OPEN ACCESS Freely available online

PLoS BIOLOGY

Essay

Where Do Introns Come From?

Francesco Catania*, Michael Lynch

In eukaryotes (and viruses), genes may be organized into coding and noncoding regions, called exons and (spliceosomal) introns, respectively (Box 1). Both types of sequences are transcribed into pre-mRNA, but whereas exons are used for protein synthesis, introns are spliced out during/immediately after transcription [1] (Figure 1). Although spliceosomal introns are widespread in the eukaryotic tree, they are unequally distributed across species as a consequence of ongoing intron gain and loss [2,3]. So for instance, 287 spliceosomal introns populate the entire genome of the baker's yeast (*Saccharomyces cerevisiae*) [4], but this number increases to ~4,760 in a different yeast species (*Schizosaccharomyces pombe*) and reaches ~38,000 and ~140,000 in the genomes of the fruit fly *Drosophila melanogaster* and *Homo sapiens*, respectively [5]. Explaining the causes and functional implications of this uneven distribution requires understanding why spliceosomal introns exist in the first place and what the evolutionary origin(s) of these sequences are—a problem that has proved a conundrum for the past 30 years [6].

Box 1. Spliceosomal Introns

Spliceosomal introns are noncoding intervening sequences of eukaryotic and viral genes that are removed during the process of pre-mRNA maturation, leaving only coding sequence (exons) to be part of the messenger RNA. Three additional classes of introns are known—the group-I, group-II, and group-III introns, all of which are capable of self-splicing. Spliceosomal introns cannot self-splice but are removed by a dynamic nuclear apparatus, the spliceosome. Such machinery typically consists of five uridylate-rich small nuclear RNAs (U1, U2, U4, U5, and U6) and a large number of associated proteins.

Did Spliceosomal Introns Insert in or Emerge from Coding Sequences?

Although the evidence is circumstantial, it is widely thought that spliceosomal introns originated from group-II introns—self-splicing introns that are widely found in fungi, plants, protists, and bacteria—which invaded the uninterrupted nuclear genes of an early eukaryote and subsequently lost the ability to self-splice, as the host genome took over this function [7–10]. However, while group-II introns may have spawned the primordial population of spliced introns and the present-day mechanism for their removal (the spliceosome), this model does not provide an explanation for either the irregular distribution of spliceosomal introns in modern-day species [11] or for recent episodes of intron gain [12–14].

Recent studies on vertebrate genome evolution have shown that numerous exons have emerged from pre-existing noncoding sequences (i.e., introns) [15,16]. The acknowledged de novo generation of exons from introns—termed "exonization"—leads us to ask the question: Does the inverse process, i.e., "intronization" or the creation of spliceable sequences from nonintronic regions, take place? A number of empirical observations suggest that intronization is more than a formal possibility [17–23] (see below).

When Nonsense Codons Become Meaningful

The process of translation normally ends when the ribosome reaches a nonsense (or stop) codon at the end of the coding mRNA. A nonsense codon that is positioned upstream of the true stop can lead to premature translation termination and thus is called a premature termination codon (PTC). PTCs found in eukaryotic coding sequences are sometimes excluded from the mature mRNA [24–31], for example by the activation of otherwise latent 5′ splice sites that act to remove the PTC from the transcript [32]. In addition, many eukaryotic PTC-containing mRNAs that are not subject to these nuclear mechanisms of PTC recognition and exclusion are nonetheless subject to degradation in the cytoplasm by the nonsense-mediated decay (NMD) pathway [33]. Notably, as PTC-harboring alleles may have harmful phenotypic effects—since aberrant gene products can lead to cell damage—these mechanisms provide potential cellular routes for reducing the negative effects of PTC-containing alleles. While secondary mutations that re-establish the reading frame are possible, changes that increase the efficiency of spliceosomal removal of the PTC will—at a minimum—also return the mRNA dosage to normal.

The First Steps of the Intronization Hypothesis

We suggest that the cell's capacity to filter out aberrant transcripts, in concert with imperfect splice-site recognition [34,35], may provide a powerful mechanism for generating spliceosomal introns. Specifically, if a PTC-containing exonic region is accidentally spliced out during mRNA maturation, and the open reading frame (ORF) of the transcript is preserved, the NMD pathway will not be elicited by that transcript, providing

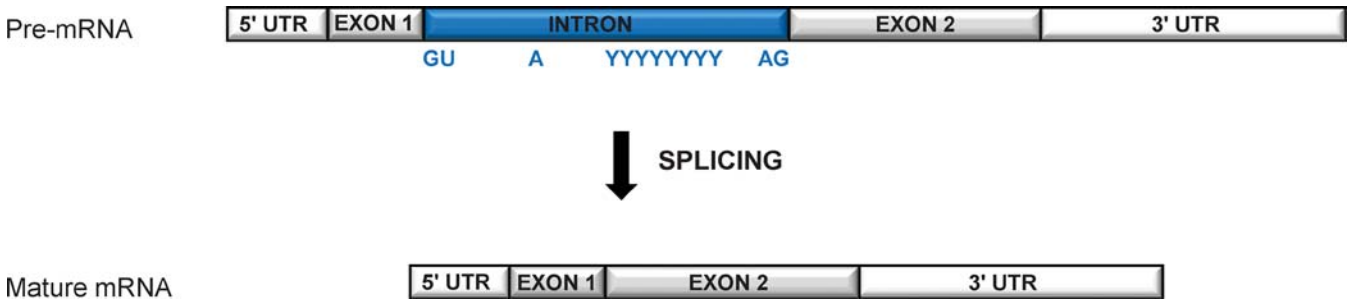
Essays articulate a specific perspective on a topic of broad interest to scientists.

Citation: Catania F, Lynch M (2008) Where do introns come from? PLoS Biol 6(11): e283. doi:10.1371/journal.pbio.0060283

Abbreviations: CBC, capping-binding complex; CPF, cleavage/polyadenylation factor; NMD, nonsense-mediated decay; ORF, open reading frame; PTC, premature translation termination codon; SF, splicing factor; UTR, untranslated region

Francesco Catania and Michael Lynch are at the Department of Biology, Indiana University, Bloomington, Indiana, United States of America.

* To whom correspondence should be addressed. E-mail: fcatania@indiana.edu



doi:10.1371/journal.pbio.0060283.g001

Figure 1. Schematic Example of an Intron-Containing Gene

The regions preceding and following the coding regions (exons) are transcribed but not translated, and are called 5′ and 3′ UTRs. The intronic sequence intervenes between the two coding exons and contains splicing signals that are recognized by a nuclear machinery, the spliceosome, which carries out splicing. Splicing signals are located at both ends of the intron (e.g., canonical GU and AG dinucleotides at the 5′ and 3′ splice site, respectively) and within the intron (an adenine residue, called the branch site, and frequently a polypyrimidine (C/U) tract.

the key first step in the establishment of a new intron.

