Molecular Cell, Vol. 14, 221–231, April 23, 2004, Copyright ©2004 by Cell Press

Minimal Conditions for Exonization of Intronic Sequences: 5′ Splice Site Formation in *Alu* Exons

Rotem Sorek,[1,2,5] Galit Lev-Maor,[1,5]
Mika Reznik,[1,5] Tal Dagan,[3] Frida Belinky,[3]
Dan Graur,[4] and Gil Ast[1,*]
[1]Department of Human Genetics and
Molecular Medicine
Sackler Faculty of Medicine
Tel Aviv University
Ramat Aviv 69978
Israel
[2]Compugen
72 Pinchas Rosen Street
Tel Aviv 69512
Israel
[3]Department of Zoology
George S. Wise Faculty of Life Sciences
Tel Aviv University
Ramat Aviv 69978
Israel
[4]Department of Biology and Biochemistry
University of Houston
Houston, Texas 77204

Summary

Alu exonization, which is an evolutionary pathway that creates primate-specific transcriptomic diversity, is a powerful tool for studying alternative-splicing regulation. Through bioinformatic analyses combined with experimental methodology, we identified the mutational changes needed to create functional 5′ splice sites in *Alu*. We revealed a complex mechanism by which the sequence composition of the 5′ splice site and its base pairing with the small nuclear RNA U1 govern alternative splicing. We show that in *Alu*-derived GC introns the strength of the base pairing between U1 snRNA and the 5′ splice site controls the skipping/inclusion ratio of alternative splicing. Based on these findings, we identified 7810 *Alu*s within the human genome that are prone to exonization. Mutations in these *Alu*s may cause genetic disorders or contribute to human-specific protein diversity.

Introduction

The availability of the complete human genome sequence has made it clear that gene number is not the sole determinant of proteome complexity. Sequencing of the human genome showed that humans possess only ~26,000 protein coding genes, which is only slightly larger than the number of genes in *C. elegans*. This surprisingly low number contrasts with the number of human proteins, estimated to be more than 90,000 (Harrison et al., 2002; Lander et al., 2001).

By producing more than one type of mRNA from a single gene, alternative splicing is a significant contributor to proteome diversification (Brett et al., 2002; Graveley, 2001). Bioinformatic analyses indicate that 35%–74% of all human genes participate in alternative splicing, which contributes significantly to human proteome complexity and explains the numerical disparity between genes and proteins (Black, 2000; Johnson et al., 2003; Modrek and Lee, 2002). Alternative splicing is often regulated according to cell type, developmental stage, sex, or in response to an external stimulus (Cartegni et al., 2002; Stoilov et al., 2002). Aberrant regulation of alternative splicing has been implicated in an increasing number of human diseases, including cancer (Hastings and Krainer, 2001; Modrek and Lee, 2002; Nissim-Rafinia and Kerem, 2002; Stoilov et al., 2002; Xu and Lee, 2003).

Approximately 75–130 million years may have passed since the human and mouse common ancestor speciated into two separate lineages (Waterston et al., 2002; Yang and Yoder, 1999). Most of the genes (99%) are orthologous, and the majority of these genes (86%) share the same intron/exon arrangement as well as a high degree of conservation (88%) in homologous exon sequences (Waterston et al., 2002). If most of the genes are highly conserved between human and mouse, what are the genomic elements that contribute to some of the unique features of humans and primates? Human-mouse comparative analysis revealed that alternative splicing is often associated with recent exon creation and/or loss (Modrek and Lee, 2003). Thus, alternative splicing has the potential of creating species-specific cassette exons.

Retrotransposons are short sequences of DNA that produce new copies of themselves by the reverse transcription of an RNA intermediate. These mobile DNA elements have had a profound influence in shaping eukaryotic genomes. As much as 46% of the human genome is made up of transposable elements, the most abundant of which is a primate-specific dimeric retrotransposon called *Alu*. The human genome contains approximately 1.4 million *Alu* copies, all derived from a single 7SL RNA-specifying gene (Brosius, 1999; Kazazian, 2000; Lander et al., 2001). *Alu* elements, which are 300 nucleotides long, currently amplify at a rate of about one insertion in every 100–200 new live births (Dewannieux et al., 2003). The transcription of *Alu* elements is usually suppressed by a large number of hypermethylated CG dinucleotides, which block a bipartite RNA polymerase III promoter (Deininger and Batzer, 2002; Makalowski et al., 1994).

We have recently shown that more than 5% of the alternatively spliced internal exons in the human genome are derived from *Alu* (Sorek et al., 2002; reviewed in Kreahling and Graveley, 2004). As far as we know, all alternatively spliced *Alu* exons (*AEx*) were created exclusively via the exonization of intronic elements. These *AEx*s enrich the transcriptome and enhance the coding capacity and regulatory versatility of primate genomes with new isoforms without compromising the integrity and original repertoire of the transcriptome and its resulting proteome. In contrast, each newly created

[*]Correspondence: gilast@post.tau.ac.il
[5]These authors contributed equally to this work.

constitutively spliced AEx will generate a new product at the expense of the original product, and such a loss may be deleterious (Sorek et al., 2002; Lev-Maor et al., 2003).

Through molecular evolutionary methodology, it was possible to align each AEx to its inferred ancestral sequence. We could, therefore, identify the changes in the Alu sequences that were most probably responsible for the exonization. This methodology has recently allowed us to identify the delicate interplay between two AG dinucleotides that maintain a weak 3′ splice site (3′ss) responsible for alternative splicing (Lev-Maor et al., 2003). We have further demonstrated that an activation of an intronic Alu sequence (silent Alu sequence never spliced in as an exon) can be induced either by point mutations in certain positions along the Alu sequence or by changing the concentration of the splicing regulatory proteins in the cell (Lev-Maor et al., 2003). Thus, Alu exonization is an evolutionary pathway that creates primate-specific genomic diversity. Here, we reveal the mutational steps that create 5′ splice site (5′ss) in alternatively spliced AExs and examine how the spliceosome selects this 5′ss.

Results

We compiled a data set of exonized Alus in which the prevalent 5′ss of these exons is selected (Figure 1, Alu position 157 being the first position of the intron) and compared their adjacent regions with homologous positions in the Alu antisense consensus sequence (Jurka and Milosavljevic, 1991). This allowed us to identify nucleotide positions along the Alu sequence that need to be changed or remain conserved in order to enable Alu exonization. Two types of 5′ss were found: introns that start with GC (Figure 1, cases 1–4) and introns that start with GT (Figure 1, cases 5–25). In the human genome, more than 98% of all introns begin with GT (Farrer et al., 2002; Lander et al., 2001; Thanaraj and Clark, 2001). The minority GC introns (0.7% of all introns) were claimed to be frequently involved in alternative splicing (Thanaraj and Clark, 2001).

The most significant change observed between exonized Alus and their ancestral sequences was at position 156 (position two of the intron), where a mutation from C to T generates a canonical GT 5′ss at positions 157-156. This occurred in 21 of the 25 (84%) exonized Alus. In those cases in which positions 157-156 remained GC as in the ancestral sequence (rows 1–4), position 155 (representing position 3 of the intron) was found to be mutated from G to A.

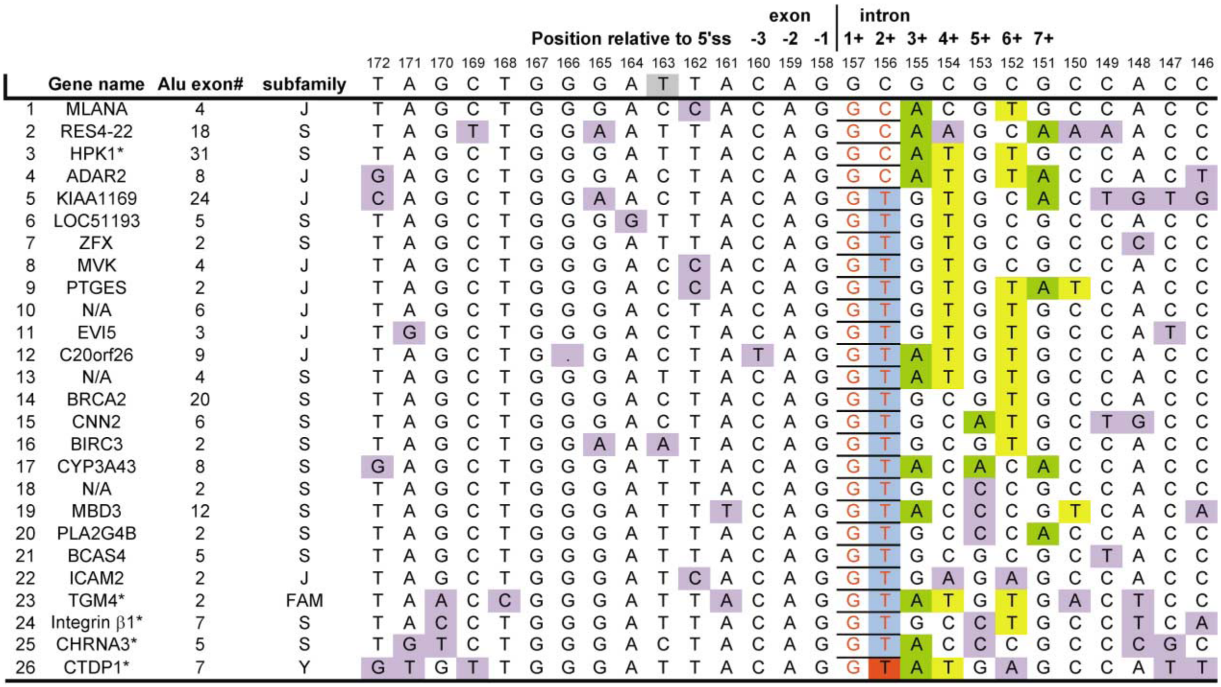
In the ancestral sequence, CG dinucleotides are found in positions 156-155, 154-153, and 152-151 (Figure 1, upper row). Since CG dinucleotides in Alu are frequently hypermethylated (Kunkel and Diaz, 2002) and mutate 9.2 times more frequently than non-CG positions (Batzer et al., 1990; Kunkel and Diaz, 2002), one may attribute the formation of the GT 5′ss to the propensity of these sites to mutate from CG to TG. The prevalent mutations in positions 156, 154, and 152 are from C to T (yellow and blue), whereas in positions 155 and 151 the prevalent mutations are from G to A (green).

