

Comparative Analysis Identifies Exonic Splicing Regulatory Sequences—The Complex Definition of Enhancers and Silencers

Amir Goren,^{1,3} Oren Ram,^{1,3} Maayan Amit,¹ Hadas Keren,¹ Galit Lev-Maor,¹ Ida Vig,¹ Tal Pupko,² and Gil Ast^{1,*}

¹Department of Human Molecular Genetics and Biochemistry

Sackler Faculty of Medicine

²Department of Cell Research and Immunology

George S. Wise Faculty of Life Sciences

Tel Aviv University

Tel Aviv 69978

Israel

Summary

Exonic splicing regulatory sequences (ESRs) are *cis*-acting factor binding sites that regulate constitutive and alternative splicing. A computational method based on the conservation level of wobble positions and the overabundance of sequence motifs between 46,103 human and mouse orthologous exons was developed, identifying 285 putative ESRs. Alternatively spliced exons that are either short in length or contain weak splice sites show the highest conservation level of those ESRs, especially toward the edges of exons. ESRs that are abundant in those subgroups show a different distribution between constitutively and alternatively spliced exons. Representatives of these ESRs and two SR protein binding sites were shown, experimentally, to display variable regulatory effects on alternative splicing, depending on their relative locations in the exon. This finding signifies the delicate positional effect of ESRs on alternative splicing regulation.

Introduction

An average human gene consists of 8.8 exons of approximately 130 nucleotides separated by 7.8 introns of ~3000 nucleotides (Lander et al., 2001). Thus, the mechanism that directs the splicing machinery to locate short exonic sequences that are flanked by long intronic sequences is of a great interest. The splicing reaction is catalyzed by the spliceosome—a complex that contains five small nuclear ribonucleoproteins (snRNPs) and at least 150 additional proteins (Hartmuth et al., 2002; Jurica and Moore, 2003; Nilsen, 2003; Zhou et al., 2002). In the process of the pre-mRNA splicing, the splicing machinery precisely recognizes the exon-intron junctions, removes the introns, and joins the exons to form a mature mRNA (Black, 2003). Four splice signals are essential for accurate splicing: the 5' and 3' splice site (5'ss and 3'ss), the polypyrimidine tract, and the branch site sequence (Black, 2003). These four splice signals alone cannot execute the splicing process correctly, and it has been estimated that these splicing signals provide only

one-half of the information required for recognition by the splicing machinery (Lim and Burge, 2001).

Studies of the molecular basis of splicing reveal the existence of exonic and intronic *cis*-acting regulatory sequences, which bind *trans*-acting factors and thus influence splice-site selection. These *cis*-acting elements are relatively short, usually 4–18 nucleotides (Cartegni et al., 2002; Fairbrother et al., 2002). These motifs are classified as exonic or intronic splicing enhancers (ESE/ISE) and silencers (ESS/ISS), and they are required for the regulation of both constitutive and alternative splicing (Blencowe, 2000; Graveley, 2000). Specific binding of splicing regulatory proteins (such as SR proteins) to ESEs and ESSs assists in the placement of the spliceosome on the appropriate splice sites (Sanford et al., 2005; Singh and Valcarcel, 2005). Alternative splicing is a process in which more than one mRNA is produced from the same precursor messenger RNA by using different 5'ss and/or 3'ss, giving rise to functionally different proteins (Maniatis and Tasic, 2002). Alternative splicing increases the diversity of the transcriptome within and between cells and adds an additional regulatory dimension to the expression pattern of an organism (Ast, 2004; Graveley, 2001).

Some recent studies employed computational methods to predict sequences that have ESE activity. The RESCUE-ESE method identified 238 hexamers preferentially associated with constitutive exons with weak splice sites (Fairbrother et al., 2002). In another study, octamers that are overrepresented in internal noncoding exons, versus unspliced pseudoexons and 5'-untranslated regions (UTRs) of transcripts of intronless genes, were compiled (Zhang and Chasin, 2004). Similarly, sequence motifs that are overabundant in genes that harbor introns over those that do not harbor introns were extracted (Fedorov et al., 2001). Putative ESSs were also identified by using an *in vivo* splicing reporter system developed to screen a library of random decanucleotides (Wang et al., 2004). Some of these studies tested, experimentally, a sample of candidate motifs for their splicing activity.

Recent results have shown that, in exon 12 of the cystic fibrosis gene, approximately 31% of the synonymous substitutions that were tested affect the splicing efficiency by inducing exon skipping and result in an inactive protein (Pagani et al., 2005). This indicates that evolution of mammalian exons is affected by at least two types of selection pressures: preservation of protein functionality and conservation of exonic splicing regulatory sequences (ESRs).

The idea underlying a computational search for regulatory sequences using comparative genomics is that selective pressure causes regulatory elements to evolve at a slower rate than less-functional or nonfunctional sequences. Therefore, highly conserved sequences in a collection of homologous regions are good candidates for functioning as regulatory elements (Blanchette and Tompa, 2002). It is estimated that human and mouse diverged 75–130 million years ago (Waterston et al., 2002). Most of the human genes (99%) have a mouse ortholog

*Correspondence: gilast@post.tau.ac.il

³These authors contributed equally to this work.

with a high level of sequence similarity of the exons (88%). However, while orthologous alternatively spliced exons show a very high degree of conservation (94%), constitutively spliced exons show a lower degree of only 89% conservation (Sorek et al., 2004c; Sugnet et al., 2004). Therefore, human-mouse orthologous exons are a good model system for such analysis.

Here, we developed a comparative genomics method that identifies putative regulatory sequences, based on the combination of two features: (1) the evolutionary conservation of wobble positions between human and mouse orthologous exons, since the wobble positions are almost free of coding constraints, and (2) the analysis of the overabundance of sequence motifs, compared with their random expectation, given by their codon relative frequency. This method resulted in 285 significant ESR motifs. We validated our putative ESRs experimentally by choosing ten motifs and placing them in two suboptimal exons (minigenes). Remarkably, we found distinctions in the inclusion/exclusion levels between the two minigenes mediated by the same ESR candidates—the same ESR sequence can function as an enhancer in one exon and a silencer in the other. We demonstrated the functionality of those ESRs in their natural environment by mutating four ESRs located in three different reporting exons and showing their effect on splicing. In addition, we slid two known SR binding sites (SRp40 and SF2/ASF) along the suboptimal alternatively spliced exon. Strikingly, we observed that those sequences could function as ESEs or ESSs, depending on their positions along the exon. Elimination of those sites by point mutations restored the original splicing level, indicating the specificity of the positional-dependent effect. Thus, we suggest that the distinction between ESE and ESS based solely on the sequence of the motif should be refined—the function of ESRs also depends on their positions in the exons and on other spatial effectors.