In this model, we propose that protein-coding sequences that fortuitously contain the minimal requisite sequence information for spliceosome-mediated recognition (see splicing signals in Figure 1) have a latent potential to undergo splicing and to intronize after acquiring PTC mutations that interrupt the ORF of the message. Although splicing to remove a PTC may initially be inefficient, degradation of the pool of unspliced PTC-containing transcripts by NMD will produce a relatively pure pool of PTC-free mature mRNAs [21], thereby maintaining the spliced allele in a possibly still active state. Such an allele can then be subject to positive selection for subsequent mutations that improve splicing of the modified region.

Mutant alleles with PTC-compensating splicing are more likely to become established if they generate proteins that retain at least some activity. Thus, we expect introns arising from this process to share certain characteristics: (1) a short length, thereby minimizing the number of lost codons; and (2) a sequence length that is a multiple of three, so as to preserve the ORF. Also, as the emergence of introns in small exons would lead to the creation of two even smaller flanking exons, whose correct splicing might be compromised [36], we expect either that introns never (or rarely) emerge in short exons or that intronization includes the whole exon. In the latter case, notably, if the small exon is not terminal, the intronization process would lead to the merger of two introns and the encompassed exon and hence to the loss, rather than a gain, of an intron. Unless excision of the newly intronized coding sequence has sufficiently large deleterious consequences, the fixation of the novel intron may be either selectively neutral or promoted by natural selection.

The Role of Alternative Splicing in the Process of Intronization

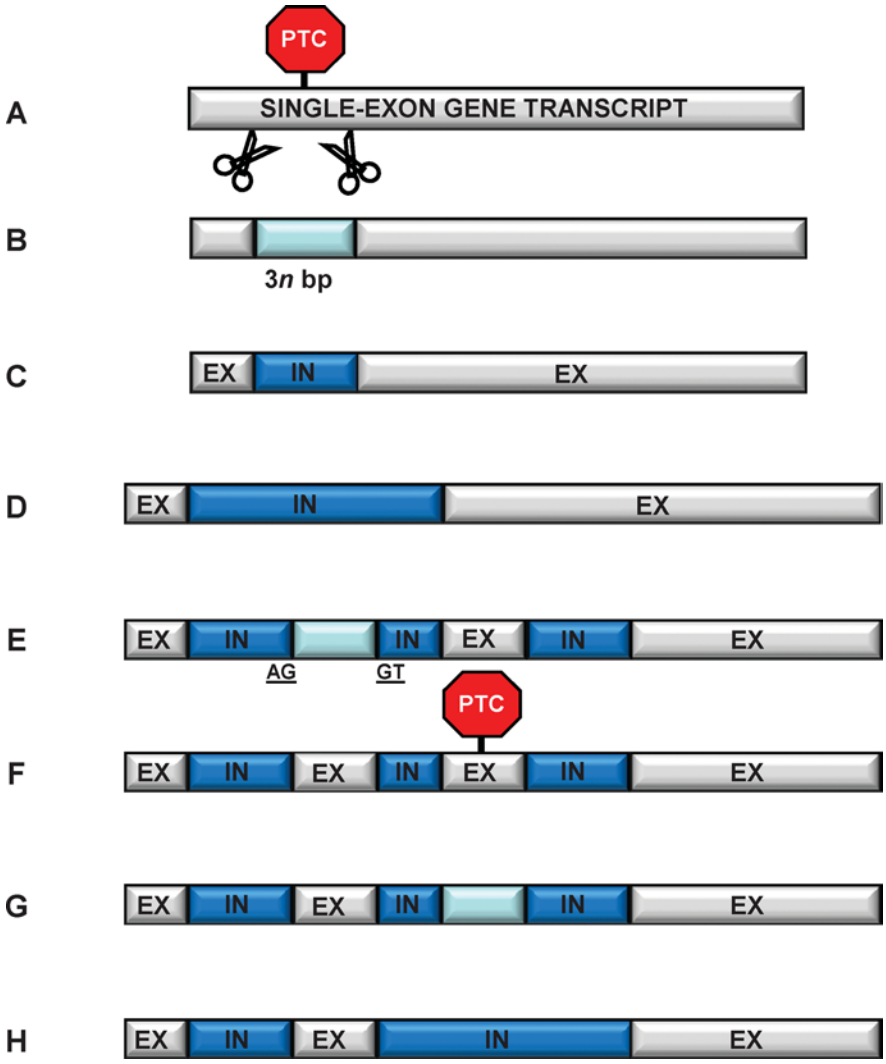
As it is unlikely that the process of intronization is instantaneous, we predict the existence of a transient phase during which gene regions are neither fully exonic nor fully intronic, essentially exhibiting the features of DNA sequences undergoing alternative splicing. Alternative splicing is a nuclear process that leads to the incorporation of noncoding regions and/or the exclusion of coding regions from mature mRNAs [37], events that bear on the original definition of exon and intron [6].

Several findings support the hypothesized gradual conversion from exonic to intronic sequences. Specifically, across vertebrates, constitutively spliced introns in one species often align to homologous alternatively spliced coding sequences of orthologous genes from another species [16]—suggesting a source–product relationship between the two types of sequences; and alternatively spliced exons in vertebrates have been shown to emerge from constitutive exons [38,39]. Similarly, alternatively spliced introns have been found to share multiple features with exonic sequences in humans [40] and to emerge from constitutive exonic sequences in nematodes [41]. All these observations are consistent with an evolutionary loop and a shared evolutionary history between at least a subset of exons and introns (Figure 2).

The gradual conversion that we envision for the process of intronization has been reported also for the process of exonization [42]. In particular, most exons that have emerged from noncoding sequences are alternatively spliced, often being minor forms, i.e., exons that are rarely included in the mature transcript [15,16,43–47]. It is interesting to note that DNA regions undergoing exonization are commonly identified under the assumption of no parallel exon losses across species (e.g., [16]). The validity of this assumption is uncertain, and if parallel events of exon loss are relatively frequent, then several reported cases of exon gain would have to be recategorized as exon losses, i.e., events of intronization of coding sequences. The latter scenario is consistent with some remarkable feature similarities of putatively young (minor-form) exons and young introns, as predicted by the intronization model. Specifically, minor-form exons tend to be: $3n$ in size when located within the coding region [48]; unusually short; fast-evolving; PTC-enriched [49,50]; and located in polypeptide segments that have no or very little immediate effect on protein structure [51,52]; they also have weaker splice sites compared to constitutively spliced exons [53–60].