Some of the changes observed in Figure 1 seem to be random (purple); others are presumably due to CG substitutions (yellow, blue, and green). However, it is unclear which of these changes are important to the exonization of Alus and which represent inconsequential intronic substitutions. To pinpoint the positions that are important for exonization, we compiled a data set of 166,276 full-length intronic Alus that are found in the antisense orientation in introns of human genes and compared them to exonized Alus (Dagan et al., 2004) (see also Experimental Procedures). By aligning each of these Alus to its ancestral sequence, we examined the percentage of each of the four nucleotides in each position along the Alu sequence (Figure 2A; positions 172 to 146 of the antisense Alu consensus sequence are shown). We similarly examined the 25 AExs shown in Figure 1 (Figure 2B). Finally, we compared the nucleotide distribution for each position of the intronic Alus to those of the AExs and looked for statistically significant deviations between the distributions (Figure 2B).

Remarkably, in only two positions along the entire exonized-Alus sequence did we find a distribution significantly different from that in the homologous positions of the intronic Alus. In position 156 (position two of the intron) C changes predominantly to T to create the aforementioned GT 5′ss; in position 153 (position 5 of the intron) G is conserved over the expected frequency (G was found in 18 instead of 10.9 expected sequences). Since sense- and antisense-oriented Alus can be under different selective pressures in the human genome, we also compiled a data set of 136,151 intronic Alus that are found in the sense orientation in introns and repeated the statistical comparison to exonized Alus. This yielded the same results, indicating positions two and five of the intron as significantly different between exonized and nonexonized Alus (Figures 2C and 2D). This indicates that these are the most important positions in the creation of functional Alu-derived 5′ss. Therefore, Alus in the antisense orientation in introns are in fact "preexons" whose exonization requires only a small number of mutations.

Significantly, the same mutation, from C to T at position 156, of an antisense Alu sequence in intron six of the CTDP1 gene was found to be the cause of CCFDN (congenital cataracts, facial dysmorphism, and neuropathy) syndrome (Figure 1, row 26; Varon et al., 2003). In this gene, the G in position 153 is indeed conserved as predicted by our analysis. Previously, only mutations leading to the constitutive exonization of Alu elements were known to be deleterious, e.g., in Alport-syndrome type I, Sly syndrome, and ornithine aminotransferase (OAT) deficiency (Vervoort et al., 1998; Mitchell et al., 1991; Knebelmann et al., 1995). CCFDN syndrome was the first reported case in which a mutation leading to the creation of an alternatively spliced AEx resulted in a genetic disease. Thus, the appearance of an aberrant Alu-containing spliced form may result in genetic disease even when the normal mRNA continues to be synthesized.

Following these results we wished to understand how the alternative splicing of these exonized Alus is regulated. In mRNA splicing, the 5′ss is recognized by three small nuclear RNPs (complexes of snRNA and proteins), one of which (U1) base pairs across the 5′ss junction (potential base pairing between positions −3 to +6;



																		exon			intron									
																	Position relative to 5'ss	-3	-2	-1	1+	2+	3+	4+	5+	6+	7+			
				172	171	170	169	168	167	166	165	164	163	162	161	160	159	158	157	156	155	154	153	152	151	150	149	148	147	146
	Gene name	Alu exon#	subfamily	T	A	G	C	T	G	G	G	A	T	T	A	C	A	G	G	C	G	C	G	C	G	C	C	A	C	C
1	MLANA	4	J	T	A	G	C	T	G	G	G	A	C	C	A	C	A	G	G	C	A	C	G	T	G	C	C	A	C	C
2	RES4-22	18	S	T	A	G	T	T	G	G	A	A	T	T	A	C	A	G	G	C	A	A	G	C	A	A	A	A	C	C
3	HPK1*	31	S	T	A	G	C	T	G	G	G	A	T	T	A	C	A	G	G	C	A	T	G	T	G	C	C	A	C	C
4	ADAR2	8	J	G	A	G	C	T	G	G	G	A	C	T	A	C	A	G	G	C	A	T	G	T	A	C	C	A	C	T
5	KIAA1169	24	J	C	A	G	C	T	G	G	A	A	C	T	A	C	A	G	G	T	G	T	G	C	A	C	T	G	T	G
6	LOC51193	5	S	T	A	G	C	T	G	G	G	G	T	T	A	C	A	G	G	T	G	T	G	C	G	C	C	A	C	C
7	ZFX	2	S	T	A	G	C	T	G	G	G	A	T	T	A	C	A	G	G	T	G	T	G	C	G	C	C	C	C	C
8	MVK	4	J	T	A	G	C	T	G	G	G	A	C	C	A	C	A	G	G	T	G	T	G	C	G	C	C	A	C	C
9	PTGES	2	J	T	A	G	C	T	G	G	G	A	C	C	A	C	A	G	G	T	G	T	G	T	A	T	C	A	C	C
10	N/A	6	J	T	A	G	C	T	G	G	G	A	C	T	A	C	A	G	G	T	G	T	G	T	G	C	C	A	C	C
11	EVI5	3	J	T	G	G	C	T	G	G	G	A	C	T	A	C	A	G	G	T	G	T	G	T	G	C	C	A	T	C
12	C20orf26	9	J	T	A	G	C	T	G	.	G	A	C	T	A	T	A	G	G	T	A	T	G	T	G	C	C	A	C	C
13	N/A	4	S	T	A	G	C	T	G	G	G	A	T	T	A	C	A	G	G	T	A	T	G	T	G	C	C	A	C	C
14	BRCA2	20	S	T	A	G	C	T	G	G	G	A	C	T	A	C	A	G	G	T	G	C	G	T	G	C	C	A	C	C
15	CNN2	6	S	T	A	G	C	T	G	G	G	A	C	T	A	C	A	G	G	T	G	C	A	T	G	C	T	G	C	C
16	BIRC3	2	S	T	A	G	C	T	G	G	A	A	A	T	A	C	A	G	G	T	G	C	G	T	G	C	C	A	C	C
17	CYP3A43	8	S	G	A	G	C	T	G	G	G	A	T	T	A	C	A	G	G	T	A	C	A	C	A	C	C	A	C	C
18	N/A	2	S	T	A	G	C	T	G	G	G	A	T	T	A	C	A	G	G	T	G	C	C	C	G	C	C	A	C	C
19	MBD3	12	S	T	A	G	C	T	G	G	G	A	T	T	T	C	A	G	G	T	A	C	C	C	G	T	C	A	C	A
20	PLA2G4B	2	S	T	A	G	C	T	G	G	G	A	T	T	A	C	A	G	G	T	G	C	C	C	A	C	C	A	C	C
21	BCAS4	5	S	T	A	G	C	T	G	G	G	A	T	T	A	C	A	G	G	T	G	C	G	C	G	C	T	A	C	C
22	ICAM2	2	J	T	A	G	C	T	G	G	G	A	T	C	A	C	A	G	G	T	G	A	G	A	G	C	C	A	C	C
23	TGM4*	2	FAM	T	A	A	C	C	G	G	G	A	T	T	A	C	A	G	G	T	A	T	G	T	G	A	C	T	C	C
24	Integrin β1*	7	S	T	A	C	C	T	G	G	G	A	T	T	A	C	A	G	G	T	G	C	C	T	G	C	C	T	C	A
25	CHRNA3*	5	S	T	G	T	C	T	G	G	G	A	C	T	A	C	A	G	G	T	A	C	C	C	G	C	C	C	G	C
26	CTDP1*	7	Y	G	T	G	T	T	G	G	G	A	T	T	A	C	A	G	G	T	A	T	G	A	G	C	C	A	T	T