Results

Computational Prediction of Splicing Regulatory Motifs

We compiled a data set of 46,103 human-mouse orthologous exons of the same length and identified the reading frame for each exon. To extract putative ESRs, we searched this data set for conserved and overabundant hexamers (codon pairs). Specifically, codon conservation at wobble positions and codon usage values were computed (Grantham et al., 1981; see Table S1 in the Supplemental Data available with this article online). We used this table as the baseline to search for hexamers that show significant statistical deviation from the expected level. Under the assumption that combinations of two adjacent codons are independent, the expected conservation rate (ECR) and the expected occurrence rate (EOR) of dicodons were computed. The ECR of a dicodon is the multiplication of the corresponding codon conservations and the number of times the dicodon was detected in the data set ($De_{dicodon}$), which was defined as follows: $ECR_{dicodon} = P_{conservation} (1st\ codon) \times P_{conservation} (2nd\ codon) \times De_{dicodon}$. Intuitively, this number reflects the expected frequency for observing a human-mouse conserved dicodon. Similarly, the

EOR of a dicodon is the multiplication of the corresponding codon occurrences and the number of times the pair of amino acids represented by the dicodon was detected in the data set ($De_{amino\ acid}$). This rate was computed using the following equation: $EOR_{dicodon} = P_{occurrence} (1st\ codon) \times P_{occurrence} (2nd\ codon) \times De_{amino\ acid}$. Once the expected dicodon scores were computed, these numbers were contrasted with the actual number of conserved/occurred dicodons (observed conservation rate, OCR; observed occurrence rate, OOR), and we calculated the statistical significance of the difference for each dicodon in terms of occurrence and conservation. For each Z score that was calculated, we assigned a p value (see Supplementary Material on http://www.tau.ac.il/~gilast/sup_mat7.htm for all 3721 calculated hexamers and p value scores). Next, we performed a multiple testing correction by using the false discovery rate (FDR) (Benjamini and Hochberg, 1995) algorithm and extracted those dicodons that were significant. We found 1058 hexamers that were significantly overrepresented and 305 hexamers that were significantly highly conserved in the wobble positions (FDR, $p < 0.0001$ for both criteria), of which 285 hexamers appeared in the intersection of both groups (Table S2).

It has been shown previously that orthologous exons that are alternatively spliced in both human and mouse are more conserved than constitutively spliced ones (Sorek et al., 2004c; Sugnet et al., 2004). In this study, the conservation at each codon position (first, second, and wobble) was evaluated. To this end, we subdivided the human-mouse orthologs database to constitutively (44,731) and alternatively (760) spliced exons (of the exon skipping type), after discarding exons that are constitutively spliced in one organism, but alternatively spliced in the other. Next, we compared the conservation level among the three coding positions between these two subgroups. Although the conservation levels at the first and second codon positions for both types of exons were in the range of 93%–98%, a striking difference in the conservation level at the wobble position was observed: 84% conservation for alternative exons and 72% for constitutive exons (Figure 1A). Similar results were obtained when human exons were compared with seven other vertebrate species (Figure S1). This finding indicates high purifying selection on alternatively spliced exons presumably unrelated to the coding constraints. This may be due to the presence of ESRs and is consistent with previous studies showing selective pressure against synonymous mutations in alternatively spliced exons (Parnley et al., 2006; Xing and Lee, 2005).

Next, we evaluated the spatial conservation of wobble positions across the exon. First, we eliminated from each exon the last three nucleotides and the first nucleotide (which are part of the 5' splice site and 3' splice site consensus sequence, respectively). We divided each exon into ten equally sized deciles. Next, the conservation level of the wobble positions was plotted against the relative position in the exon. We found that the conservation level of the wobble positions between human and mouse exons is higher toward the exon edges (Figure 1B). In order to validate the significance of this result, we compared the edges of the exons (the first and last deciles) to the “middle” (deciles 2–9) for alternatively and constitutively spliced exons. The conservation of

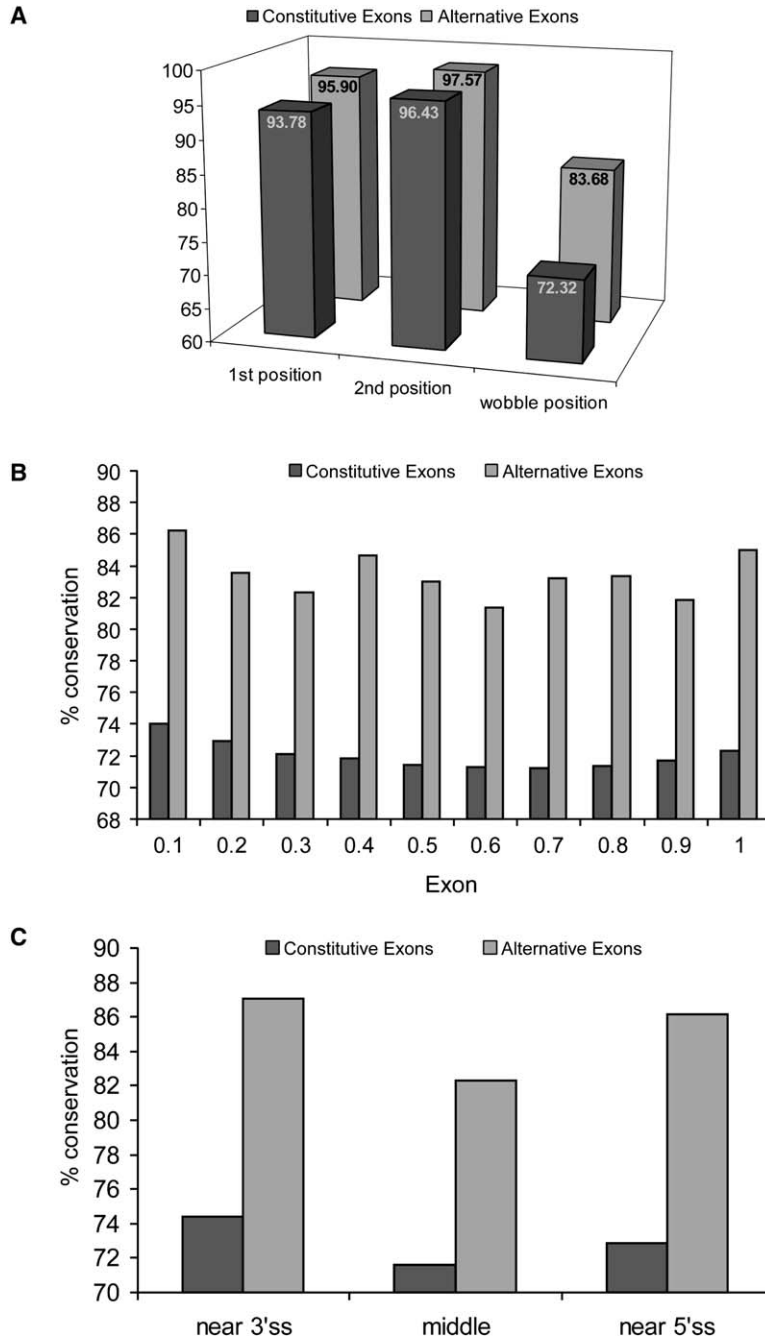


Figure 1. Conservation Level of the Wobble Positions Is Highest Adjacent to Exon Edges
The coding sequence (CDS) of conserved human-mouse orthologous exons was extracted, and the exon pairs were aligned. For each of the human-mouse exon pairs, the conservation level of the wobble positions was calculated using edit distance after removal of the first nucleotide and the last three nucleotides of each exon. (A) The conservation level of each coding position between constitutively and alternatively spliced exons is shown in dark gray and light gray, respectively. X axis, the positions of the codon; Y axis, conservation level of the wobble position between the conserved human-mouse exons. (B) Higher conservation level of the wobble positions located near the splice sites. The conserved human-mouse orthologous exons were extracted, and for each pair the conservation level was calculated using edit distance. X axis, location range from the 3'ss to the 5'ss, in 10% increments (exons are normalized between 0 and 1, see Supplemental Data); Y axis, conservation level of the wobble positions between conserved human-mouse orthologous exons. (C) Higher conservation level of the wobble positions located in the first and last ~12 nucleotides of orthologous human-mouse alternative exons. The first and last four wobble positions (without the contribution of the first and last three nucleotides of each exon) were defined as "near 3'ss" and "near 5'ss," respectively. The remaining internal wobble positions were defined as "middle." The conservation of the human-mouse orthologous exon pairs for the three groups was calculated. X axis, the location definitions; Y axis, conservation level of the wobble positions between conserved human-mouse orthologous exons.