Introns as a Result of the Crosstalk between mRNA-Associated Processes

The extensive network of interactions between mRNA-associated processes [61] suggests that other mechanisms, in addition to NMD, may be involved in the origin (and evolution) of



doi:10.1371/journal.pbio.0060283.g002

Figure 2. Schematic Example for a Hypothesized Temporal Succession of Exonization and Intronization Processes

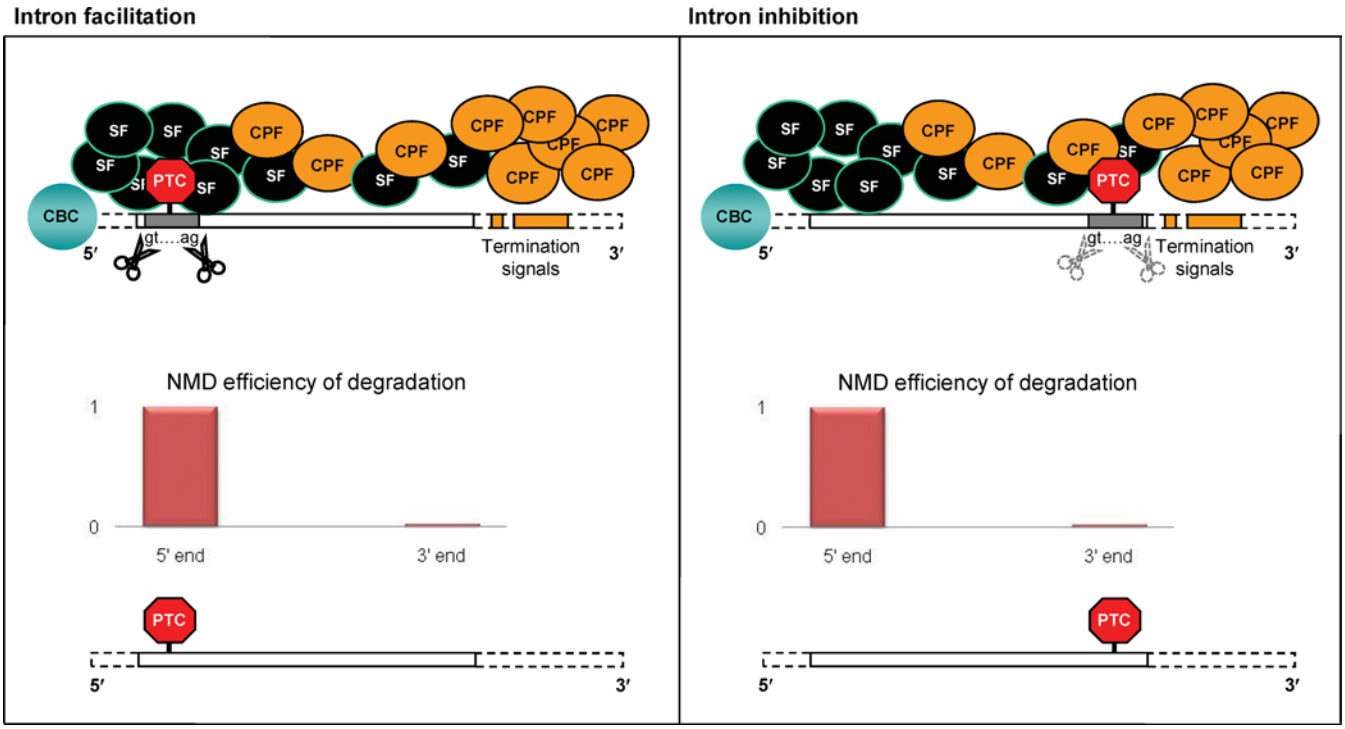
(A) A PTC-containing exonic region is fortuitously spliced. (B) If the region spliced has a length that is a multiple of 3, and its absence is not detrimental to the functionality of the coded protein, the exonic region may start being skipped, with its incorporation in the mature mRNA gradually decreasing, essentially undergoing a phase of alternative splicing (indicated in light blue). (C) The constitutive intron (IN) is created, flanked by two constitutive exons (EX); and (D) may grow larger (for example, as a consequence of insertion accumulation). (E) A second constitutive intron arises downstream in a way similar to that described for the previous intron, while a process of exonization starts within the previous intron. (F) The novel, initially alternatively spliced exon now becomes constitutively spliced, while the exon downstream of it acquires a PTC and may start the process of intronization. (G) The intronization process of the PTC-containing exon undergoes a phase of alternative splicing, which subsequently leads to (H) the technical loss of one intron and one exon and the creation of a single, larger intron.

spliceosomal introns. Here, we examine the role played by cleavage/polyadenylation factors (CPFs) and the mRNA capping-binding complex (CBC). CPFs bind 3′ untranslated region (UTR) sequence signals (positioned just past the stop codon in the coding region) and are actively involved in mRNA 3′ end formation, a process that broadly consists of cleaving the nascent transcript and adding a tail of multiple adenines to its 3′ end. The CBC is a structure that is added to the mRNA 5′ end immediately after the start of transcription and regulates several steps of mRNA metabolism [62].

Several connections have been found between the processes of cleavage/polyadenylation and splicing [63–71], and splicing factors (SFs) and CPFs have also been documented to compete or interfere with each other [72–75]. Although the targets of such competition remain unknown, in plants AU-richness, and U-richness in particular, appears to be not only a landmark for intron recognition but also a signal for CPFs [19,76–78]. The latter finding is consistent with U-rich sequences directing transcription termination in several eukaryotes and viruses [79]. Notably, U-rich sequences, such as the polypyrimidine tract (Figure 1), are also present in most eukaryotic introns and play a significant role in the splicing process [80].

We propose that CPFs regularly access U-rich tracts along the mRNA during transcription, but are antagonized (or interfered with) by SFs when the U-rich regions are located within an intron. Notably, these two sets of factors are also known to antagonize/interfere in exons, as the U1 small nuclear ribonucleoprotein (an SF) inhibits 3′-end processing when bound to the 3′-end of a pre-mRNA in the vicinity of the cleavage-polyadenylation site [81–86]. Under our hypothesis, the interaction between SFs and CPFs in exon sequences modulates the likelihood that PTC mutations will be removed by a splicing event, thereby defining the physical setting for the facilitation or inhibition of intron colonization (Figure 3). The ability of CPFs to contact U-rich sequences and block SFs is expected to be affected by diverse factors, including the distance and the strength of the 5′ splice site [19], the presence of splicing-modulating sequences (such as splicing enhancers), the local concentration of splicing proteins [87], the transcription elongation rate [88,89], and the mRNA secondary structure [90]. Optimal splicing conditions also promote transcription elongation [91,92] and termination [65,93], whereas weaker splicing conditions facilitate the binding of CPFs, inhibiting the binding of SFs to the polypyrimidine tract.

The CBC also influences splicing, acting as a splicing enhancer by increasing the population of SFs local to the 5′ end [94–96], thus favoring splicing at this end of the transcript. At the 3′ end, the presence of strong canonical termination signals favors the recruitment of CPFs to terminal U-rich DNA stretches, thus inhibiting the potential assembly of SFs in this region. As a result, the CBC and the resultant excess of SFs are expected to



doi:10.1371/journal.pbio.0060283.g003

Figure 3. A PTC in a Coding Region Typically Elicits a Translation-Dependent Surveillance Mechanism, Such As NMD, Which Leads to the Degradation of the Aberrant Transcript

If the mRNA region containing the PTC harbors fortuitous recognition elements for its spliceosome-mediated removal (e.g., latent splice sites), a PTC-containing segment may be spliced out during mRNA maturation (in grey). The likelihood with which accidental splicing of an entirely new intron may occur is expected to be higher in regions of the transcript where the concentration of SFs is naturally elevated (e.g., at the 5′ end, in proximity of the CBC), compared to the mRNA 3′ end, where strong canonical termination signals (in orange) favor the preferential binding of CPFs that, under the proposed model, compete/interfere with SFs for the binding of U-rich tracts. The fortuitous gain of introns is favored at the 5′ end because unspliced PTC-containing transcripts in this region are more efficiently degraded, thereby alleviating the negative cellular consequences of the PTC.

enhance the frequency of fortuitous splicing events at the 5′ end, while the presence of strong termination signals is expected to reduce the frequency of fortuitous splicing events at the 3′ end.