Figure 1. Selection of 5′ss of Alu-Derived Exons

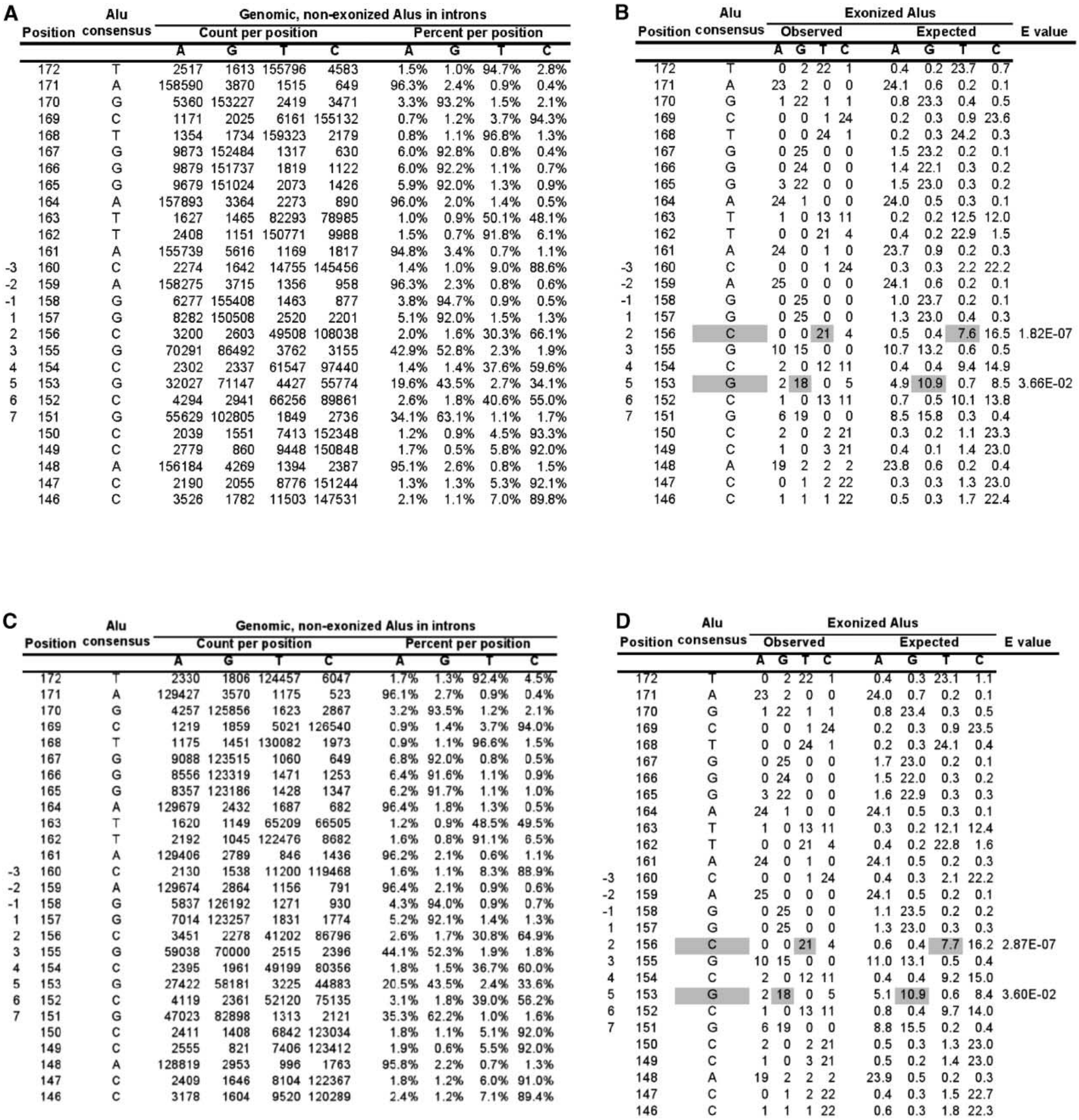
Alignment is shown for the region near the most prevalent 5′ss on the right arm of exonized Alu sequences (in the antisense orientation). Data for 26 exonized Alus (Dubbink et al., 1998; Hu et al., 1996; Mihovilovic et al., 1993; Miller and Zeller, 1997; Sorek et al., 2002; Svineng et al., 1998) are shown. The 27 nucleotides spanning positions 146–172 according to the numbering in Jurka and Milosavljevic (1991) are shown. The dinucleotides GT or GC that are selected as the 5′ss (defining the beginning of the intron) are in red. The 5′ss were inferred by alignment of expressed sequences to the human genome (Sorek et al., 2002) (Supplemental Table S1 available on Molecular Cell's website). The Alu consensus sequence appears in the first row; the position differing between Alu subfamilies S and J is marked in gray. Nucleotides that differ from the Alu consensus sequence are marked in purple; mutations that changed the dinucleotide CG to TG are marked in yellow; those for which this mutation creates a splice site are marked in blue; mutations that changed CG to CA are marked in green. Row 26 represents the 5′ss of an antisense Alu sequence in intron six of CTDP1 gene in which a mutation from C to T at position 156 resulted in CCFDN syndrome (Varon et al., 2003). This mutation (marked red) led to the exonization of an intronic Alu sequence by activation of the 5′ss and the creation of an alternatively spliced AEx. Numbers on the top mark positions relative to the 5′ss. Gene names are as in RefSeq convention. The AEx number is the serial number of the Alu-containing exon in the related gene, and the Alu subfamily type was inferred with the use of RepeatMasker (http://www.repeatmasker.org/).

for clarity, "−" and "+" indicate positions upstream or downstream of the 5′ss, respectively). This base pairing is a prerequisite step for splicing in most introns (Brow, 2002). Although the importance of the U1 snRNA:5′ss base pairing is well established in constitutive splicing, the function of this base pairing in alternative splicing is only partially understood (Cohen et al., 1993). We, therefore, set to understand the manner in which U1 affects the alternative splicing of AExs.

We used a minigene containing the genomic sequence of adenosine deaminase gene, ADAR2, from exon seven to nine, in which exon eight is an alternatively spliced AEx (Lev-Maor et al., 2003). The 5′ss of this AEx is of the GC type (Figure 1, row 4). To examine the function of the base pairing between U1 and the 5′ss in alternative splicing, we transfected 293T cells with the ADAR2 minigene containing mutations in the 5′ss and complemented these mutations with cotrasfected U1 gene containing the appropriate compensatory mutation. Therefore, the cells contained exogenous and endogenous U1s that competed with each other to bind the 5′ss (Zhuang and Weiner, 1986). Following transfection, cytoplasmic RNA was collected, and the splicing pattern of the ADAR2 minigene was examined by RT-PCR (see Experimental Procedures). We tested the effect of serial mutations on the splicing of the ADAR2 minigene when the first nucleotides of intron eight are GC (Figure 3A, representing GC 5′ss) or when the C is mutated to T creating GT (Figure 3B, representing GT 5′ss).

Our results indicate that the GC 5′ss maintains alternative rather than constitutive splicing directly because the C in position two of the 5′ss unpairs with U1. This conclusion is supported by the fact that a compensatory mutation in U1 (A to G in position seven), which restores the base pairing with position two, results in the constitutive splicing of the exon (Figure 3A, lanes 1–3; see Figure 3C for wild-type [wt] U1:5′ss base pairing).

As expected from our bioinformatic analysis (Figure 1, rows 1–4), when the 5′ss is of the GC type, an A in position three is essential for its proper selection; mutations to C, T, or G led to AEx skipping (Figure 3A, lanes 5, 6, and 8). This finding is in agreement with the weight metric of GC 5′ss where A is the prevalent nucleotide in position three (Thanaraj and Clark, 2001) and may suggest that an A at position three of the GC intron that forms A:Ψ pairing (Ψ, pseudo-uridine) with U1 is required to avoid two consecutive positions that



A

	Position	Alu consensus	Genomic, non-exonized Alus in introns							
			Count per position				Percent per position			
			A	G	T	C	A	G	T	C
	172	T	2517	1613	155796	4583	1.5%	1.0%	94.7%	2.8%
	171	A	158590	3870	1515	649	96.3%	2.4%	0.9%	0.4%
	170	G	5360	153227	2419	3471	3.3%	93.2%	1.5%	2.1%
	169	C	1171	2025	6161	155132	0.7%	1.2%	3.7%	94.3%
	168	T	1354	1734	159323	2179	0.8%	1.1%	96.8%	1.3%
	167	G	9873	152484	1317	630	6.0%	92.8%	0.8%	0.4%
	166	G	9879	151737	1819	1122	6.0%	92.2%	1.1%	0.7%
	165	G	9679	151024	2073	1426	5.9%	92.0%	1.3%	0.9%
	164	A	157893	3364	2273	890	96.0%	2.0%	1.4%	0.5%
	163	T	1627	1465	82293	78985	1.0%	0.9%	50.1%	48.1%
	162	T	2408	1151	150771	9988	1.5%	0.7%	91.8%	6.1%
	161	A	155739	5616	1169	1817	94.8%	3.4%	0.7%	1.1%
-3	160	C	2274	1642	14755	145456	1.4%	1.0%	9.0%	88.6%
-2	159	A	158275	3715	1356	958	96.3%	2.3%	0.8%	0.6%
-1	158	G	6277	155408	1463	877	3.8%	94.7%	0.9%	0.5%
1	157	G	8282	150508	2520	2201	5.1%	92.0%	1.5%	1.3%
2	156	C	3200	2603	49508	108038	2.0%	1.6%	30.3%	66.1%
3	155	G	70291	86492	3762	3155	42.9%	52.8%	2.3%	1.9%
4	154	C	2302	2337	61547	97440	1.4%	1.4%	37.6%	59.6%
5	153	G	32027	71147	4427	55774	19.6%	43.5%	2.7%	34.1%
6	152	C	4294	2941	66256	89861	2.6%	1.8%	40.6%	55.0%
7	151	G	55629	102805	1849	2736	34.1%	63.1%	1.1%	1.7%
	150	C	2039	1551	7413	152348	1.2%	0.9%	4.5%	93.3%
	149	C	2779	860	9448	150848	1.7%	0.5%	5.8%	92.0%
	148	A	156184	4269	1394	2387	95.1%	2.6%	0.8%	1.5%
	147	C	2190	2055	8776	151244	1.3%	1.3%	5.3%	92.1%
	146	C	3526	1782	11503	147531	2.1%	1.1%	7.0%	89.8%

B

	Position	Alu consensus	Exonized Alus								E value
			Observed				Expected				
			A	G	T	C	A	G	T	C	
	172	T	0	2	22	1	0.4	0.2	23.7	0.7	
	171	A	23	2	0	0	24.1	0.6	0.2	0.1	
	170	G	1	22	1	1	0.8	23.3	0.4	0.5	
	169	C	0	0	1	24	0.2	0.3	0.9	23.6	
	168	T	0	0	24	1	0.2	0.3	24.2	0.3	
	167	G	0	25	0	0	1.5	23.2	0.2	0.1	
	166	G	0	24	0	0	1.4	22.1	0.3	0.2	
	165	G	3	22	0	0	1.5	23.0	0.3	0.2	
	164	A	24	1	0	0	24.0	0.5	0.3	0.1	
	163	T	1	0	13	11	0.2	0.2	12.5	12.0	
	162	T	0	0	21	4	0.4	0.2	22.9	1.5	
	161	A	24	0	1	0	23.7	0.9	0.2	0.3	
-3	160	C	0	0	1	24	0.3	0.3	2.2	22.2	
-2	159	A	25	0	0	0	24.1	0.6	0.2	0.1	
-1	158	G	0	25	0	0	1.0	23.7	0.2	0.1	
1	157	G	0	25	0	0	1.3	23.0	0.4	0.3	
2	156	C	0	0	21	4	0.5	0.4	7.6	16.5	1.82E-07
3	155	G	10	15	0	0	10.7	13.2	0.6	0.5	
4	154	C	2	0	12	11	0.4	0.4	9.4	14.9	
5	153	G	2	18	0	5	4.9	10.9	0.7	8.5	3.66E-02
6	152	C	1	0	13	11	0.7	0.5	10.1	13.8	
7	151	G	6	19	0	0	8.5	15.8	0.3	0.4	
	150	C	2	0	2	21	0.3	0.2	1.1	23.3	
	149	C	1	0	3	21	0.4	0.1	1.4	23.0	
	148	A	19	2	2	2	23.8	0.6	0.2	0.4	
	147	C	0	1	2	22	0.3	0.3	1.3	23.0	
	146	C	1	1	1	22	0.5	0.3	1.7	22.4	

