Current Biology, Vol. 14, R351–R352, May 4, 2004, ©2004 Elsevier Ltd. All rights reserved. DOI 10.1016/j.cub.2004.04.024

Molecular Evolution: Introns Fall into Place

Dispatch

Arlin Stoltzfus

The evolutionary origin of spliceosomal introns remains elusive. The startling success of a new way of predicting intron sites suggests that the splicing machinery determines where introns are added to genes.

Two major issues dominate the debate over the origin and evolution of the spliceosomal introns of eukaryotic protein-coding genes: the time —'early' versus 'late'— at which introns became a prominent feature of gene organization, and the extent to which the putative benefits of introns justify their existence and account for their features. Before 1990, it had become a textbook dogma that introns were retained from a primordial RNA world in order to speed evolution by exon shuffling [1]. Serious disagreement in the literature erupted in 1990, with a prominent dispute over claims of pervasive exon shuffling [2], followed by disputes over the correspondence of exons to protein domains, ancient shared introns, intron 'sliding', and more recently, introns in organelle-derived nuclear genes [3] and the possibility of preferred sites for intron gain [4].

Predicting Intron Sites from Splicing Experiments

Work by Sadusky *et al.* [5], published recently in *Current Biology*, promises to stimulate a closer look at the pattern and process of intron gain. Remarkably, these authors devised an experimental method to predict where introns will be found in a gene family (Figure 1). For each of ten introns in a set of three actin genes (from human, *Arabidopsis*, and *Physarum*), the authors knocked out the original 'donor' splice junction at the 5'-end of the intron and examined the *in vivo* splicing of the resulting mutant gene. Notably, the splicing machinery finds new ways to splice the mutant transcripts, often making use of a cryptic splice junction (Figure 1). Seven of the nine junctions identified in this manner coincide in location with a known intron in one of the other actin genes. The probability of this being a coincidence is 1 in 35 million. Furthermore, as the authors relied on a slightly out of date compilation of intron data, there is an 8th match: the cryptic splice site at position 17–1 matches an intron site in a *Pneumocystis carinii* actin gene [6]. Thus, the revised probability of coincidence is an astounding 1 in 3 billion.

Sites of Intron Gain or Sites of Intron Loss?

For the cryptic splice sites identified (seven donor and two acceptor sites), the preferred nucleotides are $A_5G_8|G_8T_6$, where the subscript indicates how many times the nucleotide appears and the '|' indicates the exon-intron or intron-exon boundary. This same AG|GT motif also represents the most frequent nucleotides in naturally occurring exon-intron (donor) and intron-exon (acceptor) junctions in most eukaryotic organisms. While these nucleotide preferences are weak on the exonic side, they are strong on the intronic side, such that 95% of all spliceosomal introns begin with GT and end with AG. In addition, results from several gene families [5,7] suggest that AG|GT is a preferred context for intron gain. Indeed, Dibb and Newman were the first to present a substantive case for this interpretation [7].

Thus, the method works because actin genes are sufficiently similar that finding AGGT-like cryptic splice sites in one gene is effectively a method of predicting AGGT-like intron gain sites in the gene family as a whole. One might object that this 'intron gain site' interpretation ignores the possibility of intron loss. The AG|GT sites that are occupied by an intron in one gene, but not in another, might be sites of intron gain in the first gene, or sites of intron loss in the second. This 'intron loss site' interpretation, though intuitively unappealing, is nonetheless difficult to exclude without systematically analyzing a probabilistic model of the inheritance, loss and gain of introns.

Exactly this kind of analysis has been done recently [8]. The results reveal that most introns are gained, with gain favoring 'mAGGt' contexts ('m' can be A or C, and the upper-case letter indicates a stronger preference). Taking preferential intron gain as a given, we may ask what its causes and consequences are. Below I will focus on the causes of preferential gain and ignore the consequences, some of which are addressed elsewhere [4,9].