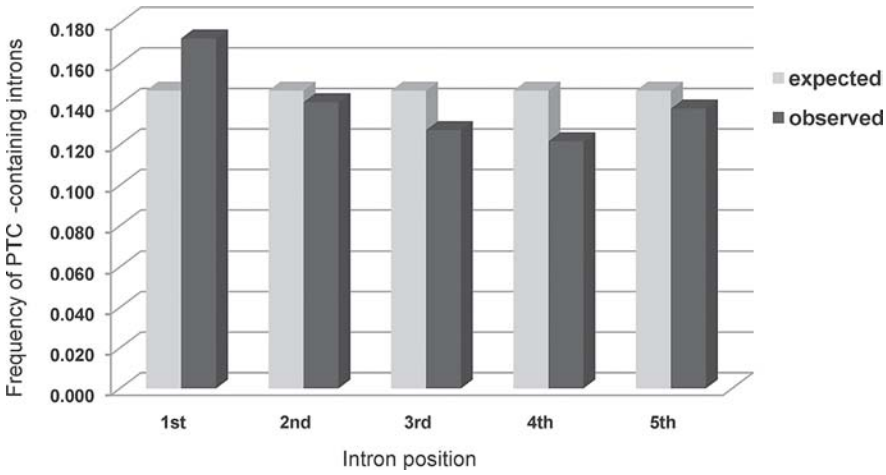
Finally, the antagonistic interactions between SFs and CPFs are likely mediated by two other major classes of competing proteins [97], namely the serine/arginine-rich proteins and the heterogeneous nuclear ribonucleoproteins [87]. Notably, heterogeneous nuclear ribonucleoproteins have also been suggested to both participate in transcription termination and bind the polypyrimidine tract [79,98–100].

Central to our hypothesis is the observation that the more favorable splicing at the 5′ end of a gene parallels the spatial pattern of the efficiency of NMD degradation of aberrant transcripts. Specifically, NMD effectiveness is maximal when a PTC is proximal to the 5′ end of the mRNA and minimal when it resides in the most 3′ exon, close to the 3′ end of the transcript [101–104] (Figure 3). Thus, not only are splicing-eliciting PTCs expected to arise more frequently in the 5′ ends of genes, but such modifications also have the greatest chance of emerging as novel introns with minimal fitness effects.

Support for the Intronization Hypothesis

The distribution of introns over the length of the coding sequence is consistent with the idea that NMD, as well as the interactions between CPFs and SFs, cooperatively guides the successful colonization by introns. In particular, the NMD pathway appears to have been lost in eukaryotic lineages that have no or nearly no introns [105]. While it is not possible to rule out that NMD could be simply lost in situations where introns are rare (e.g., as a consequence of genome reduction), we suggest that, as introns are not essential to the functioning of NMD [101,103,106–111], by increasing the costs of imperfect splicing, the loss of NMD produces an environment that inhibits intron colonization.

As for preferential intron location, assuming a steady-state process of intron birth and death, an increase in intron birth is expected to shift the age distribution to younger introns. Under our hypothesis, young introns are expected to be biased toward lengths that are multiples of three, to be relatively short, and to contain a PTC. These expectations fit the observations of a recent study where PTC-containing $3n$ introns in the ciliate *Paramecium tetraurelia* were revealed to be about twice as frequent compared to PTC-containing introns of the two other size classes [112]. Our own study of the intron dataset used in the latter study shows that this higher frequency is independent of the position occupied along the transcript (data not shown), and that PTC-containing introns are over-represented at the 5′ end of transcripts (1st intron position, $\chi^2 = 23.26$, $p = 1.41 \times 10^{-6}$) but less frequent toward the 3′ end (3rd intron position, $\chi^2 = 7.93$, $p = 4.87 \times 10^{-3}$; 4th intron position, $\chi^2 = 6.44$, $p = 0.0111$), consistent with the idea that splicing-



doi:10.1371/journal.pbio.0060283.g004

Figure 4. Frequency of PTC-Containing Introns in *P. tetraurelia*
This analysis has been performed using 15,286 EST-confirmed introns [112] and does not include single-intron genes.

eliciting PTCs arise more frequently in the 5′ ends of genes (Figure 4).

Our hypothesis specifically predicts that small 3*n* introns should be enriched with PTCs, either as a consequence of these stop codons eliciting intronization and/or because PTCs are secondarily selected for as a means to detect erroneously spliced transcripts. This prediction is supported by the significant under-representation of PTC-free 3*n* introns associated with short intron size in six different eukaryotes [112].

Explaining Introns in Untranslated DNA Sequences

Within a gene, spliceosomal introns can also reside in UTRs [113], and UTR introns show similar patterns of frequency and spatial distributions in distantly related species [114,115]: 5′-UTR introns are frequent and dispersed at random, while 3′-UTR introns are very rare, despite the fact that 3′ UTRs are typically about two to three times longer than 5′ UTRs.

In light of the intronization model, these features of introns in UTRs can be explained in two non-mutually exclusive ways. First, a significant fraction of today's intron-containing UTRs may have been coding sequences at the time of intron addition. In support of this scenario, the translatability of a number of ORFs residing in currently annotated UTRs has been shown [116–119]. Second, the emergence of introns in 5′ UTRs may be associated with the potentially deleterious effects of upstream premature translation start AUG codons. Simply put, we suggest that whereas PTCs may encourage the gain of internal introns, premature translation start codons may encourage the gain of 5′-UTR external introns. The latter scenario is consistent with an elevated abundance of AUGs in 5′-UTR introns [120].

Spliceosomal introns primarily inhabit protein-coding genes, but they also sometimes interrupt noncoding RNA genes [121,122]. Although the proposed hypothesis does not claim to explain the origin of all introns, it is worth noting that the presence of spliceosomal introns in noncoding RNA genes might also be the result of accidental splicing events and of the subsequent proofreading activity of surveillance mechanisms. In particular, although no translation has been reported for the products of these genes, experimental evidence suggests that, like mRNAs, noncoding RNAs are also subject to post-transcriptional surveillance pathways [123]. A possible beneficial effect of a splicing event is the improvement in the folding of the mature RNA, and consistent with this possibility, the noncoding RNA quality-control step appears to target molecules that are either misfolded or contain functionally deleterious mutations [124–126]. Thus, as in the case of protein-coding genes, it can be postulated that fortuitous endogenous events may on rare occasions promote splicing in noncoding RNAs, in such a way as to prevent more harmful secondary structures.

