The effect of nonsense codons on splicing: A genomic analysis

XIANG ZHANG, JAMES LEE, and LAWRENCE A. CHASIN
Department of Biological Sciences, Columbia University, New York, NY 10027, USA

ABSTRACT

The phenomenon of nonsense-associated altered splicing raises the possibility that the recognition of in-frame nonsense codons is used generally for exon identification during pre-mRNA splicing. However, nonsense codon frequencies in pseudo exons and in regions flanking 5′ splice sites are no greater than that expected by chance, arguing against the widespread use of this strategy as a means of rejecting potential splice sites.

Keywords: NAS; nonsense; pre-mRNA splicing; exon definition; latent splice sites; pseudo exons; exon skipping

Two recent papers have bolstered the idea that the translatability of an exon can influence its splicing. Wang et al. (2002) have shown that an exon in the T-cell receptor transcript is not only excluded when it contains a nonsense codon, but that this event is accompanied by the inclusion of an overlapping cryptic alternative exon. The new exon restores an open reading frame. Independently, Li et al. (2002) have shown that a latent 5′ splice site downstream of an exon can be activated if all in-frame nonsense codons are removed from the region between it and the upstream exon. In these mutated pre-mRNA molecules, a new enlarged exon is chosen for splicing, based on its newfound extended translatability. Explanations offered for these two results include nuclear recognition of the translatability of an exon before it is spliced.

The recognition of exons or introns during the splicing process cannot rest on the splice site sequences alone, because similar sequences abound in large pre-mRNA molecules (Senapathy et al. 1990). There is considerable evidence that it is the exon that is the initial element of recognition, providing a size constraint for choosing real splice sites. However, if we define "pseudo exons" as intronic regions of typical exon size (50–250 nt) bounded by sequences that closely match the consensuses for 3′ and 5′ splice sites, their abundance still outweighs that of real exons by an order of magnitude (Sun and Chasin 2000). Splicing enhancers, assayed mostly in the context of alternatively spliced exons, are also quite varied and degenerate, and candidate sequences are easily found in pseudo exons (data not shown). How then does the cell distinguish real exons from pseudo exons? Exon translatability could be providing the missing information.

To test this idea, we examined the predicted translatability of pseudo exons. If pseudo exons are being generally discriminated against because they contain nonsense codons, then each should contain at least one in-frame nonsense codon. If, on the other hand, nonsense codons play no such role, then nonsense codons should occur at a frequency dictated simply by chance. We culled pseudo exons from a human intron–exon database (Saxonov et al. 2000) after eliminating redundant and predicted genes. We defined these pseudo exons as intronic sequences that have at the upstream end a pseudo splice site with a consensus 3′ matrix score of 78 and at the downstream end a pseudo splice site with a consensus 5′ matrix score of 75; using these criteria, 75% of real exons are captured. In addition, pseudo exons had to be far from real exons (>400 nt), not overlap, be free of highly repeated sequences, and not resemble any sequence found in the human EST database. Only exons of a length of less than 110 nt were considered; this size limitation was imposed because the chance occurrence of a nonsense codon approaches 100% for longer sequences. In-frame nonsense codons were counted in these pseudo exons, the reading frame being defined by that of the upstream real exon. Out-of-frame nonsense codons were also tallied for comparison. We calculated the proportion of pseudo exons expected to have at least one in-frame stop codon by chance based on the overall frequency of stop

Reprint requests to: Lawrence Chasin, Columbia University, 912 Fairchild–Mail Code 2433, 1212 Amsterdam Avenue, New York, NY 10027, USA; e-mail: lac2@columbia.edu.

Article and publication are at http://www.rnajournal.org/cgi/doi/10.1261/rna.5060403.

triplets in repeat-free intron sequences (0.049 per triplet, or slightly more than 3/64 = 0.047). The results for these 2850 pseudo exons are shown as a function of pseudo exon size in Figure 1A. The frequency of in-frame nonsense codons was not different from that expected by chance alone (indicated by the line in Fig. 1A). For instance, for pseudo exons up to length 70, the proportion with at least one in-frame nonsense codon was 0.46, that expected by chance is 0.44 ($p = 0.11$), and that expected from the translatability hypothesis is 1.0 ($p < 10^{-10}$ even taking the expectation to be 0.75 rather than 1.0). We conclude that nonsense codons do not play a role in the cellular exclusion of these pseudo exons.