the wobble positions near the 3'ss (first decile) was significantly higher than the middle for both alternative exons (86.2% compared to 82.9%, χ^2 , $p < 0.01$) and constitutive exons (74.1% compared to 71.7%, χ^2 , $p < 0.01$). Similarly, the conservation of the wobble positions near the 5'ss (last decile) was significantly higher than the middle for both alternative exons (85% compared to 82.9%, χ^2 , $p < 0.02$) and constitutive exons (72.3% compared to 71.7%, χ^2 , $p < 0.01$; see Figure 1B).

To further examine this phenomenon, we used a fixed distance from the splice sites. Therefore, the conservation level was calculated for the first four wobble positions and the last four wobble positions (~12 nucleotides) of orthologous human-mouse alternative and

constitutive exons (without the contribution of the first nucleotide and the last three nucleotides of each exon). The first four wobble positions and the last four wobble positions were defined as "near 3'ss" and "near 5'ss," respectively. The remaining internal wobble positions were defined as "middle." The conservation levels of the wobble positions in the human-mouse orthologous exon pairs for the near 3'ss and near 5'ss groups were found to be significantly higher when compared with the middle group in both alternatively and constitutively spliced exons (χ^2 , $p < 0.01$ for both types of exons, Figure 1C). This observation was found to be more pronounced in alternatively spliced exons, compared with constitutively spliced exons. The relative

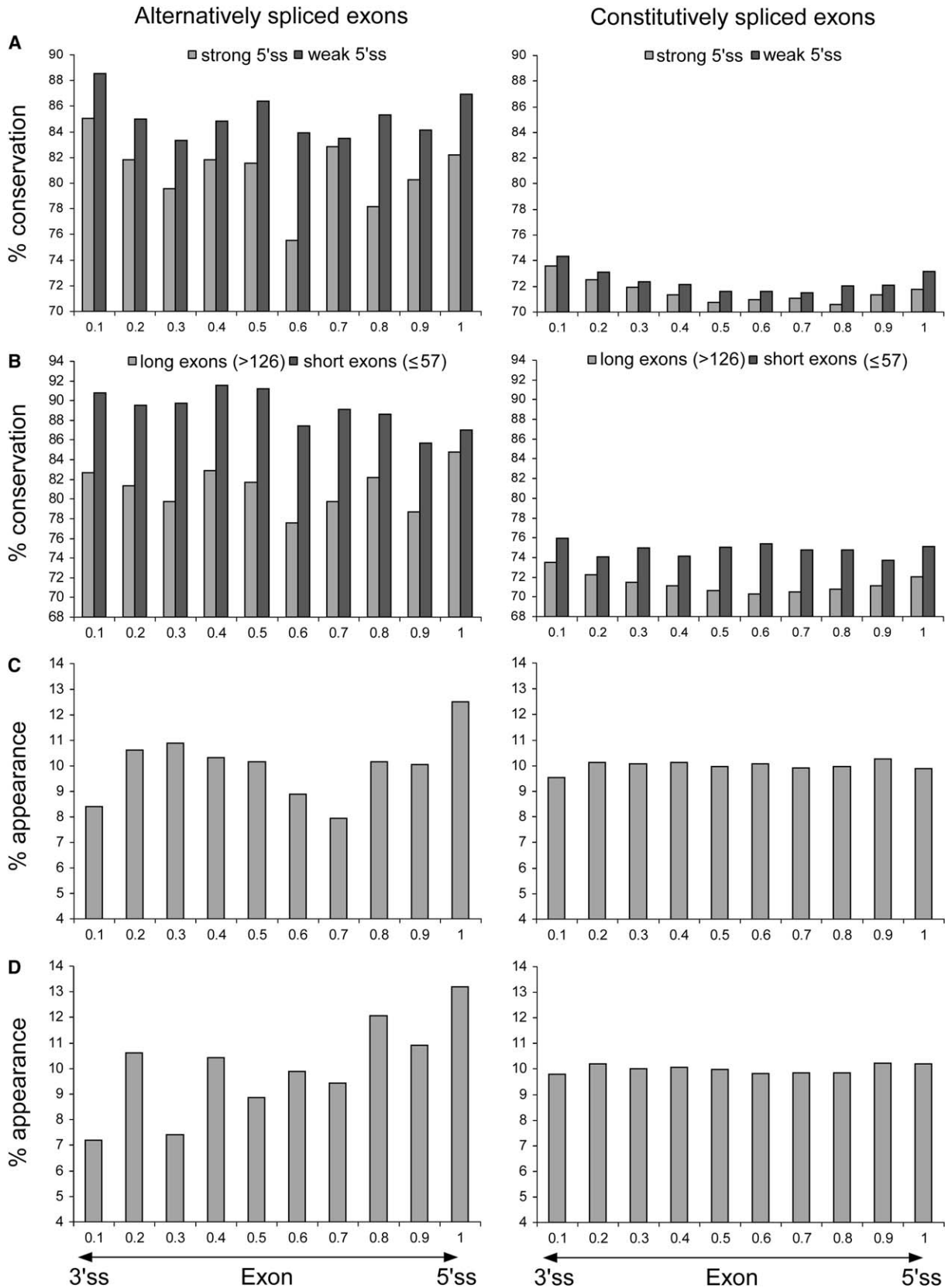


Figure 2. Conservation Level of the Wobble Positions and the Distribution of ESR Candidates along the Exons
The exons were divided into subgroups according to strong and weak 5'ss with respect to U1 binding (A) and short and long exons (B). Left and right panels are alternatively and constitutively spliced exons, respectively. X axis for all panels is the location range from the 3'ss to the 5'ss, in 10% increments (exons are normalized between 0 and 1, see Supplemental Data); Y axis is the conservation level of the wobble positions

decrease in the conservation level of the wobble position in the alternative exons was 4.9% (from 86.6% at the edges to 82.3% in the middle), compared with a relative decrease of 2.7% in constitutively spliced exons (from 73.6% at the edges to 71.6% in the middle).