Why Don't All Eukaryotic and Viral Genes Contain Introns?

Two possible explanations for the existence of intronless genes are: (1) that introns can simply be lost, so that a subset of intron-free genes is to be expected; and (2) that some intronless genes may be derived retrogenes, i.e., mature mRNAs that are reverse transcribed into DNA copies and inserted into the genome [127]. However, splicing is known to affect mRNA export into the cytoplasm, as unspliced transcripts usually accumulate in the nucleus [128,129]. How then can transcripts of intronless genes accumulate in the cytoplasm? A number of eukaryotic and viral single-exon genes have been found to contain sequence elements that favor nucleus-cytoplasm export [130–132]. Notably, results both from in vivo and in vitro experiments show that such elements not only play a major role in nuclear export but also enhance polyadenylation and strongly inhibit splicing, thereby inhibiting intron colonization [133–135]. These findings suggest that (ancestrally or derived) intronless genes that contain the aforementioned sequence elements are unlikely to gain introns, simply because of their intrinsic resistance to the splicing apparatus. Although it remains to be proven, it is possible that the relative abundance of these elements that inhibit splicing plays a role in establishing different levels of intron-richness between eukaryotic species (e.g., between *Sa. cerevisiae* and *Sc. pombe*).

Conclusions

We have proposed a novel hypothesis for the origin of spliceosomal introns, invoking endogenous production within translatable sequences (at least in the case of protein-coding genes), facilitated by the activity of cellular surveillance mechanisms. Despite the mutational hazard associated with intron presence and proliferation [136], we argue that, at least initially, introns might represent a favorable life line for an allele that has acquired an ORF-disrupting mutation. In this sense, in-frame stop codons need not be dead ends, as often believed, but rather sequences that occasionally facilitate the evolution of eukaryotic gene structure, possibly favoring not only intronization, but also processes

such as exonization (following a PTC loss [137]). Further experimental validation of our hypothesis would not only support the idea that intron birth/death rates depend on both the population-genetic [136] and the intracellular environment, but also shed light on a surprising aspect of the evolution of eukaryotic gene structure, i.e., the ongoing, stochastic process of mutual conversion between exons and introns within genes. ■

Acknowledgments

We are grateful to Thomas G. Doak, J. Ignasi Lucas, M. Virginia Sanchez Puerta, Xiang Gao, and Andy Alverson for constructive discussions and to anonymous reviewers for suggestions for the improvement of the manuscript. We thank Eric Meyer for providing the EST-confirmed set of *P. tetraurelia* introns.

Funding. This work was supported by the National Science Foundation grant MCB-0342431 to ML and MetaCyte funding from the Lilly Foundation to Indiana University.