Similarly, we tested the idea that nonsense codons are interposed between real 5′ splice sites and downstream latent (pseudo) 5′ splice sites so as to subvert the use of the downstream site. If this were a general mechanism for 5′ splice site definition, then the proportion of such intron regions containing at least one in-frame nonsense codon should be 1.0. If not, nonsense codons should occur simply by chance. To calculate the proportion based on chance, we considered the intron flank to be made up of two components: the real splice site itself, which has a high probability of harboring a nonsense codon due to the consensus (C or A)AG/GURAGU, which carries a nonsense URA triplet at positions +2 to +4; and the remaining sequence. For the former, we used the weight matrix of real sites to calculate the probability of finding a stop codon. For the latter, we used the overall frequency of stop codons in the downstream 100-nt flanks of the real exons in our database. This frequency was 0.038 per triplet, which is less than the 3/64 (0.047) expected on a random basis, reflecting the nonrandom nature of sequences flanking exons (Nussinov 1989; Engelbrecht et al. 1992). We analyzed 1164 real 5′ splice sites situated ≤100 nt upstream of a latent 5′ splice site that had a consensus matrix score greater than the median of real sites. The occurrence of nonsense codons as a function of the distance between the splice site and the downstream latent 5′ splice site is shown in Figure 1B. The proportion of intervening intron sequences having at least one in-frame nonsense codon was not greater than that expected by chance (the line in Fig. 1B), and considerably less than the 1.0 predicted by a translatability hypothesis. For a statistical treatment of this data we considered the 598 sequences with false sites within 50 nt in the downstream flanks. Of this set, 0.42 have at least one stop codon between the false sites and the 5′ splice sites (0.18 within the splice site and 0.24 in the region beyond). The corresponding expectation for the total based on chance is 0.43. The values for the totals are not statistically different ($p = 0.71$). In contrast, the observed value of 0.42 is different from the expectation based on a nonsense requirement to inactivate the latent splice site of 1.0 ($p < 10^{-10}$ even taking the expectation to be 0.75 rather than 1.0). We conclude that nonsense codons do not act in a general way to stifle potential competition by a neighboring latent 5′ splice site.

Miriami et al. (2002) also performed a statistical test of the idea that stop codons are associated with latent 5′ splice sites. They found support for this proposition in observing that the density of in-frame stop codons was significantly higher in introns bearing latent splices sites than in those devoid of such sequences. Using the same criteria as these authors for latent 5′ splice sites and using our database, we obtained the same difference in densities of in-frame stop

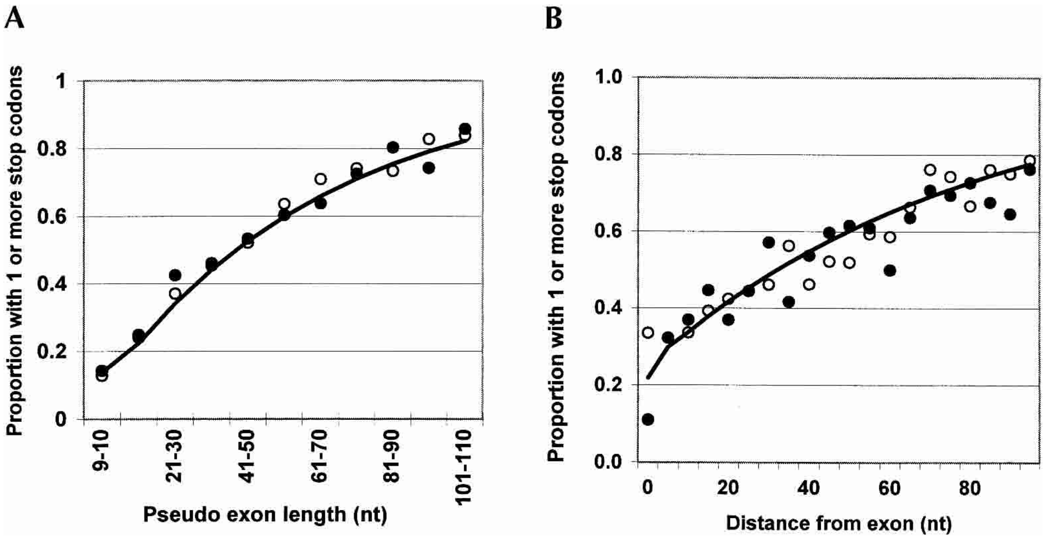

FIGURE 1. The occurrence of in-frame (solid circles) nonsense codons in intron regions. The average for the two types of out-of-frame nonsense codons are included for comparison (open circles). The line represents the result expected by chance alone for an in-frame nonsense codon, using a Poisson distribution. (*A*) Pseudo exons; windows of 10 were grouped. (*B*) Flanks downstream from exons; windows of 5 were grouped; the starting point of each window is plotted. Lists of the sequences underlying this data can be found at www.columbia.edu/data/cu/biology/faculty/chasin/rnajournal.