C

	Position	Alu consensus	Genomic, non-exonized Alus in introns							
			Count per position				Percent per position			
			A	G	T	C	A	G	T	C
	172	T	2330	1806	124457	6047	1.7%	1.3%	92.4%	4.5%
	171	A	129427	3570	1175	523	96.1%	2.7%	0.9%	0.4%
	170	G	4257	125856	1623	2867	3.2%	93.5%	1.2%	2.1%
	169	C	1219	1859	5021	126540	0.9%	1.4%	3.7%	94.0%
	168	T	1175	1451	130082	1973	0.9%	1.1%	96.6%	1.5%
	167	G	9088	123515	1060	649	6.8%	92.0%	0.8%	0.5%
	166	G	8556	123319	1471	1253	6.4%	91.6%	1.1%	0.9%
	165	G	8357	123186	1428	1347	6.2%	91.7%	1.1%	1.0%
	164	A	129679	2432	1687	682	96.4%	1.8%	1.3%	0.5%
	163	T	1620	1149	65209	66505	1.2%	0.9%	48.5%	49.5%
	162	T	2192	1045	122476	8682	1.6%	0.8%	91.1%	6.5%
	161	A	129406	2789	846	1436	96.2%	2.1%	0.6%	1.1%
-3	160	C	2130	1538	11200	119468	1.6%	1.1%	8.3%	88.9%
-2	159	A	129674	2864	1156	791	96.4%	2.1%	0.9%	0.6%
-1	158	G	5837	126192	1271	930	4.3%	94.0%	0.9%	0.7%
1	157	G	7014	123257	1831	1774	5.2%	92.1%	1.4%	1.3%
2	156	C	3451	2278	41202	86796	2.6%	1.7%	30.8%	64.9%
3	155	G	59038	70000	2515	2396	44.1%	52.3%	1.9%	1.8%
4	154	C	2395	1961	49199	80356	1.8%	1.5%	36.7%	60.0%
5	153	G	27422	58181	3225	44883	20.5%	43.5%	2.4%	33.6%
6	152	C	4119	2361	52120	75135	3.1%	1.8%	39.0%	56.2%
7	151	G	47023	82898	1313	2121	35.3%	62.2%	1.0%	1.6%
	150	C	2411	1408	6842	123034	1.8%	1.1%	5.1%	92.0%
	149	C	2555	821	7406	123412	1.9%	0.6%	5.5%	92.0%
	148	A	128819	2953	996	1763	95.8%	2.2%	0.7%	1.3%
	147	C	2409	1646	8104	122367	1.8%	1.2%	6.0%	91.0%
	146	C	3178	1604	9520	120289	2.4%	1.2%	7.1%	89.4%

D

	Position	Alu consensus	Exonized Alus								E value
			Observed				Expected				
			A	G	T	C	A	G	T	C	
	172	T	0	2	22	1	0.4	0.3	23.1	1.1	
	171	A	23	2	0	0	24.0	0.7	0.2	0.1	
	170	G	1	22	1	1	0.8	23.4	0.3	0.5	
	169	C	0	0	1	24	0.2	0.3	0.9	23.5	
	168	T	0	0	24	1	0.2	0.3	24.1	0.4	
	167	G	0	25	0	0	1.7	23.0	0.2	0.1	
	166	G	0	24	0	0	1.5	22.0	0.3	0.2	
	165	G	3	22	0	0	1.6	22.9	0.3	0.3	
	164	A	24	1	0	0	24.1	0.5	0.3	0.1	
	163	T	1	0	13	11	0.3	0.2	12.1	12.4	
	162	T	0	0	21	4	0.4	0.2	22.8	1.6	
	161	A	24	0	1	0	24.1	0.5	0.2	0.3	
-3	160	C	0	0	1	24	0.4	0.3	2.1	22.2	
-2	159	A	25	0	0	0	24.1	0.5	0.2	0.1	
-1	158	G	0	25	0	0	1.1	23.5	0.2	0.2	
1	157	G	0	25	0	0	1.3	23.0	0.3	0.3	
2	156	C	0	0	21	4	0.6	0.4	7.7	16.2	2.87E-07
3	155	G	10	15	0	0	11.0	13.1	0.5	0.4	
4	154	C	2	0	12	11	0.4	0.4	9.2	15.0	
5	153	G	2	18	0	5	5.1	10.9	0.6	8.4	3.60E-02
6	152	C	1	0	13	11	0.8	0.4	9.7	14.0	
7	151	G	6	19	0	0	8.8	15.5	0.2	0.4	
	150	C	2	0	2	21	0.5	0.3	1.3	23.0	
	149	C	1	0	3	21	0.5	0.2	1.4	23.0	
	148	A	19	2	2	2	23.9	0.5	0.2	0.3	
	147	C	0	1	2	22	0.4	0.3	1.5	22.7	
	146	C	1	1	1	22	0.6	0.3	1.8	22.3	

Figure 2. Comparison of Exonized Alus to Nonexonized Intronic Alus

(A) Profile of 166,276 full-length antisense intronic Alus. For each position, the number of appearances of each nucleotide (count per position) is shown as well as the frequency of each base in the position ("percent per position").

(B) Per position comparison between the 25 exonized Alus (observed) and nonexonized intronic antisense Alus. Expected profile was calculated by multiplying the profile matrix of nonexonized antisense intronic Alus from (A) by 25. Significant deviations between the observed and expected profiles are apparent in positions 156 and 153. In position 156, T is expected 7.6 times (marked gray) but observed 21 times (gray), indicating a strong tendency of this position to change from C to T in exonized Alus. In position 153, G is expected 10.9 times but appears 18 times, indicating the importance of G in position 5 of the splice site.

(C) Profile as in (A) but for 136,151 full-length sense intronic Alus.

(D) Per position comparison between the 136,151 full-length sense intronic Alus and 25 exonized Alus. Comparison as in (B), indicating positions 156 and 153 as having significant deviations between the observed and expected profiles, with only slight differences between the E values in (B) and (D).

unpair with U1 (see also Freund et al., 2003). A U1 containing a compensatory mutation restoring the base pairing in position three of the 5′ss restores AEx inclusion (Figure 3A, lane 9). In contrast, a T:A compensatory mutation in the same position failed to restore AEx inclusion (Figure 3A, lane 7). This suggests that G:C rather than A:U pairing in position three of GC 5′ss contains the sufficient energy for U1 binding to the 5′ss. The failure of this A:T pairing to promote AEx inclusion is in contrast to the Ψ:A U1:5′ss pairing that is required for

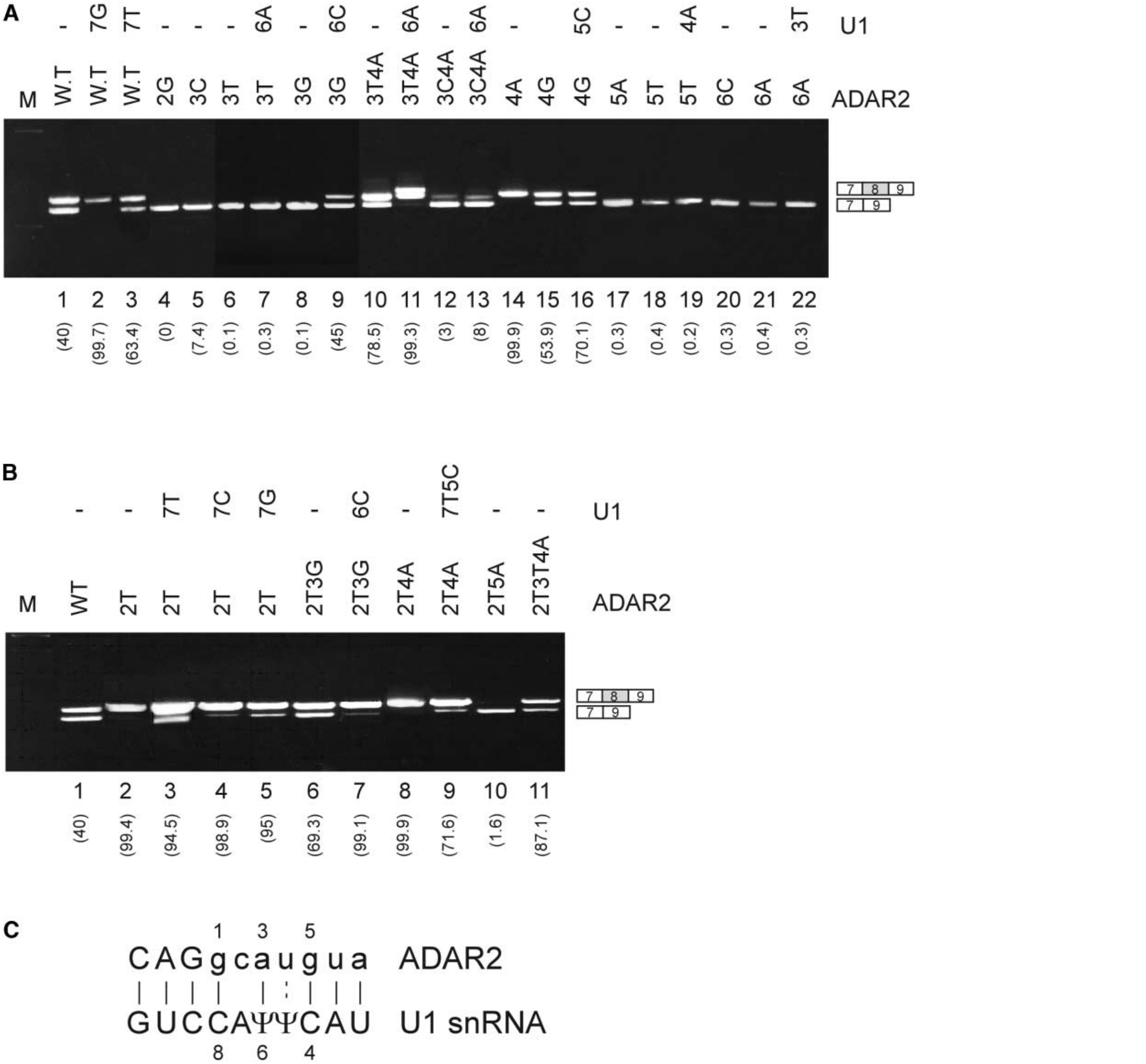


Figure 3. Splicing Assays on ADAR2 Minigene Mutants Using Compensatory U1 Mutants