References

1. Burge CB, Tuschl T, Sharp PA (1999) Splicing of precursors to mRNAs by the spliceosomes. In: Gesteland RF, Cech T, Atkins JF, editors. The RNA world. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press. pp. 525-560.
2. Qiu WG, Schisler N, Stoltzfus A (2004) The evolutionary gain of spliceosomal introns: Sequence and phase preferences. Mol Biol Evol 21: 1252-1263.
3. Roy SW, Fedorov A, Gilbert W (2003) Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. Proc Natl Acad Sci U S A 100: 7158-7162.
4. Juneau K, Palm C, Miranda M, Davis RW (2007) High-density yeast-tiling array reveals previously undiscovered introns and extensive regulation of meiotic splicing. Proc Natl Acad Sci U S A 104: 1522-1527.
5. Mourier T, Jeffares DC (2003) Eukaryotic intron loss. Science 300: 1393.
6. Gilbert W (1978) Why genes in pieces. Nature 271: 501-501.
7. Cavalier-Smith T (1991) Intron phylogeny—A new hypothesis. Trends Genet 7: 145-148.
8. Sharp PA (1985) On the origin of RNA splicing and introns. Cell 42: 397-400.
9. Cech TR (1986) The generality of self-splicing RNA—Relationship to nuclear messenger-RNA splicing. Cell 44: 207-210.
10. Jacquier A (1990) Self-splicing group-ii and nuclear pre-messenger-rna introns—How similar are they. Trends Biochem Sci 15: 351-354.
11. Roy SW, Gilbert W (2006) The evolution of spliceosomal introns: Patterns, puzzles and progress. Nat Rev Genet 7: 211-221.
12. Coulombe-Huntington J, Majewski J (2007) Intron loss and gain in *Drosophila*. Mol Biol Evol 24: 2842-2850.
13. Sverdlov AV, Babenko VN, Rogozin IB, Koonin EV (2004) Preferential loss and gain of introns in 3′ portions of genes suggests a reverse-transcription mechanism of intron insertion. Gene 338: 85-91.
14. Roy SW, Penny D (2007) A very high fraction of unique intron positions in the intron-rich diatom *Thalassiosira pseudonana* indicates widespread intron gain. Mol Biol Evol 24: 1447-1457.
15. Wang W, Zheng HK, Yang S, Yu HJ, Li J, et al. (2005) Origin and evolution of new exons in rodents. Genome Res 15: 1258-1264.
16. Alekseyenko AV, Kim N, Lee CJ (2007) Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes. RNA 13: 661-670.
17. Gromoll J, Lahrmann L, Godmann M, Muller T, Michel C, et al. (2007) Genomic checkpoints for exon 10 usage in the luteinizing hormone receptor type 1 and type 2. Mol Endocrinol 21: 1984-1996.
18. Zhuo D, Madden R, Elela SA, Chabot B (2007) Modern origin of numerous alternatively spliced human introns from tandem arrays. Proc Natl Acad Sci U S A 104: 882-886.
19. Luehrsen KR, Walbot V (1994) Intron creation and polyadenylation in maize are directed by AU-rich RNA. Genes Dev 8: 1117-1130.
20. Simpson CG, Brown JW (1993) Efficient splicing of an AU-rich antisense intron sequence. Plant Mol Biol 21: 205-211.
21. Rushforth AM, Anderson P (1996) Splicing removes the *Caenorhabditis elegans* transposon Tc1 from most mutant pre-mRNAs. Mol Cell Biol 16: 422-429.
22. Fridell RA, Pret AM, Searles LL (1990) A retrotransposon 412 insertion within an exon of the *Drosophila melanogaster* vermilion gene is spliced from the precursor RNA. Genes Dev 4: 559-566.
23. Giroux MJ, Clancy M, Baier J, Ingham L, McCarty D, et al. (1994) *De novo* synthesis of an intron by the maize transposable element Dissociation. Proc Natl Acad Sci U S A 91: 12150-12154.
24. Cartegni L, Chew SL, Krainer AR (2002) Listening to silence and understanding nonsense: Exonic mutations that affect splicing. Nat Rev Genet 3: 285-298.
25. Valentine CR (1998) The association of nonsense codons with exon skipping. Mutat Res 411: 87-117.
26. Dietz HC, Kendzior RJ Jr. (1994) Maintenance of an open reading frame as an additional level of scrutiny during splice site selection. Nat Genet 8: 183-188.
27. Dietz HC, Valle D, Francomano CA, Kendzior RJ Jr., Pyeritz RE, et al. (1993) The skipping of constitutive exons in vivo induced by nonsense mutations. Science 259: 680-683.
28. Wang J, Hamilton JI, Carter MS, Li S, Wilkinson MF (2002) Alternatively spliced TCR mRNA induced by disruption of reading frame. Science 297: 108-110.
29. Gersappe A, Burger L, Pintel DJ (1999) A premature termination codon in either exon of minute virus of mice P4 promoter-generated pre-mRNA can inhibit nuclear splicing of the intervening intron in an open reading frame-dependent manner. J Biol Chem 274: 22452-22458.
30. Naeger LK, Schoborg RV, Zhao Q, Tullis GE, Pintel DJ (1992) Nonsense mutations inhibit splicing of MVM RNA in cis when they interrupt the reading frame of either exon of the final spliced product. Genes Dev 6: 1107-1119.
31. Hentze MW, Kulozik AE (1999) A perfect message: RNA surveillance and nonsense-mediated decay. Cell 96: 307-310.
32. Li BH, Wachtel C, Miriami E, Yahalom G, Friedlander G, et al. (2002) Stop codons affect 5′ splice site selection by surveillance of splicing. Proc Natl Acad Sci U S A 99: 5277-5282.
33. Chang YF, Imam JS, Wilkinson MF (2007) The nonsense-mediated decay RNA surveillance pathway. Annu Rev Biochem 76: 51-74.
34. Lu QL, Morris GE, Wilton SD, Ly T, Artem'yeva OV, et al. (2000) Massive idiosyncratic exon skipping corrects the nonsense mutation in dystrophic mouse muscle and produces functional revertant fibers by clonal expansion. J Cell Biol 148: 985-996.
35. Matsuzaki Y, Nakano A, Jiang QJ, Pulkkinen L, Uitto J (2005) Tissue-specific expression of the ABCC6 gene. J Invest Dermatol 125: 900-905.
36. Dominski Z, Kole R (1991) Selection of splice sites in pre-mRNAs with short internal exons. Mol Cell Biol 11: 6075-6083.
37. Ast G (2004) How did alternative splicing evolve? Nat Rev Genet 5: 773-782.
38. Lev-Maor G, Goren A, Sela N, Kim E, Keren H, et al. (2007) The "alternative" choice of constitutive exons throughout evolution. PLoS Genet 3(11): e203. doi:10.1371/journal.pgen.0030203
39. Koren E, Lev-Maor G, Ast G (2007) The emergence of alternative 3′ and 5′ splice site exons from constitutive exons. PLoS Comput Biol 3(5): e95. doi:10.1371/journal.pcbi.0030095
40. Galante PA, Sakabe NJ, Kirschbaum-Slager N, de Souza SJ (2004) Detection and evaluation of intron retention events in the human transcriptome. RNA 10: 757-765.
41. Irimia M, Rukov JL, Penny D, Vinther J, Garcia-Fernandez J, et al. (2008) Origin of introns by 'intronization' of exonic sequences. Trends Genet 24: 378-381.
42. Modrek B, Lee CJ (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat Genet 34: 177-180.
43. Zhang XHF, Chasin LA (2006) Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. Proc Natl Acad Sci U S A 103: 13427-13432.
44. Sela N, Mersch B, Gal-Mark N, Lev-Maor G, Hotz-Wagenblatt A, et al. (2007) Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. Genome Biol 8: R127.
45. Wang W, Kirkness EF (2005) Short interspersed elements (SINEs) are a major source of canine genomic diversity. Genome Res 15: 1798-1808.
46. Krull M, Petrusma M, Makalowski W, Brosius J, Schmitz J (2007) Functional persistence of exonized mammalian-wide interspersed repeat elements (MIRs). Genome Res 17: 1139-1145.
47. Sorek R, Ast G, Graur D (2002) Alu-containing exons are alternatively spliced. Genome Res 12: 1060-1067.
48. Resch A, Xing Y, Alekseyenko A, Modrek B, Lee C (2004) Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. Nucleic Acids Res 32: 1261-1269.
49. Xing Y, Lee C (2005) Assessing the application of Ka/Ks ratio test to alternatively spliced exons. Bioinformatics 21: 3701-3703.
50. Sorek R, Shamir R, Ast G (2004) How prevalent is functional alternative splicing in the human genome? Trends Genet 20: 68-71.
51. Romero PR, Zaidi S, Fang YY, Uversky VN, Radivojac P, et al. (2006) Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. Proc Natl Acad Sci U S A 103: 8390-8395.
52. Kriventseva EV, Koch I, Apweiler R, Vingron M, Bork P, et al. (2003) Increase of functional diversity by alternative splicing. Trends Genet 19: 124-128.
53. Clark F, Thanaraj TA (2002) Categorization and characterization of transcript-confirmed