The Conservation of the Wobble Positions Can Differentiate Unique Characteristics of Alternatively and Constitutively Spliced Exons

It has been shown that there are some characteristics that distinguish between conserved alternatively and constitutively spliced exons, such as the 5'ss strength and exon length (Roca et al., 2005; Carmel et al., 2004; Sorek et al., 2004c; Zheng et al., 2005). These differences presumably indicate that alternatively spliced exons are suboptimal for detection by the splicing machinery (Ast, 2004). We further analyzed the correlation between the conservation in the wobble positions and the 5'ss strength and the correlation between the conservation in the wobble positions and the exon length. In alternatively spliced exons, the level of conservation in wobble positions was found to be higher in exons with a weak 5'ss, compared with exons with a strong 5'ss (Mann-Whitney, $p < 0.01$; Figure 2A, left panel). This trend was observed in constitutively spliced exons as well, although in a milder manner (Mann-Whitney, $p < 0.05$; Figure 2A, right panel). Similarly, the conservation in wobble positions was found to be significantly higher in short exons, compared with long ones, for both alternatively and constitutively spliced exons (Mann-Whitney, $p < 0.01$). However, the difference in the conservation was much higher in alternatively spliced exons than in constitutively spliced exons, with an average difference of $\sim 7.9\%$ and $\sim 3.4\%$, respectively (Figure 2B). This implies that short exons are under high purifying selection.

Different Distributions of ESR Groups along the Alternatively Spliced Exons Compared to the Constitutive Ones

The 285 putative ESRs are a nonhomogenous group in terms of their interacting proteins, binding affinities, etc. Suboptimal alternative exons—those with a weak 5'ss and short alternatively spliced exons—are likely to rely heavily on their ESRs for splicing regulating. Thus, from the 285 putative ESRs, we extracted ESRs that are overrepresented in suboptimal alternative exons. Approximately 10% of the top prevalent ESRs in each subgroup (the weak 5'ss and short exons) of alternatively spliced exons were tested for their distribution across all the alternatively and constitutively spliced exons (26 and 25 ESRs were tested from the weak 5'ss and short alternatively spliced exons, respectively; see Tables S3 and S4). We statistically tested the (null) hypothesis that the chosen hexamers are uniformly distributed along the exons. We divided each exon into ten equally sized deciles and calculated the percentage of the candidates' appearances at each 10% interval. For example, for weak 5'ss, the 26 ESRs appeared 109,691 times in constitutive exons. Therefore, we are able to

calculate the percentage of appearances out of these cases at each interval. A significant variation in the ESRs' distribution was observed across the alternatively and constitutively spliced exons for both ESR groups (χ^2 , $p < 0.01$ for the top 25 short ESRs in alternative exons; $p < 0.02$ for the top 26 weak ESRs in alternative exons; $p < 0.01$ for the constitutive exons in both short and weak subgroups), with a higher level of appearance near the 5'ss of alternatively spliced exons (Figures 2C and 2D, left panels). We further tested the (null) hypothesis that the spatial distributions of the chosen ESR along the exon are the same for alternatively and constitutively spliced exons. The results revealed a significant difference in the distribution of those ESRs between constitutively and alternatively spliced exons (χ^2 , $p < 0.01$ for both selected ESR subgroups). This may indicate the importance of the ESRs' positions along the alternative exons.

Our findings reject the hypothesis that a higher conservation level of alternatively spliced orthologous exons suggests a higher prevalence of ESRs. When we counted the number of appearances of ESRs (for all 285 ESRs) in alternatively and constitutively spliced exons, we found a significantly lower amount of ESRs in the alternatively spliced exons (7,344 total ESRs appeared, out of 66,027 checked hexamers), compared with the constitutively spliced exons (627,741 total ESRs appeared, out of 5,223,101 checked hexamers) (χ^2 , $p < 0.01$). These results are consistent with previous studies (Fairbrother et al., 2002; Wang et al., 2005; Zhang and Chasin, 2004) and therefore imply that the higher conservation level of the alternatively spliced exons, compared with constitutively spliced exons, cannot be explained by a high prevalence of ESRs in alternatively spliced exons.

ESRs Validation

A sample of ten putative ESRs (inferred using the bioinformatics analysis) was selected for experimental evaluation. These ESRs were tested for their effect on splicing, as well as the effect of their positions across the exon on alternative splicing. To this end, we selected the ten putative ESRs for in vivo splicing assays by searching those that are more abundant near the 5'ss (ESRs H-TATGCA and I-ATCCAG); near the 3'ss (ESRs E-TCTGGT, F-ATCATG, and G-ACTGTA); in alternative exons with a weak 5'ss (and near the 5'ss), compared with strong 5'ss; and also in short alternative exons, compared with long alternative ones (ESRs A-AAAGGA, B-AAGAAA, and J-CAAGAA). The last two hexamers were randomly selected from the 285 putative ESRs (ESR C-CTTGTA and D-TTTGCA) (see also Table S5). Briefly, the ESRs were inserted into two different reporter exons by substituting the wt sequence. Those exons are the middle of a genomic sequence containing three exons separated by two introns (minigenes). After transfection of the minigenes to a 293T cell line, total cytoplasmic RNA was extracted, and splicing products were separated on 2% agarose gel after reverse transcription polymerase chain reaction (RT-PCR). The first

between conserved human-mouse orthologous exons. The distribution of selected putative ESRs were chosen according to their prevalence in alternative exons with weak 5'ss (C) and short exons (D) along all alternatively and constitutively spliced exons (left and right panels, respectively). Y axis is the percentage of the candidates' appearance in each 10% interval.