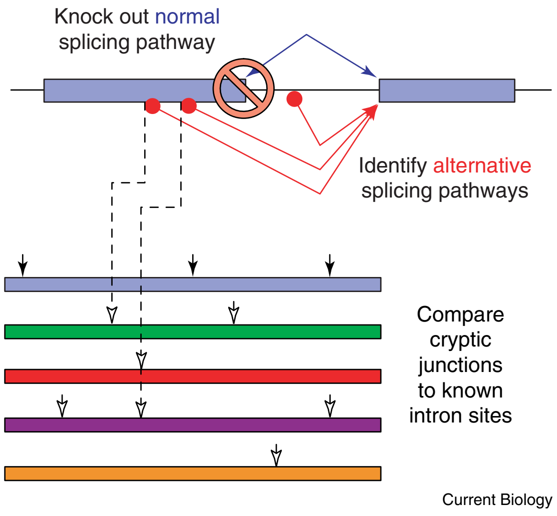
Knock out normal splicing pathway
Identify alternative splicing pathways
Compare cryptic junctions to known intron sites
Current Biology

Figure 1.
Schematic diagram of the experiments of Sadusky *et al.* [5].

Center for Advanced Research in Biotechnology (CARB), 9600 Gudelsky Drive, Rockville, Maryland 20850, USA.
E-mail: arlin@carb.nist.gov

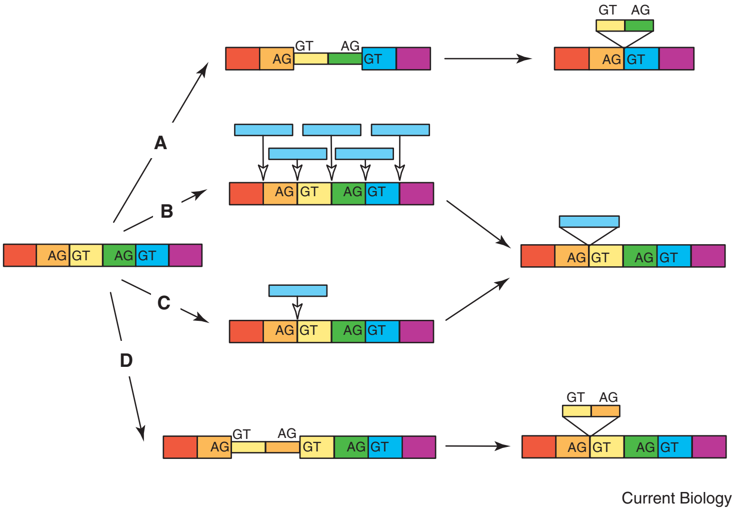

Figure 2. Models of intron gain at AG|GT sites.

An unsplit gene is shown at the left, divided into colored blocks to highlight preferred and non-preferred sites.

(A) In the 'preferential reassignment model', an exonic GT…AG sequence in the preferred context is recognized by the splicing machinery as an intron. (B) In the 'preferential retention model', natural selection eliminates randomly inserted introns that lack the favorable context. (C) In the 'targeted insertion model', introns insert preferentially at AG|GT sites. Models (B) and (C) generate a pattern in which exogenous sequences are inserted at preferred sites. (D) Duplication of an AGGT-containing exonic segment creates a distinctive pattern in which introns added in preferred contexts are duplicate copies of flanking sequences.

Mechanisms of Targeted Intron Gain

Four types of models of intron gain are consistent with a preference for AG|GT-like contexts (Figure 2), given the supposition that this context tends to enhance splicing. Only one type of model can be excluded. The 'reassignment model' (Figure 2, model A) implies that the exonic parts of a gene shrink as the gene accumulates introns, which is not the case: intron number increases with the length of exons [10,11]. In principle, computer searches of introns against a genome sequence could uncover homologies that implicate either the 'endogenous duplication model' of Dibb [12] (Figure 2D) or the insertional models (Figure 2B,C). So far, such searches have failed [13]. However, homology searches involving non-coding sequences are difficult, and unless a pervasive pattern is uncovered, even the apparently successful outcome of detecting the homolog of an intron is not a magic bullet that solves the evolutionary problem of its origin. If it were, the problem would be solved already, because there is an example of a clear intron gain by insertion: a variant of the *Dissociation* transposon family that carries its own splice signals inserted into the maize *sh2* gene [14]. Though this observation is beyond doubt, it does not resolve the issue of which mechanism is responsible for the great mass of introns gained in the last one or two billion years in diverse eukaryotic genomes.