(A) Analysis of GC introns. The top line shows the mutants in the U1 gene, and the second line shows the mutants in the ADAR2 5′ss sequence. The mutations are numbered according to the numbers indicated in (C). The indicated plasmid mutants were transfected or cotransfected into 293T cells, total cytoplasmic RNA was extracted, and splicing products were separated in 2% agarose gel after reverse transcription-polymerase chain reaction (RT-PCR). The leftmost lane is the DNA size marker. Lane 1, splicing products of wt ADAR2, and lanes 2–22, splicing products of the indicated ADAR2 minigene mutated at the 5′ss. The two mRNA isoforms are shown on the right. Numbers in parentheses in the bottom indicate percentage of the Alu-containing mRNA isoform as determined by TINA2 (100% corresponds to the total of both mRNA isoforms). Lane 11 contains two closely joined bands; the sequences of both were found to be identical and correspond to exon inclusion. This phenomenon may be attributed to migration of the overexpressed U1 together with the splicing product.

(B) Analysis of GT introns. Leftmost lane is the DNA size marker. Lane 1, splicing products of wt ADAR2, and lanes 2–11 are the splicing products of the indicated mutants.

(C) Schematic illustration of the base paring between the 5′ss of exon/intron eight of ADAR2 and U1. ADAR2 and U1 positions are numbered forward and reverse, respectively. Watson-Crick and non-Watson-Crick base pairing are marked by solid or dashed line, respectively.

AEx inclusion. This might be related to the stabilizing effect of Ψ on the backbone and the stem structure (Arnez and Steitz, 1994) or to the dynamics it allows for a noncanonical interaction at the base of a stem (Sundaram et al., 2000).

Position four of the 5′ss affects the level of alternative splicing of the AEx. Mutation of that position from T to A, which strengthens the base pairing with position five in U1, resulted in constitutive inclusion of the AEx (Figure 3A, lane 14). Mutation of the same position from T to G increased the level of the AEx inclusion from 40% to 54%. The inclusion level increased to 70% following cotransfection with U1 containing a compensatory mutation that base pairs with that G (Figure 3A, lanes 15 and 16).

To understand whether or not a cooperative effect between positions three and four exists, we produced double mutants and tested their inclusion ratios. Mutations that introduced a mispairing between U1 and position three (A to T) but allowed better base pairing of U1

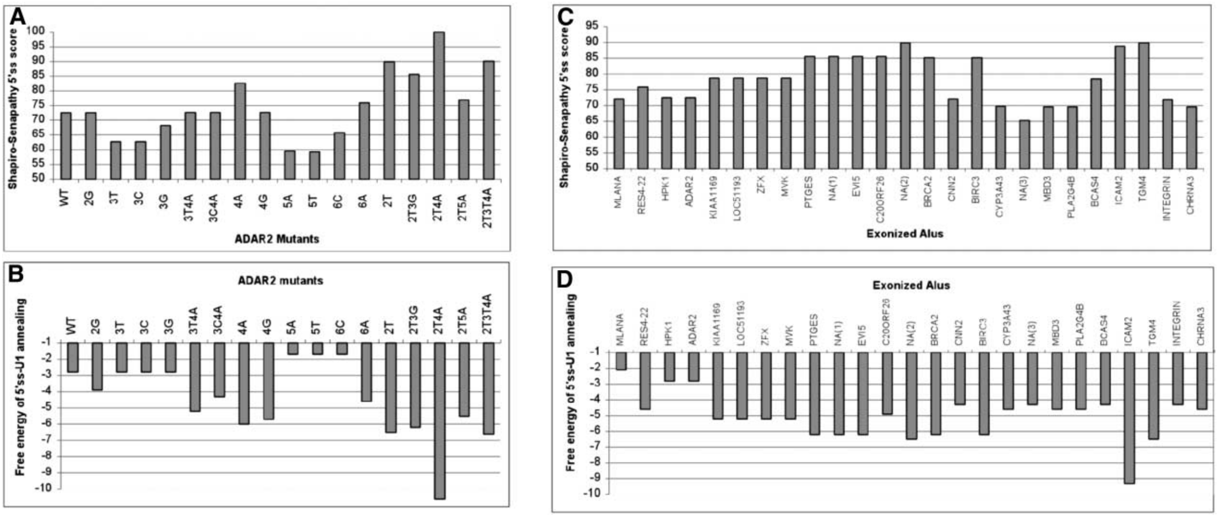


Figure 4. Strengths of 5′ss

(A) Shapiro-Senapathy score for 5′ss strength (Shapiro and Senapathy, 1987), calculated for the ADAR2 5′ss mutants indicated in Figures 3A and 3B.

(B) Free energy (ΔG) for 5′ss:U1 binding in ADAR2 5′ss mutants. Free energy was calculated using the software Mfold (Zuker, 1989), version 3.1 and according to Carmel et al. (2004).

(C) Shapiro-Senapathy score for the 25 exonized Alus indicated in Figure 1.

(D) Free energy (ΔG) for 5′ss:U1 binding in 5′ss of the 25 exonized Alus.

with position four (T to A) restored alternative splicing of the AEx with a ratio of inclusion of 78% (Figure 3A, lane 10). This splicing became constitutive when the cells were cotransfected with U1 containing a compensatory mutation that base pairs with the T in position three (Figure 3A, lane 11). Our results show that positions three and four are involved in controlling the level of exon inclusion in GC 5′ss. We note that position three in GC 5′ss may be T if position four is A; however, no such case was found in the set of exonized Alus, probably because such an event would require two sequential transversions (from GC to TA), a combination of events that have a very low probability of occurrence in the genome.

The cooperative effect between positions three and four did not occur in double mutants in which position three was mutated to C and position four to A (Figure 3A, lanes 12 and 13), implying that U but not C, in position three of the 5′ss, can form a noncanonical Ψ:U pairing with Ψ in position six of U1. Indeed, Ψ:U base pairing was recently reported to be important in position four of 5′ss in yeast splicing (Libri et al., 2002). The above results suggest a hierarchy in the strength of the pairing between U1 and position three of the 5′ss: Ψ:A > Ψ:U > Ψ:C. This hierarchy seems to determine the level of alternative splicing.

We further studied the importance of positions five and six in GC 5′ss. Mutations in these positions resulted in the total skipping of the AEx, which compensatory mutations in U1 failed to restore (Figure 3A, lanes 17–22). This suggests that other splicing factors also recognize these positions in 5′ss.

To study the regulation of AExs with GT in their 5′ss, we mutated the C in position 156 of the ADAR2 AEx to T, creating a GT 5′ss (Figure 3B). This mutation resulted in a shift from alternative to constitutive inclusion of the AEx (Figure 3B, lanes 1 and 2), further supporting our conclusion that the weak base pairing with U1 maintains the alternative splicing of the GC 5′ss. Cotransfection with U1 that unpairs with the T at position two had only a minor effect of ~5% exon skipping (Figure 3B, compare lane 2 to lanes 3–5), presumably reflecting the effect of U5(p220) binding to that nucleotide (Reyes et al., 1996). In addition, as predicted from our bioinformatical analysis (Figure 2), G in position five of the intron is essential for the alternative splicing of AExs, as a G to A mutation in that position resulted in total AEx skipping (Figure 3B, lane 10).

In contrast to the GC 5′ss, which required an obligatory A in position three, mutation of that position to G in the GT 5′ss reduced the splicing from constitutive to alternative but did not eliminate exon inclusion entirely (Figure 3B, compare lane 2 to 6). This probably stems from the fact that GT introns are generally stronger 5′ss than GC introns, and a G in position three of the GT 5′ss can form a G:Ψ pairing with U1 (see also Freund et al., 2003). Cotransfection with U1 containing a compensatory mutation that base pairs with the G in position three restored a constitutive splicing, suggesting that the base pairing of that position with U1 is a main factor affecting the alternative splicing ratio (Figure 3B, lanes 6 and 7, and also Figure 3A, lane 10). A similar preference for A over G in position three was also observed in other 5′ss where the base pairing between U1 and the 5′ss was suboptimal (Burge and Karlin, 1997; Freund et al., 2003).