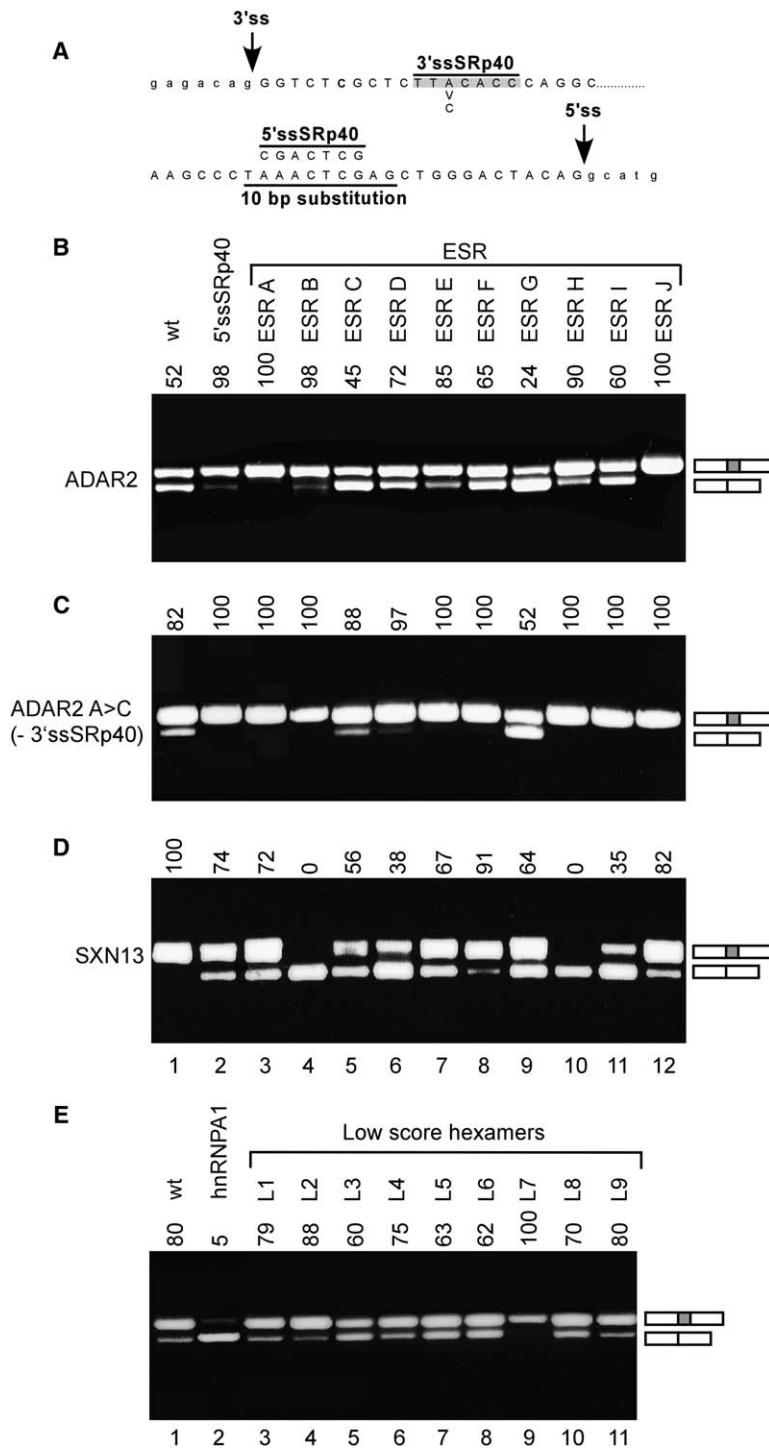


Figure 3. Splicing Assays on ADAR2 and SXN13 Minigene Mutants

(A) Either SRp40 ESR or the indicated putative ESR was inserted 15 bp upstream of the 5'ss in exon 8 of the ADAR2 minigene, by substituting the underlined sequence (marked as 10 bp substitution) with the ten selected ESRs. For the SRp40 binding site, the mutation used to insert that site is shown above the line. The position of the SRp40 binding site near the 3'ss is highlighted in gray and marked "3'ss-SRp40." The mutation (A to C) used to eliminate that site is indicated under the sequence.

(B) Plasmid mutants were introduced into 293T cells by transfection. Total cytoplasmic RNA was extracted, and splicing products were separated in 2% agarose gel (stained with EtBr) after RT-PCR. Lane 1, wt ADAR2; lane 2, insertion of SRp40 RNA binding site; lanes 3–12, insertion of ten different ESR A–J. (C) Similar to (B), except that SRp40 ESR near the 3'ss was mutated, so that the wt is an exon without 3'ssSRp40.

(D) Similar to (B), except for the use of the SXN13 minigene; the ESRs were inserted into the middle of a short exon, 15 nucleotides upstream of the 5'ss.

(E) Similar to (C), except that the hnRNP A1 binding site (TAGGGT in lane 2) or low-scoring hexamers (lanes 3–11) substituted for the wt sequence. The two minigene mRNA products are shown on the right; empty and filled boxes define constitutive and alternative exons, respectively. The inclusion level of the reporter exon is shown above each lane in percent.

reported minigene was obtained from the human genomic ADAR2 gene, exons 7–9, in which exon 8 is alternatively spliced. Exon 8 contains suboptimal 5'ss and 3'ss (Lev-Maor et al., 2003; Sorek et al., 2004a) and was shortened from 120 to 81 nucleotides to further suboptimize the exon. The ESRs were inserted 15 nucleotides upstream from the 5'ss (Figure 3A). Insertion of the SRp40 binding site in that position (CGACTCG marked as 5'ssSRp40 in Figure 3A and with a score of 4.6 according to ESEfinder) changed the inclusion level from

52% to 98%, indicating that this position is sensitive to the effect of ESR on alternative splicing (Figure 3B, compare lanes 1 and 2). All ten tested ESRs have an effect on alternative splicing. Most of them functioned as strong enhancers (Figure 3B, ESRs A, B, E, H, and J), others were moderate enhancers (Figure 3B, ESRs D, F, and I), and two functioned as silencers (Figure 3B, ESRs C and G).

The ten ESRs described above were found to induce different inclusion levels. What determines these

inclusion levels? We characterized another ESR, located near the 3' ss that is predicted to bind SRp40 (Cartegni et al., 2003) (highlighted in gray and marked 3' ssSRp40 in Figure 3A). A point mutation that eliminated that 3' ssSRp40 binding site revealed that it had a silencing effect on the wt and on ESR A–J transcripts (Figure 3C). A comparison of Figure 3B (active 3' ssSRp40 near the 3' ss) and Figure 3C (mutated 3' ssSRp40) shows that this ESR accounts for up to a 43% decrease in alternative splicing. We concluded that, in this exon, both ESRs (near the 3' ss and 5' ss) affect the inclusion level.

When the same 11 ESRs (including SRp40) were inserted into a different reporting exon—SXN13 (Coulter et al., 1997)—that is constitutively spliced, we observed mostly different levels of repression. For example, ESRs B and H are strong ESEs in ADAR2 but strong ESSs in SXN13; and ESR G, which is the strongest ESS in ADAR2, exerts only mild ESS effect in SXN13 (compare Figure 3D with Figures 3B and 3C). Similar results were also obtained when the minigenes were inserted in HeLa cells (data not shown). This shows that the “ESR environment” (the intron/exon structure and possible interacting ESRs/ISRs) affects the inclusion level considerably.

Is it possible to evaluate how many functional ESRs are not detected by our analysis? To address this issue, we repeated the analysis, shown in Figure 3C, by substituting the wt sequence with the hnRNP A1 binding site TAGGGT (Burd and Dreyfuss, 1994) and with nine hexamers that are not part of the 285 ESRs (see Supplemental Data for those sequences, marked L1–L9). We ensured that the substitution of the wt sequence with that of the hexamers did not create ESE/ESS in the newly created junctions between the hexamers and the wt sequence, according to previous studies (Cartegni et al., 2003; Fairbrother et al., 2002; Wang et al., 2004; Zhang and Chasin, 2004) and our 285 ESRs. As expected, the hnRNP A1 binding site exerts an ESS activity (Figure 3E, lane 2). Most of the low-score hexamers have a mild effect on splicing (up to 20%), and this is with respect to the strong effect observed by the ten representative ESRs (compare Figure 3C and 3E). However, hexamer L7 functions as a strong ESE (Figure 3E, lane 9), suggesting that one out of nine hexamers, not identified so far by this and other studies (Cartegni et al., 2003; Fairbrother et al., 2002; Wang et al., 2004; Zhang and Chasin, 2004), may be functional ESRs. The mild effect of most of the low-score hexamers on splicing may be because the demands we imposed on the bioinformatic analysis (high conservation and prevalence) select only strong and prevalent sites. However, this analysis does not select functional sequences with a low binding affinity or with low prevalence. Regarding hexamer L7 that functions as a strong ESE and is not part of ESEs/ESSs identified previously (Cartegni et al., 2003; Fairbrother et al., 2002; Wang et al., 2004; Zhang and Chasin, 2004), it may be that this is a human-specific enhancer that evolved after the split between the human and mouse lineages and thus is not covered by our analysis. Alternatively, it shows the limitations of each of the bioinformatic methodologies used so far in identifying all ESRs.