constitutively and alternatively spliced introns and exons from human. Hum Mol Genet 11: 451-464.
54. Zavolan M, Kondo S, Schonbach C, Adachi J, Hume DA, et al. (2003) Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. Genome Res 13: 1290-1300.
55. Baek D, Green P (2005) Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. Proc Natl Acad Sci U S A 102: 12813-12818.
56. Zheng CL, Fu XD, Gribskov M (2005) Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. RNA 11: 1777-1787.
57. Itoh H, Washio T, Tomita M (2004) Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes. RNA 10: 1005-1018.
58. Xing Y, Lee CJ (2005) Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. PLoS Genet 1(3): e34. doi:10.1371/journal.pgen.0010034
59. Garg K, Green P (2007) Differing patterns of selection in alternative and constitutive splice sites. Genome Res 17: 1015-1022.
60. Xing Y, Lee C (2005) Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. Proc Natl Acad Sci U S A 102: 13526-13531.
61. Maniatis T, Reed R (2002) An extensive network of coupling among gene expression machines. Nature 416: 499-506.
62. Banerjee AK (1980) 5′-terminal cap structure in eucaryotic messenger ribonucleic acids. Microbiol Rev 44: 175-205.
63. Cooke C, Hans H, Alwine JC (1999) Utilization of splicing elements and polyadenylation signal elements in the coupling of polyadenylation and last-intron removal. Mol Cell Biol 19: 4971-4979.
64. Ortiz DF, Strommer JN (1990) The Mu1 maize transposable element induces tissue-specific aberrant splicing and polyadenylation in two Adh1 mutants. Mol Cell Biol 10: 2090-2095.
65. Dye MJ, Proudfoot NJ (1999) Terminal exon definition occurs cotranscriptionally and promotes termination of RNA polymerase II. Mol Cell 3: 371-378.
66. Pandey NB, Chodchoy N, Liu TJ, Marzluff WF (1990) Introns in histone genes alter the distribution of 3′ ends. Nucleic Acids Res 18: 3161-3170.
67. Vagner S, Vagner C, Mattaj IW (2000) The carboxyl terminus of vertebrate poly(A) polymerase interacts with U2AF 65 to couple 3′-end processing and splicing. Genes Dev 14: 403-413.
68. Gunderson SI, Beyer K, Martin G, Keller W, Boelens WC, et al. (1994) The human U1A snRNP protein regulates polyadenylation via a direct interaction with poly(A) polymerase. Cell 76: 531-541.
69. Lutz CS, Murthy KG, Schek N, O'Connor JP, Manley JL, et al. (1996) Interaction between the U1 snRNP-A protein and the 160-kD subunit of cleavage-polyadenylation specificity factor increases polyadenylation efficiency in vitro. Genes Dev 10: 325-337.
70. Kyburz A, Friedlein A, Langen H, Keller W (2006) Direct interactions between subunits of CPSF and the U2 snRNP contribute to the coupling of pre-mRNA 3′ end processing and splicing. Mol Cell 23: 195-205.
71. Proudfoot NJ, Furger A, Dye MJ (2002) Integrating rnRNA processing with transcription. Cell 108: 501-512.
72. Peterson ML, Perry RP (1989) The regulated production of mu m and mu s mRNA is dependent on the relative efficiencies of mu s poly(A) site usage and the c mu 4-to-M1 splice. Mol Cell Biol 9: 726-738.
73. Takagaki Y, Seipelt RL, Peterson ML, Manley JL (1996) The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. Cell 87: 941-952.
74. Leff SE, Evans RM, Rosenfeld MG (1987) Splice commitment dictates neuron-specific alternative RNA processing in calcitonin/CGRP gene expression. Cell 48: 517-524.
75. Lou H, Neugebauer KM, Gagel RF, Berget SM (1998) Regulation of alternative polyadenylation by U1 snRNPs and SRp20. Mol Cell Biol 18: 4977-4985.
76. Ko CH, Brendel V, Taylor RD, Walbot V (1998) U-richness is a defining feature of plant introns and may function as an intron recognition signal in maize. Plant Mol Biol 36: 573-583.
77. Goodall GJ, Filipowicz W (1989) The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing. Cell 58: 473-483.
78. Gniadkowski M, Hemmings-Mieszczak M, Klahre U, Liu HX, Filipowicz W (1996) Characterization of intronic uridine-rich sequence elements acting as possible targets for nuclear proteins during pre-mRNA splicing in *Nicotiana plumbaginifolia*. Nucleic Acids Res 24: 619-627.
79. Zhao J, Hyman L, Moore C (1999) Formation of mRNA 3′ ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. Microbiol Mol Biol Rev 63: 405-445.
80. Singh R, Valcarcel J, Green MR (1995) Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. Science 268: 1173-1176.
81. Furth PA, Baker CC (1991) An element in the bovine papillomavirus late 3′ untranslated region reduces polyadenylated cytoplasmic RNA levels. J Virol 65: 5806-5812.
82. Furth PA, Choe WT, Rex JH, Byrne JC, Baker CC (1994) Sequences homologous to 5′ splice sites are required for the inhibitory activity of papillomavirus late 3′ untranslated regions. Mol Cell Biol 14: 5278-5289.
83. Ashe MP, Griffin P, James W, Proudfoot NJ (1995) Poly(A) site selection in the HIV-1 provirus: inhibition of promoter-proximal polyadenylation by the downstream major splice donor site. Genes Dev 9: 3008-3025.
84. Ashe MP, Pearson LH, Proudfoot NJ (1997) The HIV-1 5′ LTR poly(A) site is inactivated by U1 snRNP interaction with the downstream major splice donor site. Embo J 16: 5752-5763.
85. Vagner S, Ruegsegger U, Gunderson SI, Keller W, Mattaj IW (2000) Position-dependent inhibition of the cleavage step of pre-mRNA 3′-end processing by U1 snRNP. RNA 6: 178-188.
86. Abad X, Vera M, Jung SP, Oswald E, Romero I, et al. (2008) Requirements for gene silencing mediated by U1 snRNA binding to a target sequence. Nucleic Acids Res 36: 2338-2352.
87. Smith CWJ, Valcarcel J (2000) Alternative pre-mRNA splicing: the logic of combinatorial control. Trends Biochem Sci 25: 381-388.
88. Cramer P, Pesce CG, Baralle FE, Kornblihtt AR (1997) Functional association between promoter structure and transcript alternative splicing. Proc Natl Acad Sci U S A 94: 11456-11460.
89. de la Mata M, Alonso CR, Kadener S, Fededa JP, Blaustein M, et al. (2003) A slow RNA polymerase II affects alternative splicing in vivo. Mol Cell 12: 525-532.
90. Eperon LP, Graham IR, Griffiths AD, Eperon IC (1988) Effects of RNA secondary structure on alternative splicing of pre-mRNA: Is folding limited to a region behind the transcribing RNA polymerase? Cell 54: 393-401.
91. Ares M Jr., Grate L, Pauling MH (1999) A handful of intron-containing genes produces the lion's share of yeast mRNA. RNA 5: 1138-1139.
92. Fong YW, Zhou Q (2001) Stimulatory effect of splicing factors on transcriptional elongation. Nature 414: 929-933.
93. McCracken S, Lambermon M, Blencowe BJ (2002) SRm160 splicing coactivator promotes transcript 3′-end cleavage. Mol Cell Biol 22: 148-160.
94. Lewis JD, Izaurralde E, Jarmolowski A, McGuigan C, Mattaj IW (1996) A nuclear cap-binding complex facilitates association of U1 snRNP with the cap-proximal 5′ splice site. Genes Dev 10: 1683-1698.
95. Inoue K, Ohno M, Sakamoto H, Shimura Y (1989) Effect of the cap structure on pre-mRNA splicing in *Xenopus* oocyte nuclei. Genes Dev 3: 1472-1479.
96. Ohno M, Sakamoto H, Shimura Y (1987) Preferential excision of the 5′ proximal intron from mRNA precursors with two introns as mediated by the cap structure. Proc Natl Acad Sci U S A 84: 5187-5191.
97. Caceres JF, Stamm S, Helfman DM, Krainer AR (1994) Regulation of alternative splicing in vivo by overexpression of antagonistic splicing factors. Science 265: 1706-1709.
98. Garcia-Blanco MA, Jamison SF, Sharp PA (1989) Identification and purification of a 62,000-dalton protein that binds specifically to the polypyrimidine tract of introns. Genes Dev 3: 1874-1886.
99. Wagner EJ, Garcia-Blanco MA (2001) Polypyrimidine tract binding protein antagonizes exon definition. Mol Cell Biol 21: 3281-3288.
100. Sharma S, Falick AM, Black DL (2005) Polypyrimidine tract binding protein blocks the 5′ splice site-dependent assembly of U2AF and the prespliceosomal E complex. Mol Cell 19: 485-496.
101. van Hoof A, Green PJ (1996) Premature nonsense codons decrease the stability of phytohemagglutinin mRNA in a position-dependent manner. Plant J 10: 415-424.
102. Baumann B, Potash MJ, Kohler G (1985) Consequences of frameshift mutations at the immunoglobulin heavy chain locus of the mouse. Embo J 4: 351-359.
103. Longman D, Plasterk RH, Johnstone IL, Caceres JF (2007) Mechanistic insights and identification of two novel factors in the *C. elegans* NMD pathway. Genes Dev 21: 1075-1085.
104. Silva AL, Pereira FJ, Morgado A, Kong J, Martins R, et al. (2006) The canonical UPF1-dependent nonsense-mediated mRNA decay is inhibited in transcripts carrying a short open reading frame independent of sequence context. RNA 12: 2160-2170.
105. Lynch M, Hong X, Scofield DG (2006) NMD and the evolution of eukaryotyic gene structure. In: Maquat LE, editor. Nonsense-mediated mRNA decay. Georgetown (TX): Landes Bioscience. pp. 197-211.
106. Mendell JT, Medghalchi SM, Lake RG, Noensie EN, Dietz HC (2000) Novel Upf2p orthologues suggest a functional link between translation initiation and nonsense surveillance complexes. Mol Cell Biol 20: 8944-8957.
107. Pulak R, Anderson P (1993) Messenger-RNA surveillance by the *Caenorhabditis-elegans smg* genes. Genes Dev 7: 1885-1897.
108. Gatfield D, Unterholzner L, Ciccarelli FD, Bork P, Izaurralde E (2003) Nonsense-mediated mRNA decay in *Drosophila*: At the intersection of the yeast and mammalian pathways. Embo J 22: 3960-3970.