To test the importance of position four in the GT 5′ss, we mutated it to A, which base pairs with the wt U1. This maintained the constitutive inclusion of the AEx (Figure 3B, lane 8). When a mutated U1 that unpairs with position four was cotransfected, the splicing became alternative (Figure 3B, lane 9). A return to alternative splicing was also observed when an additional mutation, which unpairs with U1, was introduced in position three (Figure 3B, lane 11). This further supports the results from the GC 5′ss analysis, which showed that the nucle-

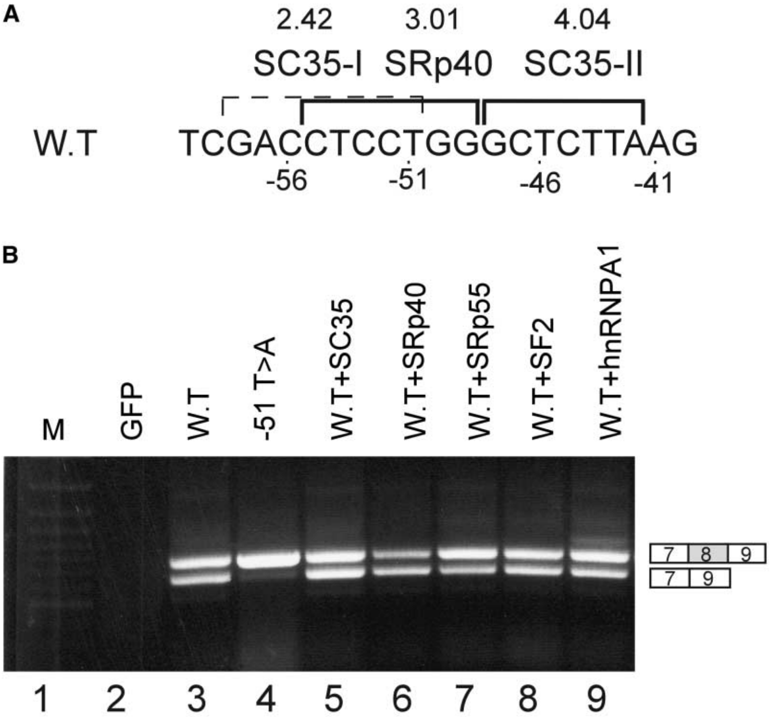


Figure 5. A Potential ESE Site in Exon Eight of ADAR2 Is Not Involved in Alternative Splicing Regulation

(A) The ADAR2 sequence from position −40 to −60 upstream of the 5′ss. SR protein potential sites are marked with a solid line; the broken line indicates a potential site that was enhanced by the mutation in position −51 from T to A. The potential sites were detected by ESEfinder (Roca et al., 2003); the type and binding score of each site is indicated.

(B) Transfection or cotransfection was performed in 293T cells. Total RNA and RT-PCR were performed as described in Figure 3. Lane 1, DNA size marker; lane 2, vector only (pEGFP-C3); lane 3, splicing products of wt ADAR2; lane 4, splicing products of the T to A mutant in position −51 that enhanced the ESE of SC35-I from a score of 2.4 to 4.2 (−51 T > A); lanes 5–9, splicing products of wt ADAR2 with the indicated SR/A1 hnRNP protein.

otide composition in positions three and four affects the delicate balance of skipping or inclusion of alternatively spliced exons.

From these serial mutations in the 5′ss and the compensatory mutations in U1 we conclude the following. (1) U1:5′ss base pairing is involved in both GC and GT 5′ss selection. (2) The alternative splicing of ADAR2 AEx is maintained due to the unpairing of U1 with position two of the 5′ss. (3) The nucleotide composition of positions three and four in the intron, both in GT 5′ss and in GC 5′ss, control the delicate skipping/inclusion ratio depending on the canonical/noncanonical base pairing of these positions with U1. (4) An A in position three of the intron is more important in GC 5′ss than in GT 5′ss, possibly due to the need to avoid two successive nonpairing positions. (5) G in position five is essential for the selection of 5′ss in the two types of introns.

The results for position four of the 5′ss indicate that when this position is mutated to A the ADAR2 exon becomes constitutive. However, there are two cases among the 25 alternatively spliced Alu exons in our compilation that contain A in position four (Figure 1, rows 2 and 22). A similar inconsistency is observed for the mutation of position six from T to C, which makes the ADAR2 exon inactive, but is found in several exonized Alus (Figure 1). To further examine this inconsistency, we calculated splice site scores and free energy of U1:5′ss binding (ΔG) in exonized and mutated Alus (Figure 4). The splice site score is a measure of how “close” the splice site is to the consensus sequence profile of 5′ss (Shapiro and Senapathy, 1987). The free energy is a measure of the U1:5′ss binding strength, taking also into consideration the differences between G:C, A:U, and G:U base pairing and stacking energy––lower ΔG values stand for stronger binding and might indicate higher exon inclusion/exclusion ratio (Carmel et al., 2004).

When position six in ADAR2 is mutated to C, the 5′ss score is reduced from 73 to 65 (Figure 4A, mutant 6C) and the free energy is increased from −2.8 to −1.7, which indicates that binding is inefficient (Figure 4B, mutant 6C). In exonized Alus where position six is C the 5′ss score can also get as low as 65 [for example NA(3) in Figure 4C], but the U1:5′ss free binding energy in these exons is never higher than −4.0 [Figure 4D, see for example ZFX, MVK, NA(3), and MBD3]. The lower ΔG stands for more efficient U1:5′ss binding, which explains why these exons are still recognized.

When position four in ADAR2 is mutated to A, the 5′ss score increases from 73 to 83, and the ΔG becomes −6.0 (Figures 4A and 4B, mutant 4A), which is in agreement with the fact that this mutation results in a constitutively spliced exon. In the exonized Alu in gene RES4-22 position four is A, but the 5′ss score is only 75 (similar to the score of the wt alternative ADAR2) and the ΔG is −4.6 (Figures 4C and 4D). This can explain the fact that the AEx of RES4-22 is alternatively spliced. In the AEx of ICAM2, position four is also A, and the 5′ss score and ΔG indicate a very high affinity to U1. However, that AEx is alternatively spliced, which might indicate that sequence elements other than the ones in the actual 5′ss are involved in its regulation.

Overall, the 5′ss score for exonized Alus was similar to that for non-Alu alternative cassette exons, averaging 78.3 in the 25 AExs and 79.8 in the set of 243 non-Alu cassette exons from Sorek and Ast (2003). The free energy value was also similar between the sets, with an average of −5.08 in the 25 AExs and −5.29 in the non-Alu cassette exons.

The results presented here and in Lev-Maor et al. (2003) indicate that the sequence composition of both 3′ss and 5′ss determines the alternative splicing ratio of AExs. In addition, our statistical comparison between intronic Alus and exonized ones shows that only two

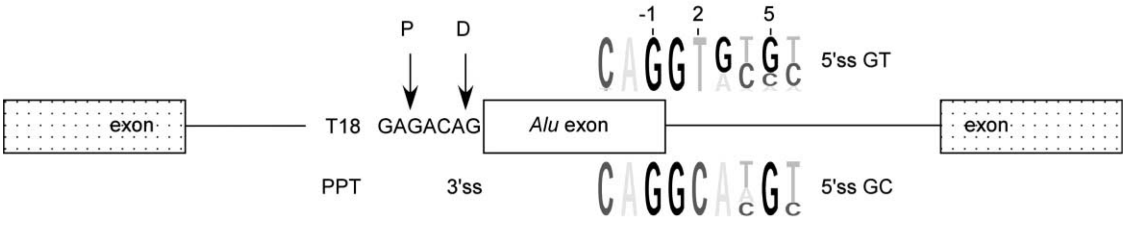


Figure 6. Sequence Elements Needed for the Creation of an Alternatively Spliced *Alu* Exon

The prevalent 3′ splice site (3′ss) is found in position 275 of the antisense *Alu* and is composed of a polypyrimidine tract (PPT) averaging 18 bases in length, followed by a 3′ss motif of GAGACAG containing a proximal and a distal AG (P and D, respectively) (Lev-Maor et al., 2003). The prevalent 5′ss can either begin with GT or with GC. Pictograms (http://genes.mit.edu/pictogram.html) depict the profiles of 21 GT-exonized and 4 GC-exonized *Alu*s, above and below the box, respectively.

positions along the *Alu* sequence substantially differ between intronic and exonized *Alu*s, both in the 5′ss sequence (Figure 2). These results imply that only the sequences of the 3′ and 5′ss of the AExs are important for exonization. This conclusion is supported by the finding that, so far, all the mutations leading to exonization of intronic *Alu*s that cause genetic disorders were only found to affect 5′ss or 3′ss sequences (references in Figure 1 and in Lev-Maor et al., 2003).

Still, it is possible that the splicing of AExs is regulated by splicing enhancers or silencers residing outside the splice sites. We used the ESE finder (Roca et al., 2003) to search the sequence of the AEx of ADAR2 for such potential sequences. Seven potential binding sites for SR proteins were found, three of them in close proximity (Figure 5A). SR proteins are known to be involved in alternative splicing regulation through binding to short sequences on the RNA molecule (Graveley, 2001). The score of all potential SR binding sites was low, but such sites can still be functional (Roca et al., 2003). To test whether these sites are functional, we serially mutated positions −48 to −54. We found no effect on the splicing pattern of the AEx, indicating that these potential ESEs are not functional (data not shown). Mutations in position −11, embedded within another weak potential ESE, gave similar results (data not shown).

To examine whether other sites are involved in the regulation of the AEx splicing, we cotransfected 293T cells with the ADAR2 minigene and various plasmids containing the most abundant splicing regulatory proteins (SR proteins and hnRNP A1, the latter was shown to promote exon skipping; Figure 5B, lanes 5–9) (Cartegni et al., 2002). In principle, if one of these proteins was involved in the regulation of AEx splicing, then increasing its nuclear concentrations might affect the skipping/inclusion ratio of the AEx. This was not the case: none of these proteins affected the alternative splicing of the AEx. This suggests that the AEx of ADAR2 has no splicing enhancers or silencers that can bind to one of these proteins and regulate the AEx splicing, but we cannot rule out the possibility that other splicing regulatory proteins, which are not part of the panel in Figure 5, affect the inclusion/skipping ratio of this exon. As a control to that experiment, we artificially created a strong binding site for SC35 by mutating T in position −51 to A and thus increasing the ESE score from 2.4 to 4.2. This mutation led to a shift toward exon inclusion, indicating that, in principle, SR proteins have the potential to modulate the splicing of AExs but are presumably not involved in the case of the exon under study (Figure 5B, lanes 3 and 4). This further supports the analysis in Figure 2, implying that in the general exonization process of AExs there is no selective pressure for the creation or loss of a splicing regulatory sequence/s located outside of the splice sites and suggesting that only the sequences of the splice sites are involved in that process.

Discussion

We have analyzed the factors influencing *Alu* exonization through a combined computational and experimental approach. We showed that 5′ss can be created de novo within *Alu* sequences in a process requiring only minimal base substitutions, involving positions two and five of the intron. By mutational analysis we extensively studied the delicate balance between the different positions of the 5′ss and U1 and showed how this balance controls the inclusion/skipping ratio of the exon in both GC and GT introns.