Resolving this issue will require examination of other implications of more specific mechanistic hypotheses. For example, let us consider a specific case of the targeted insertion model, namely the possibility that spliceosomal intron mobility, by analogy with group II introns [15], is initiated by reverse-splicing of an intron into a novel site. If this process is completed by recombination with the parent locus of a reverse-transcribed cDNA that contains the novel intron, one expects a taxonomic difference in the rates of intron gain, as in fungi a piece of modified DNA typically recombines homologously with the parent locus, whereas in mammalian cells a piece of modified DNA typically inserts ectopically, leaving the parent locus unchanged. Furthermore, as reverse transcription is highly processive, the gain of a short intron by this mechanism is more likely than that of a long intron. If the initiating event is not a complete reversal of splicing, but only a reversal of the second step, then the target-site preferences may be subtly different from the preferences for the complete reaction. If such predictions can be tested using modern computational analyses and further experiments, the mechanism of evolutionary intron gain may be, at long last, within our grasp.

References

1. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J.D. (1994). Molecular Biology of the Cell. (3rd edn) (Garland Science Publishing).
2. Doolittle, R.F. (1991). Counting and discounting the universe of exons. Science *253*, 677–679.
3. Wolf, Y.I., Kondrashov, F.A., and Koonin, E.V. (2001). Footprints of primordial introns on the eukaryotic genome: still no clear traces. Trends Genet. *17*, 499–501.
4. Long, M., and Rosenberg, C. (2000). Testing the "proto-splice sites" model of intron origin: evidence from analysis of intron phase correlations. Mol. Biol. Evol. *17*, 1789–1796.
5. Sadusky, T., Newman, A.J., and Dibb, N.J. (2004). Exon junction sequences as cryptic splice sites: implications for intron origin. Curr. Biol. *14*, 505-509.
6. Miyahira, Y., Hiraoka, Y., Komatsu, N., Takeuchi, T., and Aiso, S. (1997). Genomic structure of the actin-encoding gene of *Pneumocystis carinii*. Parasitol. Int. *46*, 289–295.
7. Dibb, N.J., and Newman, A.J. (1989). Evidence that introns arose at proto-splice sites. EMBO J. *8*, 2015–2021.
8. Qiu, W.G., Schisler, N.J., and Stoltzfus, A. (2004). Spliceosomal intron gain: sequence and phase preferences. Mol. Biol. Evol., in press.
9. Hickey, D.A., Benkel, B.F., and Abukashawa, S.M. (1989). A general model for the evolution of nuclear pre-mRNA introns. J. Theor. Biol. *137*, 41–53.
10. Blake, C. (1983). Exons— present from the beginning? Nature *306*, 535–537.
11. Lynch, M., and Kewalramani, A. (2003). Messenger RNA surveillance and the evolutionary proliferation of introns. Mol. Biol. Evol. *20*, 563–571.
12. Dibb, N.J. (1989). Proto-splice site model of intron origin. J. Theor. Biol. *151*, 405-416.
13. Fedorov, A., Roy, S., Fedorova, L., and Gilbert, W. (2003). Mystery of intron gain. Genome Res. *13*, 2236–2241.
14. Giroux, M.J., Clancy, M., Baier, J., Ingham, L., McCarty, D., and Hannah, L.C. (1994). De novo synthesis of an intron by the maize transposable element Dissociation. Proc. Natl. Acad. Sci. USA *91*, 12150–12154.
15. Dickson, L., Huang, H.R., Liu, L., Matsuura, M., Lambowitz, A.M., and Perlman, P.S. (2001). Retrotransposition of a yeast group II intron occurs by reverse splicing directly into ectopic DNA sites. Proc. Natl. Acad. Sci. USA *98*, 13207–13212.