To examine the functionality of ESRs in their natural environment, we scanned for putative ESRs in three re-

porting exons (SLC35B3, IMP, and Talin; see also Zhang et al. [2005]). In SLC35B3, we replaced CCCAAG ESR (marked K in Figure 4A) with a hexamer that is not one of the 285 ESRs (GTAAAG). In IMP, we generated a single point mutation from ATCATC to ATTATC that eliminated ESR M, whereas ESR sequences L and N were deleted (Figure 4A). Those ESRs are not shared by previous compiled data sets, and the mutations did not create an overlapping known ESR (according to Cartegni et al. [2003], Fairbrother et al. [2002], Wang et al. [2004], Zhang and Chasin [2004], and our 285 ESRs). Under the assumption that the splicing of alternative exons is more affected by regulatory sequences than constitutively spliced exons, we used a mutant from A to G in the branch site sequence of the IMP minigene that shifted the splicing of the relevant exon from constitutive to alternative (Kol et al., 2005). Figure 4B shows that ESRs K, M, and N function as ESEs, whereas ESR L functions as an ESS. We therefore concluded that those ESRs are functional in their natural environment. To examine the specificity of the ESRs' effect, we either substituted random hexamers or heptamers from the SLC35B3 and the Talin reporting exons with hexamers that are not part of the 285 ESRs, or we deleted presumably nonfunctional sequences (Figure 4C). Those mutations have no or minor effect on splicing, reinforcing the confidence that the 285 ESRs are bona fide functional sequences.

Position-Dependent Effect of ESRs on Alternative Splicing

The different effect of the same ESRs between the two reporting exons (Figure 3) may be explained by a different relative ESR location along the exon. Furthermore, this spatial effect was suggested by the bioinformatic analysis as well, which demonstrated both a higher purifying selection toward the edges of the suboptimal alternative exons and an ESR distribution that is not uniform along the alternative exons (Figure 2). Therefore, we examined whether the distinct location of an ESR along the exon may exert different effects on alternative splicing. To this end, we slid two known SR binding sites—SF2/ASF and SRp40—along exon 8 of ADAR2 (Figure 5A). Remarkably, the different positions of these two SR binding sites along the exon exert different effects on the alternative splicing pattern. SF2/ASF ESR functions mostly as an enhancer, and SRp40 functions mostly as a silencer (Figures 5B and 5C). Although SRp40 typically functions as a silencer, in seven cases it was found to function as a strong or moderate enhancer. Similarly, although SF2/ASF mostly functions as an enhancer, in ten cases it was found to function as a strong or mild silencer. In addition, it was demonstrated that there are significant variations in the enhancement/silencing level as a function of the location of both SF2/ASF and SRp40 binding sites along the exon (Figure 5D). As before, when the SRp40 ESR binding site near the 3' ss was restored, a repression effect was observed for most transcripts, excluding positions Q and T for both SF2/ASF and SRp40, and positions R and O for SF2/ASF (compare Figures 5B and 5C to Figure 5E). This strengthens the notion that more than one ESR regulates each splicing event (Smith and Valcarcel, 2000).

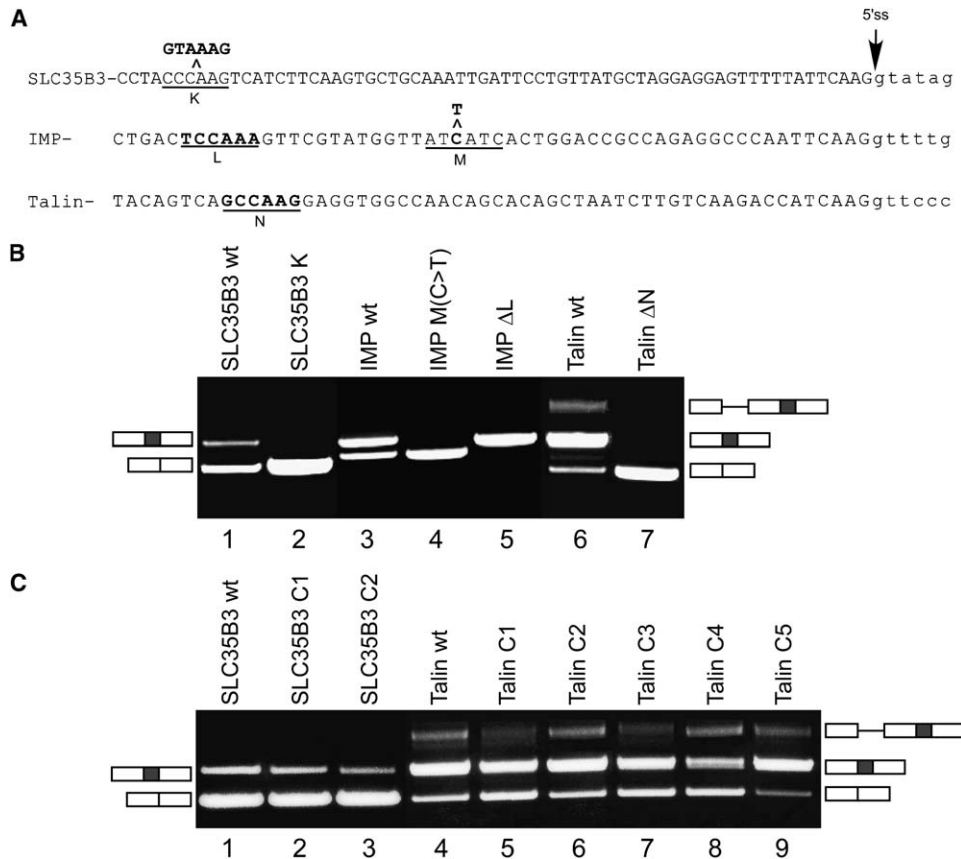


Figure 4. ESRs Are Functional in Their Natural Environment

(A) The sequence of the exonic regions of SLC35B3, IMP, and Talin that were mutated is shown. Four ESRs were mutated in those exons (underlined K–N). ESR K was substituted by a low-score hexamer (GTAAAG), ESRs L and N were deleted, and ESR M was eliminated by a C-to-T mutation. (B) Wt and mutant plasmids were treated as in Figure 3. Lane 1, wt SLC35B3; lane 2, SLC35B3 after substitution of ESR K with a low-score hexamer; lane 3, wt IMP; lane 4, mutation of C to T in ESR M; lane 5, deletion of ESR L; lane 6, wt Talin; lane 7, deletion of ESR N. The two minigene mRNA products of SLC35B3 and IMP are shown on the left; the three minigene products of Talin are shown on the right. Empty and filled boxes define constitutive and alternative exons, respectively. Intron retention product is marked by a line. (C) The specificity of the ESRs' effect. Wt and mutant plasmids were treated as in (B). Lane 1, wt SLC35B3; lanes 2 and 3, TGAATT was substituted with CGCTCT; lane 3, TGAATT was deleted; lanes 4–9, Talin; lane 4, wt; lane 5, GCACTGT was substituted with CAGATCA; lane 6, GCACTGT was deleted; lane 7, GTACCA was substituted with CGCTGT; lane 8, TAATCT was substituted with ACTTTT; lane 9, TAATCT was deleted. No substituted sequences are part of known ESRs.