Our results suggest that the fast decay of CG dinucleotides to TG or CA in the human genome is a major driving force for *Alu* exonizations. This decay causes a CAGGCG motif in positions 160-154 of antisense *Alu*s to become either CAGGTG (CG becomes TG, creating functional GT 5′ss) or CAGGCA (CG becomes CA, creating functional GC 5′ss with A in position three). Ironically, CG decay is considered one of the evolutionary mechanisms promoting the silencing of retroposition of active *Alu*s in the genome, as there is a correlation between the number of CG dinucleotides in the *Alu* sequence and its retroposition activity (Deininger and Batzer, 2002). Therefore, we would expect a negative correlation between retropositional activities of *Alu*s and exonization, i.e., if an *Alu* became an exon, there is less probability for it to be transpositionally active.

The results presented in this article and the previous analysis of 3′ss formation and regulation in exonized *Alu*s (Lev-Maor et al., 2003) provide the molecular basis for Alu exonization. Figure 6 summarizes these findings and shows a model for the composition of the prevalent 3′ss and 5′ss in exonized *Alu*s. The prevalent 3′ss is found in position 275 of the antisense *Alu*. It is composed of a polypyrimidine tract (PPT) averaging 18 bases in length, followed by a 3′ss motif of GAGACAG. In the 3′ss motif the distal AG is selected, and there is a delicate interplay between the two AGs. The G at position −7 (the seventh nucleotide upstream of the distal AG) suppresses the selection of the proximal AG: when that G is mutated there are two effects––the AEx becomes constitutive and the proximal AG is selected. The proximal AG is essential to weaken the selection of the distal AG, thus maintaining alternative splicing. The four nucleotides distance between the proximal and distal AG also ensure that mode of alternative splicing; increasing that distance leads to AEx skipping, and AEx inclusion can be restored by a high concentration of the second step splicing factor hSlu7 when the distance between the two AGs is over eight nucleotides. On the other side of the AEx the prevalent 5′ss is in position 158. There are two types of 5′ss that can be used in that same position: GT 5′ss (more common) and GC 5′ss, in which position three has to be A (less common).

This model, as well as the results presented in Figure 5, indicate that exonization of Alu sequences depends almost solely on the sequence composition of the potential 3′ and 5′ splice sites. Alus containing the 5′ss described here and the 3′ss described in Lev-Maor et al. (2003) are predicted to become alternatively spliced exons. We scanned the genome for Alus that have the potential to undergo exonization (see Experimental Procedures). We found 244,472 Alus in the antisense orientation located in introns of genes (not necessarily being full-length Alus). Of these, 46,518 had GT or GC 5′ss (GT: 15,413 and GC: 31,105) and 7810 also had the ADAR2-like 3′ss with a strong polypyrimidine tract preceding it. This set represents Alus that have either been exonized already or are "on the verge" of exonization (Supplemental Database S1 at http://www.molecule.org/cgi/content/full/14/2/221/DC1).

Presumably, the 7810 Alus should be exonized. We manually examined several examples from this set: only a minority of these Alus were found in expressed sequences (ESTs or mRNAs), and therefore, we lack evidence for exonization in the majority of cases. In one case of a "perfect" candidate Alu found in intron 20 of the inhibitor of κ-light polypeptide gene enhancer in B cell's kinase complex-associated protein (IKBKAP) gene, we experimentally tested the possibility of exonization. Mutations in this gene are known to lead to familial dysautonomia (FD), an autosomal recessive congenital neuropathy disorder. The major FD-causing mutation (99.5%) affects the 5′ss of exon 20 in IKBKAP, changing position +6 from T to C and leading to aberrant splicing (Fini and Slaugenhaupt, 2002). The perfect Alu elment in intron 20 contains a putative PPT of 18 Ts, followed by the GAGACAG motif. In the putative 5′ss, it contains a CAGgtgtgc sequence having a 5′ss score of 78.8 and ΔG of -5.2, which is similar to that of exonized Alus. This Alu is located more than 100 nucleotides from the flanking splice sites, so that there are probably no intron size constraints on the exonization process. In spite of these facts, we could detect no exonization of this Alu (data not shown).

There may be several reasons for these findings. First, the inclusion ratio of most AExs is 10% or lower (Sorek et al., 2002), and for many of these Alus, EST sampling may be more limited than the level required for discovery of lowly expressed exonized Alus. ESTs that may become available in the future may uncover additional instances of Alu exonization. Second, some Alus might lack the proper branch point upstream of their 3′ss (although the branch point consensus sequence is degenerate). Third, other factors, such as reading frame compatibility, have been shown to affect spliceosome selection of exons (Li et al., 2002). Finally, intronic sequence elements that reside outside the Alu sequence itself might have an inhibitory effect on the splicing of Alu. Indeed, Fairbrother and Chasin (2000) demonstrated that a substantial fraction of the human genome contains sequences that have splicing-inhibition properties. In any case, revealing why some of these Alus are selected as exons and some are not could be a major step toward understanding how mammalian exons are defined.

Based on the aforementioned mutation leading to CCFDN syndrome, we can predict that C-to-T mutation in position two of the GC 5′ss of some of these 7810 perfect intronic Alus may, in some cases, lead to constitutive exonization and may result in genetic disease. These Alus on the verge of exonization may have an additional evolutionary role: they may serve as a "reservoir" for new human-specific exons that may one day be exapted (i.e., adapted to a function different from their original) into promoting speciation of the human lineage.

Experimental Procedures

Building a Genomic Alu Profile

The AluGene database (http://alugene.tau.ac.il/; Dagan et al., 2004) was used to extract full-length genomic Alu sequences that are found within introns in the antisense and sense orientations of Alu. This search yielded 166,276 and 136,151 full-length Alus in the antisense and sense orientations, respectively. Each of these Alus was aligned to the Alu consensus sequence, using ClustalW 1.8 (Thompson et al., 1994). A nucleotide frequency profile was built for each site along the Alu consensus sequence (Figures 2A and 2C). In a similar manner, the sequences of 25 Alus that were exonized using position 157 as 5′ss (Figure 1) were aligned to their consensus and an "observed" profile was built (Figures 2B and 2D). A parallel "expected" profile was calculated, by multiplying the frequency profile matrix from the nonexonized intronic Alus by 25 (the number of exonized sequences). Statistical significance between the observed and expected profile was calculated using chi-square statistical test. AluGene was further queried in the same manner to find the 244,472 Alus in the antisense orientation (not necessarily full-length) located in introns of known genes.

Calculation of Splice Site Strength

Splice site scores in Figures 4A and 4C were calculated using the matrix from Shapiro and Senapathy (1987). The free energy in Figures 4B and 4D was calculated using the Mfold software (Zuker, 1989) version 3.1, which predicts the free energy (ΔG) of a single folded RNA strand. To predict the ΔG in U1:5′ss annealing, we concatenated their sequence strands into one RNA strand as follows: "NN"-5′ss-"NNNNN"-reversed U1-"NN." This sequence was given as the input for Mfold. Since this software does not hybridize the "N" sequence because it is considered neutral, the resulting calculated free energy was that of the U1:5′ss hybrid. A more detailed description is found in Carmel et al. (2004). For both score and energy calculations, the splice site was set to be at positions -3 to $+6$ relative to the 5′ss.

Plasmid Constructs

Oligonucleotide primers were designed to amplify a minigene that contains the exons 7, 8, and 9 of the gene adenosine deaminase (ADAR2). Each primer contained an additional sequence encoding a restriction enzyme. The PCR product (2.2 kb) was restriction designed and inserted between the KpnI/BglII sites in the pEGFP-C3 plasmid (Clonthech). The U1 gene was cloned in the pCR vector.

Site-Directed Mutagenesis

Oligonucleotide primers containing the desired mutations were used to amplify the mutation-containing replica of either the wt ADAR2 minigene plasmid or the U1 gene, respectively. The products were treated with DpnI restriction enzyme (12 U) (New England Biolabs) at 37°C for 1 hr. The mutant DNA (1–4 μL) was transformed into E. coli DH5α strain. Colonies were picked, followed by miniprep (QIAgene) and midiprep (BRL). All plasmids were confirmed by sequencing.

Transfection, RNA Isolation, and RT-PCR Amplification

293T cell line was cultured in Dulbecco's Modification of Eagle medium, supplemented with 4.5 gr/mL glucose (Renium) and 10% fetal calf serum (Biological Industries). Cells were cultured in 60 mm dishes under standard conditions at 37 °C with 5% CO_2. Cells were grown to 50% confluence, and transfection was performed using 12 μL FuGENE6 (Roche) with 4 μg of plasmid DNA. After 48 hr, cells were harvested. Total cytoplasmic RNA was extracted using Tri Reagent (Sigma), followed by treatment with 2 U DNase RNase-free (Ambion). Reverse transcription (RT) was performed on 2 μg total cytoplasmic RNA for 1 hr at 42 °C, using the pEGFP-C3-specific reverse primer and 2 U reverse transctiptase of avian myeloblastosis virus (RT-AMV, Roche).

The spliced cDNA products derived from the expressed minigene were detected by PCR, using the pEGFP-C3-specific reverse primer and an exon seven forward primer. Amplification was performed for 30 cycles, consisting of 94 °C for 30 s, 61 °C for 45 s, and 72 °C for 1 min using high-fidelity Taq (Roche). The products were resolved on 2% agarose gel.

Acknowledgments

We thank M. Kupiec for critical reading, we thank Alan Weiner for the U1 gene, and we thank Adrian Krainer and Stefan Stemm for the SR plasmids. We also thank the Bioinformatics Unit at Tel Aviv University for providing technical assistance and computation facilities, and we thank Amir Goren for generating the website. R.S., G.L.-M., M.R., and G.A. were supported by a grant from the Israel Science Foundation, FD Hope, Israel Cancer Association, Chief Scientist of Israel Health Ministry, and the MOP India-Israel. R.S., T.D., F.B., and D.G. were supported by the Norman and Rose Lederer Chair of Biology at Tel Aviv University.