codons. However, selection for introns lacking a latent splice site (or any particular sequence) necessarily selects for small introns, and small introns are known to be GC-rich (Lander et al. 2001). We found the average size and GC content of all introns in our database to be 1843 and 49%, respectively, whereas the corresponding values for introns without latent sites were 370 and 55%. Since stop codons are AT-rich, their density would be expected to be lower in GC-rich introns. Moreover, there was no difference in the densities of in-frame and out-of-frame stop codons in either set. We believe our approach, leading to the opposite conclusion, represents a more direct test of the hypothesis in question.

These conclusions are in agreement with several genetic studies in which nonsense mutations that did not affect splicing were described, for example, in the genes for dhfr (Urlaub et al. 1989), aprt (Kessler and Chasin 1996), or hprt (Valentine 1998). Similarly, no nonsense mutations were found among 70 mutants of a three-exon dhfr minigene selected for skipping of the central exon (Chen and Chasin 1993).

The work that instigated this analysis (Li et al. 2002; Wang et al. 2002) has provided impressive evidence that translatability can affect splicing decisions. Nuclear recognition of translatability presumably involves a complex mechanism. Our analysis now adds the question of why such complexity would emerge if it were not to be used generally.

ACKNOWLEDGMENTS

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

Received January 13, 2003; accepted February 20, 2003.

REFERENCES

Chen, I.T. and Chasin, L.A. 1993. Direct selection for mutations affecting specific splice sites in a hamster dihydrofolate reductase minigene. *Mol. Cell. Biol.* **13:** 289–300.

Engelbrecht, J., Knudsen, S., and Brunak, S. 1992. G + C-rich tract in 5′ end of human introns. *J. Mol. Biol.* **227:** 108–113.

Kessler, O. and Chasin, L.A. 1996. Effects of nonsense mutations on nuclear and cytoplasmic adenine phosphoribosyltransferase RNA. *Mol. Cell. Biol.* **16:** 4426–4435.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Li, B., Wachtel, C., Miriami, E., Yahalom, G., Friedlander, G., Sharon, G., Sperling, R., and Sperling, J. 2002. Stop codons affect 5′ splice site selection by surveillance of splicing. *Proc. Natl. Acad. Sci.* **99:** 5277–5282.

Miriami, E., Motro, U., Sperling, J., and Sperling, R. 2002. Conservation of an open-reading frame as an element affecting 5′ splice site selection. *J. Struct. Biol.* **140:** 116–122.

Nussinov, R. 1989. Conserved signals around the 5′ splice sites in eukaryotic nuclear precursor mRNAs: G-runs are frequent in the introns and C in the exons near both 5′ and 3′ splice sites. *J. Biomol. Struct. Dyn.* **6:** 985–1000.

Saxonov, S., Daizadeh, I., Fedorov, A., and Gilbert, W. 2000. EID: The Exon-Intron Database—An exhaustive database of protein-coding intron-containing genes. *Nucleic Acids Res.* **28:** 185–190.

Senapathy, P., Shapiro, M.B., and Harris, N.L. 1990. Splice junctions, branch point sites, and exons: Sequence statistics, identification, and applications to genome project. *Methods Enzymol.* **183:** 252–278.

Sun, H. and Chasin, L.A. 2000. Multiple splicing defects in an intronic false exon. *Mol. Cell. Biol.* **20:** 6414–6425.

Urlaub, G., Mitchell, P.J., Ciudad, C.J., and Chasin, L.A. 1989. Nonsense mutations in the dihydrofolate reductase gene affect RNA processing. *Mol. Cell. Biol.* **9:** 2868–2880.

Valentine, C.R. 1998. The association of nonsense codons with exon skipping. *Mutat. Res.* **411:** 87–117.

Wang, J., Hamilton, J.I., Carter, M.S., Li, S., and Wilkinson, M.F. 2002. Alternatively spliced TCR mRNA induced by disruption of reading frame. *Science* **297:** 108–110.