It may be claimed that the spatial effect of the ESRs is an indirect effect due to the elimination/creation of other functional ESRs along the exon. We therefore eliminated some of the created SF2/ASF and SRp40 binding sites by point mutations and examined the effect on splicing (Figure 5G). All point mutations restored the approximate inclusion level of the wt transcript. We therefore concluded that the positional-dependent effect is specific.

To gain more confidence that, indeed, ESRs located in different positions along the exon can exert different effects on splicing, we selected five putative ESRs and inserted them in the ADAR2 minigene next to the 3'ss and in the middle of the exon (Figure 6A, positions α and β , respectively). Figure 6A shows that the same ESR can have different effects on alternative splicing, depending on its position along the exon. For example, ESRs B, D, G, and I exert different effects on alternative splicing in the three positions (compare Figure 6A, lanes 2–5 and lanes 7–10, and Figure 3C, lanes 4, 6, 9, and 11 for the same ESRs near the 3'ss, in the middle of the exon, and near the 5'ss, respectively).

To examine whether the position-dependent effect is a general phenomenon, we used another minigene (SLC35B3) in which exon 5 is skipped in 87% of the mRNA products (Figure 6B, lane 1). We inserted different ESRs into two different positions of the alternatively spliced exon of that minigene, by substitution of the wt sequence (Figure 6B, positions γ and δ). The alternative exon contains SRp40 ESR (CTTCAAG) in position γ that has a binding score of 4 (Cartegni et al., 2003). A point mutation that eliminated that ESR led to 100% exon skipping, indicating that it functions as a weak ESE (Figure 6B, compare lane 1 with lane 11). Replacing that SRp40 ESR with the SF2/ASF binding site (CGGAGGA, with a score of 5.3) or with our ESRs named B, D, or I, led to 60%, 5%, 0%, and 21% inclusion levels, respectively (Figure 6B, lanes 2–5). We then moved to another position in that exon, located 14 nucleotides upstream of the 5'ss (termed position δ). Substitution of the wt sequence with that of SRp40 ESR after a point mutation that eliminated the SRp40 ESR in position γ led to 56% inclusion, indicating that SRp40 in two different

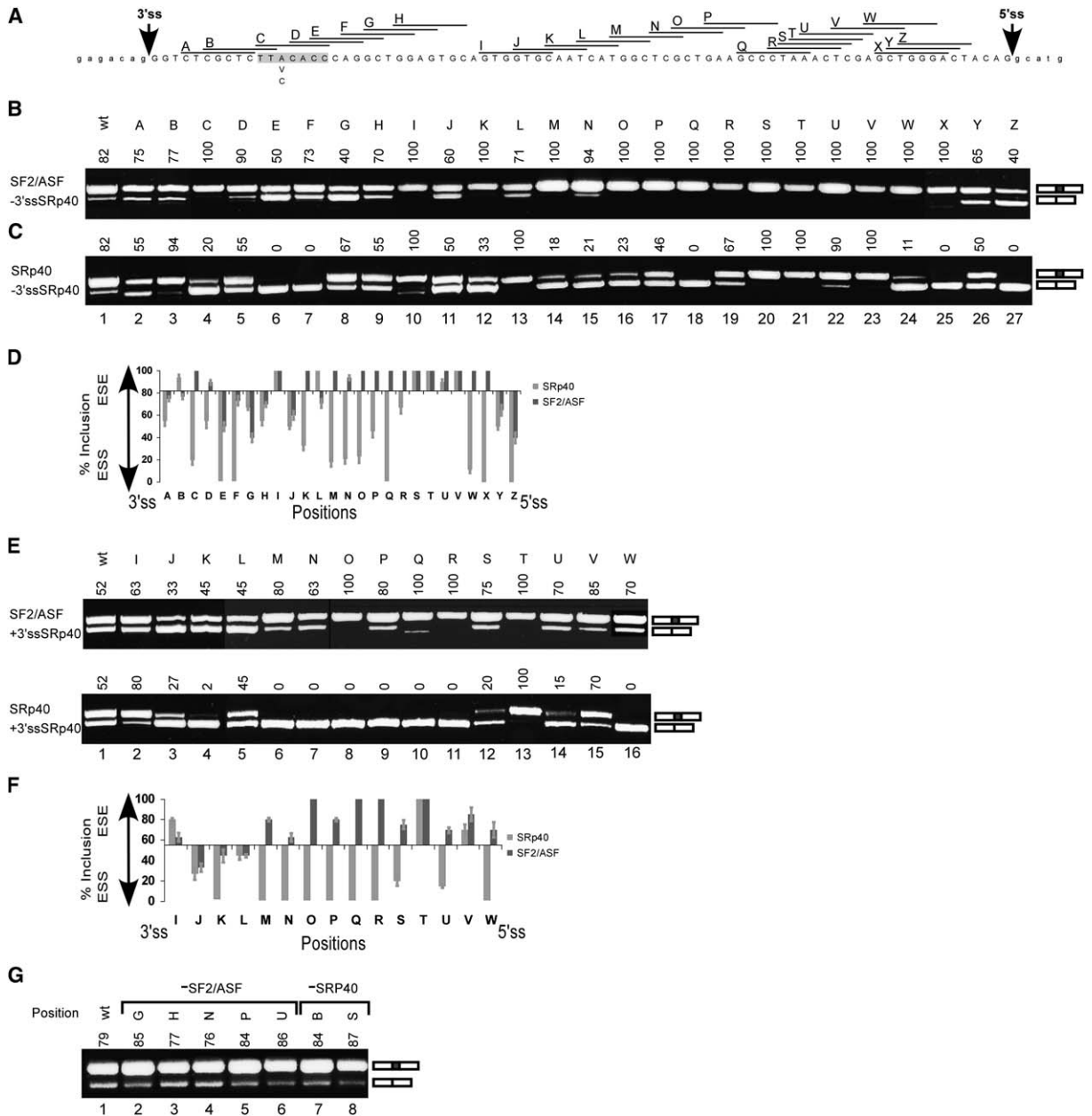


Figure 5. Position-Dependent Effect of ESRs on Alternative Splicing

(A) Sequence of exon 8 of ADAR2 with the 3' ss and 5' ss marked by arrows; SRp40 near the 3' ss is highlighted in gray, A-to-C mutation used to eliminate that site is shown; 26 positions of SF2/ASF (CTCACGA) or SRp40 (TCACTGC) ESRs replaced the wt sequence and are indicated by A–Z lines above the sequence.

(B) In vivo splicing assay as in Figure 3. Lane 1, ADAR2 with mutated SRp40 near the 3' ss; lanes 2–27, splicing products with SF2/ASF ESR corresponding to A–Z positions.

(C) Similar to (B), but with SRp40 ESR sliding along the exon.