Received: January 26, 2004
Revised: March 17, 2004
Accepted: March 23, 2004
Published: April 22, 2004

References

Arnez, J.G., and Steitz, T.A. (1994). Crystal structure of unmodified tRNA(Gln) complexed with glutaminyl-tRNA synthetase and ATP suggests a possible role for pseudo-uridines in stabilization of RNA structure. Biochemistry 33, 7560–7567.

Batzer, M.A., Kilroy, G.E., Richard, P.E., Shaikh, T.H., Desselle, T.D., Hoppens, C.L., and Deininger, P.L. (1990). Structure and variability of recently inserted Alu family members. Nucleic Acids Res. 18, 6793–6798.

Black, D.L. (2000). Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. Cell 103, 367–370.

Brett, D., Pospisil, H., Valcarcel, J., Reich, J., and Bork, P. (2002). Alternative splicing and genome complexity. Nat. Genet. 30, 29–30.

Brosius, J. (1999). RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238, 115–134.

Brow, D.A. (2002). Allosteric cascade of spliceosome activation. Annu. Rev. Genet. 36, 333–360.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 78–94.

Carmel, I., Tal, S., Vig, I., and Ast, G. (2004). Comparative analysis detects dependencies among the 5′ splice-sites positions. RNA 10, 828–840.

Cartegni, L., Chew, S.L., and Krainer, A.R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat. Rev. Genet. 3, 285–298.

Cohen, J.B., Broz, S.D., and Levinson, A.D. (1993). U1 small nuclear RNAs with altered specificity can be stably expressed in mammalian cells and promote permanent changes in pre-mRNA splicing. Mol. Cell. Biol. 13, 2666–2676.

Dagan, T., Sorek, R., Sharon, E., Ast, G., and Graur, D. (2004). AluGene: a database of Alu elements incorporated within protein-coding genes. Nucleic Acids Res. 32, D489–D492.

Deininger, P.L., and Batzer, M.A. (2002). Mammalian retroelements. Genome Res. 12, 1455–1465.

Dewannieux, M., Esnault, C., and Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. Nat. Genet. 35, 41–48.

Dubbink, H.J., de Waal, L., van Haperen, R., Verkaik, N.S., Trapman, J., and Romijn, J.C. (1998). The human prostate-specific transglutaminase gene (TGM4): genomic organization, tissue-specific expression, and promoter characterization. Genomics 51, 434–444.

Fairbrother, W.G., and Chasin, L.A. (2000). Human genomic sequences that inhibit splicing. Mol. Cell. Biol. 20, 6816–6825.

Farrer, T., Roller, A.B., Kent, W.J., and Zahler, A.M. (2002). Analysis of the role of Caenorhabditis elegans GC-AG introns in regulated splicing. Nucleic Acids Res. 30, 3360–3367.

Fini, M.E., and Slaugenhaupt, S.A. (2002). Enzymatic mechanisms in corneal ulceration with specific reference to familial dysautonomia: potential for genetic approaches. Adv. Exp. Med. Biol. 506, 629–639.

Freund, M., Asang, C., Kammler, S., Konermann, C., Krummheuer, J., Hipp, M., Meyer, I., Gierling, W., Theiss, S., Preuss, T., et al. (2003). A novel approach to describe a U1 snRNA binding site. Nucleic Acids Res. 31, 6963–6975.

Graveley, B.R. (2001). Alternative splicing: increasing diversity in the proteomic world. Trends Genet. 17, 100–107.

Harrison, P.M., Kumar, A., Lang, N., Snyder, M., and Gerstein, M. (2002). A question of size: the eukaryotic proteome and the problems in defining it. Nucleic Acids Res. 30, 1083–1090.

Hastings, M.L., and Krainer, A.R. (2001). Pre-mRNA splicing in the new millennium. Curr. Opin. Cell Biol. 13, 302–309.

Hu, M.C., Qiu, W.R., Wang, X., Meyer, C.F., and Tan, T.H. (1996). Human HPK1, a novel human hematopoietic progenitor kinase that activates the JNK/SAPK kinase cascade. Genes Dev. 10, 2251–2264.

Johnson, S., Halford, S., Morris, A.G., Patel, R.J., Wilkie, S.E., Hardcastle, A.J., Moore, A.T., Zhang, K., and Hunt, D.M. (2003). Genomic organisation and alternative splicing of human RIM1, a gene implicated in autosomal dominant cone-rod dystrophy (CORD7). Genomics 81, 304–314.

Jurka, J., and Milosavljevic, A. (1991). Reconstruction and analysis of human Alu genes. J. Mol. Evol. 32, 105–121.

Kazazian, H.H., Jr. (2000). Genetics. L1 retrotransposons shape the mammalian genome. Science 289, 1152–1153.

Knebelmann, B., Forestier, L., Drouot, L., Quinones, S., Chuet, C., Benessy, F., Saus, J., and Antignac, C. (1995). Splice-mediated insertion of an Alu sequence in the COL4A3 mRNA causing autosomal recessive Alport syndrome. Hum. Mol. Genet. 4, 675–679.

Kreahling, J., and Graveley, B.R. (2004). The origins and implications of Aluternative splicing. Trends Genet. 20, 1–4.

Kunkel, T.A., and Diaz, M. (2002). Enzymatic cytosine deamination: friend and foe. Mol. Cell 10, 962–963.

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.

Lev-Maor, G., Sorek, R., Shomron, N., and Ast, G. (2003). The birth of an alternatively spliced exon: 3′ splice-site selection in Alu exons. Science 300, 1288–1291.

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. USA 99, 5277–5282.

Libri, D., Duconge, F., Levy, L., and Vinauger, M. (2002). A role for the Psi-U mismatch in the recognition of the 5′ splice site of yeast introns by the U1 small nuclear ribonucleoprotein particle. J. Biol. Chem. 277, 18173–18181.

Makalowski, W., Mitchell, G.A., and Labuda, D. (1994). Alu sequences in the coding regions of mRNA: a source of protein variability. Trends Genet. 10, 188–193.

Mihovilovic, M., Mai, Y., Herbstreith, M., Rubboli, F., Tarroni, P., Clementi, F., and Roses, A.D. (1993). Splicing of an anti-sense Alu sequence generates a coding sequence variant for the alpha-3 subunit of a neuronal acetylcholine receptor. Biochem. Biophys. Res. Commun. 197, 137–144.

Miller, M., and Zeller, K. (1997). Alternative splicing in lecithin:cholesterol acyltransferase mRNA: an evolutionary paradigm in humans and great apes. Gene 190, 309–313.

Mitchell, G.A., Labuda, D., Fontaine, G., Saudubray, J.M., Bonnefont, J.P., Lyonnet, S., Brody, L.C., Steel, G., Obie, C., and Valle, D. (1991). Splice-mediated insertion of an Alu sequence inactivates ornithine delta-aminotransferase: a role for Alu elements in human mutation. Proc. Natl. Acad. Sci. USA 88, 815–819.

Modrek, B., and Lee, C. (2002). A genomic view of alternative splicing. Nat. Genet. 30, 13–19.

Modrek, B., and Lee, C.J. (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.

Nissim-Rafinia, M., and Kerem, B. (2002). Splicing regulation as a potential genetic modifier. Trends Genet. 18, 123–127.

Reyes, J.L., Kois, P., Konforti, B.B., and Konarska, M.M. (1996). The canonical GU dinucleotide at the 5′ splice site is recognized by p220 of the U5 snRNP within the spliceosome. RNA 2, 213–225.

Roca, X., Sachidanandam, R., and Krainer, A.R. (2003). Intrinsic differences between authentic and cryptic 5′ splice sites. Nucleic Acids Res. 31, 6321–6333.

Shapiro, M.B., and Senapathy, P. (1987). RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Res. 15, 7155–7174.

Sorek, R., and Ast, G. (2003). Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. Genome Res. 13, 1631–1637.

Sorek, R., Ast, G., and Graur, D. (2002). Alu-containing exons are alternatively spliced. Genome Res. 12, 1060–1067.

Stoilov, P., Meshorer, E., Gencheva, M., Glick, D., Soreq, H., and Stamm, S. (2002). Defects in pre-mRNA processing as causes of and predisposition to diseases. DNA Cell Biol. 21, 803–818.

Sundaram, M., Durant, P.C., and Davis, D.R. (2000). Hypermodified nucleosides in the anticodon of tRNA(Lys) stabilize a canonical U-turn structure. Biochemistry 39, 15652.

Svineng, G., Fassler, R., and Johansson, S. (1998). Identification of beta1C-2, a novel variant of the integrin beta1 subunit generated by utilization of an alternative splice acceptor site in exon C. Biochem. J. 330, 1255–1263.

Thanaraj, T.A., and Clark, F. (2001). Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. Nucleic Acids Res. 29, 2581–2593.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Varon, R., Gooding, R., Steglich, C., Marns, L., Tang, H., Angelicheva, D., Yong, K.K., Ambrugger, P., Reinhold, A., Morar, B., et al. (2003). Partial deficiency of the C-terminal-domain phosphatase of RNA polymerase II is associated with congenital cataracts facial dysmorphism neuropathy syndrome. Nat. Genet. 35, 185–189.

Vervoort, R., Gitzelmann, R., Lissens, W., and Liebaers, I. (1998). A mutation (IVS8 +0.6kbdelTC) creating a new donor splice site activates a cryptic exon in an Alu-element in intron 8 of the human beta-glucuronidase gene. Hum. Genet. 103, 686–693.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562.

Xu, Q., and Lee, C. (2003). Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. Nucleic Acids Res. 31, 5635–5643.

Yang, Z., and Yoder, A.D. (1999). Estimation of the transition/transversion rate bias and species sampling. J. Mol. Evol. 48, 274–283.

Zhuang, Y., and Weiner, A.M. (1986). A compensatory base change in U1 snRNA suppresses a 5′ splice site mutation. Cell 46, 827–835.

Zuker, M. (1989). On finding all suboptimal foldings of an RNA molecule. Science 244, 48–52.