109. Rajavel KS, Neufeld EF (2001) Nonsense-mediated decay of human HEXA mRNA. Mol Cell Biol 21: 5512-5519.
110. Buhler M, Steiner S, Mohn F, Paillusson A, Muhlemann O (2006) EJC-independent degradation of nonsense immunoglobulin-mu mRNA depends on 3′ UTR length. Nat Struct Mol Biol 13: 462-464.
111. Singh G, Rebbapragada I, Lykke-Andersen J (2008) A competition between stimulators and antagonists of Upf complex recruitment governs human nonsense-mediated mRNA decay. PLoS Biol 6(4): e111. doi:10.1371/journal.pbio.0060111
112. Jaillon O, Bouhouche K, Gout JF, Aury JM, Noel B, et al. (2008) Translational control of intron splicing in eukaryotes. Nature 451: 359-362.
113. Pesole G, Mignone F, Gissi C, Grillo G, Licciulli F, et al. (2001) Structural and functional features of eukaryotic mRNA untranslated regions. Gene 276: 73-81.
114. Hong X, Scofield DG, Lynch M (2006) Intron size, abundance, and distribution within untranslated regions of genes. Mol Biol Evol 23: 2392-2404.
115. Roy SW, Penny D, Neafsey DE (2007) Evolutionary conservation of UTR intron boundaries in *Cryptococcus*. Mol Biol Evol 24: 1140-1148.
116. Oyama M, Itagaki C, Hata H, Suzuki Y, Izumi T, et al. (2004) Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. Genome Res 14: 2048-2052.
117. Oyama M, Kozuka-Hata H, Suzuki Y, Semba K, Yamamoto T, et al. (2007) Diversity of translation start sites may define increased complexity of the human short ORFeome. Mol Cell Proteomics 6: 1000-1006.
118. Diba F, Watson CS, Gametchu B (2001) 5′UTR sequences of the glucocorticoid receptor 1A transcript encode a peptide associated with translational regulation of the glucocorticoid receptor. J Cell Biochem 81: 149-161.
119. Crowe ML, Wang XQ, Rothnagel JA (2006) Evidence for conservation and selection of upstream open reading frames suggests probable encoding of bioactive peptides. BMC Genomics 7: 16.
120. Mignone F, Gissi C, Liuni S, Pesole G (2002) Untranslated regions of mRNAs. Genome Biol 3: REVIEWS0004.
121. Bhattacharya D, Lutzoni F, Reeb V, Simon D, Nason J, et al. (2000) Widespread occurrence of spliceosomal introns in the rDNA genes of ascomycetes. Mol Biol Evol 17: 1971-1984.
122. Takahashi Y, Urushiyama S, Tani T, Ohshima Y (1993) An mRNA-type intron is present in the *Rhodotorula hasegawae* U2 small nuclear RNA gene. Mol Cell Biol 13: 5613-5619.
123. Reinisch KM, Wolin SL (2007) Emerging themes in non-coding RNA quality control. Curr Opin Struct Biol 17: 209-214.
124. LaCava J, Houseley J, Saveanu C, Petfalski E, Thompson E, et al. (2005) RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. Cell 121: 713-724.
125. LaRiviere FJ, Cole SE, Ferullo DJ, Moore MJ (2006) A late-acting quality control process for mature eukaryotic rRNAs. Mol Cell 24: 619-626.
126. Wyers F, Rougemaille M, Badis G, Rousselle JC, Dufour ME, et al. (2005) Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. Cell 121: 725-737.
127. Brosius J (1991) Retroposons—Seeds of evolution. Science 251: 753.
128. Chang DD, Sharp PA (1989) Regulation by HIV Rev depends upon recognition of splice sites. Cell 59: 789-795.
129. Legrain P, Rosbash M (1989) Some cis- and trans-acting mutants for splicing target pre-mRNA to the cytoplasm. Cell 57: 573-583.
130. Bray M, Prasad S, Dubay JW, Hunter E, Jeang KT, et al. (1994) A small element from the Mason-Pfizer monkey virus genome makes human immunodeficiency virus type 1 expression and replication Rev-independent. Proc Natl Acad Sci U S A 91: 1256-1260.
131. Huang Y, Carmichael GG (1997) The mouse histone H2a gene contains a small element that facilitates cytoplasmic accumulation of intronless gene transcripts and of unspliced HIV-1-related mRNAs. Proc Natl Acad Sci U S A 94: 10104-10109.
132. Tang H, Gaietta GM, Fischer WH, Ellisman MH, Wong-Staal F (1997) A cellular cofactor for the constitutive transport element of type D retrovirus. Science 276: 1412-1415.
133. Guang S, Mertz JE (2005) Pre-mRNA processing enhancer (PPE) elements from intronless genes play additional roles in mRNA biogenesis than do ones from intron-containing genes. Nucleic Acids Res 33: 2215-2226.
134. Huang Y, Wimler KM, Carmichael GG (1999) Intronless mRNA transport elements may affect multiple steps of pre-mRNA processing. Embo J 18: 1642-1652.
135. Liu X, Mertz JE (1995) HnRNP L binds a cis-acting RNA sequence element that enables intron-dependent gene expression. Genes Dev 9: 1766-1780.
136. Lynch M (2002) Intron evolution as a population-genetic process. Proc Natl Acad Sci U S A 99: 6118-6123.
137. Lev-Maor G, Sorek R, Levanon EY, Paz N, Eisenberg E, et al. (2007) RNA-editing-mediated exon evolution. Genome Biol 8: R29.