rich tracts along the mRNA during transcription but are antagonized (or interfered with) by splicing factors (SFs), when U-rich regions are located within an intron. Under this scenario, the interaction between SFs and CPFs modulates the likelihood of imperfect splice-site recognition, thereby defining the physical setting for the facilitation or inhibition of intron colonization (Catania and Lynch 2008).

**The Crucial Role of Cellular Surveillance Systems in the Endogeneous Origin of Introns**

Our endogenous models rest on several forms of gene structural disruptions. Thus, we expect that transcript surveillance systems, such as nonsense-mediated decay (NMD), must play a significant role in the generation of new endogeneous introns. NMD is a virtually ubiquitous cellular surveillance system in eukaryotes that recognizes and selectively degrades so-called “nonsense” transcripts that contain PTCs upstream of the true stop codon (Maquat 2004, 2006; Conti and Izaurralde 2005; Lynch et al. 2006). Translation of these nonsense transcripts would result in truncated polypeptides, and the failure of NMD surveillance has been implicated in a number of human diseases (Frischmeyer and Dietz 1999; Holbrook et al. 2004). Nonsense transcripts can occur via a variety of mechanisms, including point mutations, indels, transposition/viral insertions, and splicing mistakes and other errors during transcript processing (Mendell et al. 2004) as well as during production of alternatively spliced transcripts (Green et al. 2003; Lewis et al. 2003).

Regardless of the exact molecular mechanism by which degradation of transcripts is initiated (Isken and Maquat 2008), the near ubiquity of general NMD mechanisms in eukaryotes is consistent with the elevated frequency, as well as the likely negative selective importance, of the basic forms of gene structural disruptions that underlie our endogeneous models of intron origin. On appearance, transcripts featuring either internal duplications or de novo intronization are stabilized via NMD degradation of PTC-containing isoforms and the rapidity of fixation is dependent on the accumulation of additional structural modifications that stabilize non-PTC-containing isoforms.

**Interconnection between Internal Gene Duplication and Intronization and the Role of Alternative Splicing**

Although distinct, it is worth emphasizing that the two mechanisms for endogeneous intron origins we present are
Alternative splicing

**Figure 1.** Creation of coding sequences from introns (Exonization) and creation of spliceosomal introns from coding exons (Intronization) are proposed to be reversible processes that occur not instantaneously but through a transient phase of gradual conversion (Alternative splicing).

not mutually exclusive. If fortuitously an internal duplication does not introduce a PTC, then a population of functional transcripts is maintained without NMD. However, if (a portion of) a duplicated exonic region acquires a PTC (see Letunic et al. 2002), then NMD would degrade the primary PTC-containing isoform and the mutated coding sequence could undergo intronization (Catania and Lynch 2008).

A joint action of intronization and internal gene duplication could be conjectured to contribute to maintain optimal protein length. Specifically, because of the frequent occurrence of events of internal gene duplication across several eukaryotes (Kondrashov and Koonin 2001; Letunic et al. 2002; Gao X, unpublished data), molecular mechanisms must exist that assist the removal of the newly created redundant sequences and thereby prevent proteins from growing extremely large. Intronization could represent one of these mechanisms. Once a duplicated sequence converts into an intron, that intron could be lost through the frequent but poorly understood process of intron loss, thus contributing to reduce the increase in gene length resulted by the prior event of duplication.

Finally, intronization of translatable sequences does not occur instantaneously but is predicted to undergo a transient phase involving alternative splicing (Catania and Lynch 2008). Under this view, intronization would not only complement the established process of exonization, where exons arise gradually from intronic sequences (Alekseyenko et al. 2007), but could also be coupled with internal gene duplication, in that young introns that are by-products of tandem duplication have been found to often undergo alternative splicing (Zhuo et al. 2007).

An interesting question associated with such a scenario—in which sequences may gradually and reversibly shift from an exonic to an intronic state over time (Figure 1)—is whether the time interval between sequence type conversion is comparable among eukaryotes. Although it remains unclear how levels of alternative splicing varies across lineages and what fraction of alternative isoforms are truly functional and thus maintained by natural selection, it is tempting to speculate that the duration of the processes of exonization and intronization may vary across species. In particular, the length of the transient phase of alternative splicing could be expected to be reduced in species with large \( N_e \), where natural selection would more efficiently promote the fixation of optimal splicing signals—under the assumption that maintaining alternative splicing is typically mildly deleterious. Under the aforementioned suggestion that exonic sequences may intronize and be subsequently lost, the relative fitness disadvantage that is typically associated with introns (Lynch 2002) may be offset to a greater or lesser degree by the simultaneously acquired selective advantage associated with the production of shorter but potentially more functional proteins. The latter idea is consistent with the observation that 1) lineages having a large \( N_e \) contain shorter protein length than lineages with relatively smaller \( N_e \) (e.g., Xu et al. 2006) and 2) introns abound in lineages with small \( N_e \) (Lynch and Conery 2003).

**Conclusions**

We have briefly outlined a number of observations, published and novel (Gao X, unpublished data), as well as a theoretical model (the intronization model, Catania and Lynch 2008), which 1) are consistent with the idea that spliceosomal introns may have originated endogenously and that 2) shed a new light both on the evolution of the eukaryotic gene structure and on the biological role of alternative splicing. Also, our endogeneous models of intron origin introduce additional mechanisms whereby gene structural evolution can be mediated by NMD (Lynch and Kewalramani 2003; Lynch et al. 2006; Scofield et al. 2007). Finally, we have conjectured that introns may act as a link between internal gene duplication and intronization, where intron loss (successive to intronization) contributes to counteract the expansion of gene length mediated by events of internal gene duplication.

**Funding**

National Science Foundation grant (MCB-0342431) supported F.C., (DBI-0434671) to D.G.S.; National Institutes of Health (F32GM083550) to X.G.; MetaCyte funding from the Lilly Foundation to Indiana University.

**References**


Chabot B, Elèa SA, Zhuo D. 2008. Comment on “when good transcripts go bad: artificial RT-PCR ‘splicing’ and genome analysis”. Bioessays. 30:1256author reply 1257–1258


Rogers JH. 1989. How were introns inserted into nuclear genes. Trends Genet. 5:213–216.


Corresponding Editor: Michael Lynch