(D) Quantifications of SF2/ASF and SRp40 effects on splicing.

(E) Similar to (B) and (C), except the presence of SRp40 binding site near the 3' ss.

(F) Quantifications of the SF2/ASF and SRp40 effects on splicing from (E).

(G) Similar to (B) and (C), except that SF2/ASF or SRp40 binding sites in the indicated positions were mutated: SF2/ASF to ATCACGA, and SRp40 to TCGATGC. Error bars represent the range of inclusion level of three independent experiments.

positions in that same exon exerts a different effect on splicing—a weak and a strong ESE activity in position γ and δ , respectively (Figure 6B, lanes 1 and 7). Similarly, ESR I, which is a weak ESE in position γ , is a strong ESE in position δ . Although ESRs B and D have a similar effect in both positions (Figure 6B, compare lanes 3–5 to

8–10), the combination of two ESRs in the two exonic positions revealed that when SRp40 and a strong SF2/ASF ESR are located in positions γ and δ , respectively, they lead to 68% exon inclusion. However, placing the two ESRs in the opposite positions led to 100% inclusion (Figure 6B, lanes 6 and 12). This confirmed that

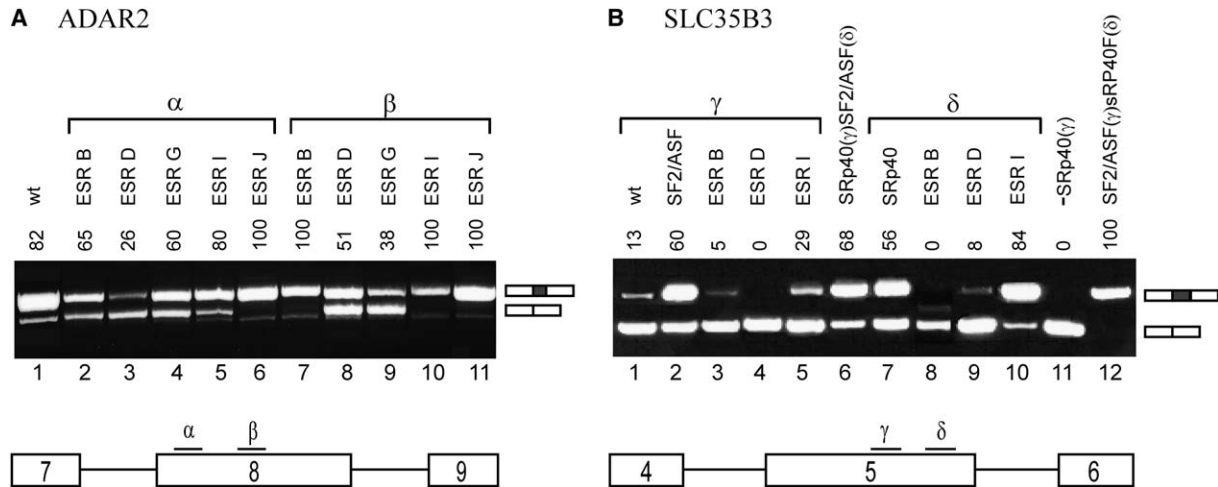


Figure 6. The Position-Dependent Effect Is a General Phenomenon
(A) The effects on splicing of five ESRs that substituted positions α and β of the ADAR2 minigene.
(B) Similar *in vivo* splicing assays as in **(A)**, using the SLC35B3 minigene. The indicated ESRs were substituted in two exonic positions (marked as γ and δ). On the bottom of that panel, a schematic drawing of the two minigenes and their substituted exonic positions is shown. In ADAR2, positions α and β begin 72 and 40 nucleotides upstream of the 5'ss. In SLC35B3, positions γ and δ begin 47 and 14 nucleotides upstream of the 5'ss (see Supplemental Data).

the position-dependent effect of ESRs on alternative splicing is a general phenomenon.

Discussion

We have developed a computational method that identifies putative regulatory sequences, based on the combination of two features: (1) the evolutionary conservation of wobble positions between human and mouse orthologous exons and (2) the analysis of the overabundance of sequence motifs, compared with their random expectation, due to their codon relative frequency. This method resulted in 285 significant motifs; more than 60% of these are unknown with respect to other ESR compiled data sets (Fairbrother et al., 2002; Wang et al., 2004; Zhang and Chasin, 2004). We experimentally validated our putative ESRs by choosing ten motifs and placing them in two suboptimal exons (minigenes). We also chose four ESRs and demonstrated their effects on splicing in their natural environment. Remarkably, we found distinctions in the inclusion/exclusion levels between the two minigenes mediated by the same candidate—the same sequence can function as an enhancer in one exon and a silencer in the other. It may be that, in some of the cases tested in Figures 3–5, the effect on the inclusion/exclusion level was due to elimination or creation of other regulatory sequences. However, the insertion of low-score hexamers (Figures 3E and 4C) and the restoration of the wt inclusion level after a point mutation that eliminated the SR binding site (Figure 5E) strongly support the observations that the position-dependent effect is specific and that the definition of a regulatory sequence as a strong/weak enhancer or silencer cannot be predicted based on the sequence of the ESR alone.

We and others have recently demonstrated that alternatively spliced exons contain several characteristics

that distinguish them from constitutively spliced exons, such as the 5'ss strength and the exon length. These characteristics demonstrate that alternatively spliced exons are generally suboptimal for spliceosome recognition and are under different selective pressures from constitutively spliced ones. Our results suggest that the higher conservation level of the wobble position adjacent to splice sites may be a result of selective pressure that is required to preserve the ESRs' ability to assist in the placement of U1 and U2AF on the correct 5'ss and PPT, respectively. This is in agreement with previous studies showing that selective pressure is stronger on ESEs located in proximity to the splice sites (Fairbrother et al., 2004; Sugnet et al., 2004).

Although synonymous mutations have no effect on the encoded protein, about 40% of such mutations have been opposed by selection, and some of those mutations are involved in diseases (Chamary et al., 2006). It was previously estimated that 90% of synonymous mutations are under weak selection but that the selection is so weak that it has a negligible effect on substitution rate (Lu and Wu, 2005). Therefore, several models were proposed to explain this phenomenon, such as the following: selective force to maintain a certain secondary structure in the RNA; the preference of particular nucleotides at the wobble position that optimized base pairing with the cognate tRNA during translation; prevention of cotranslational misfolding; transcription regulatory sequences, such as pausing sequences of RNA polymerase II; and other options as well (Chamary et al., 2006). Here we examined the possibility that some of the selective force on the wobble positions is due to the presence of ESRs. The findings that ~30% of the synonymous mutations in exon 12 of the cystic fibrosis gene affect the splicing efficiency (Pagani et al., 2005) and that the mutation in a synonymous site that abolishes ASF/SF2 ESR and generates hnRNP A1 ESR in

SMN2 causes spinal muscular atrophy (Cartegni and Krainer, 2002; Kashima and Manley, 2003), along with the results presented in this manuscript, strongly suggest that the presence of ESRs along constitutively and alternatively spliced exons may account for a large fraction of the selective force on synonymous mutations, especially in alternatively spliced exons.

Most of the information for the ESRs is obtained from constitutively spliced exons. So why does the above method select both enhancers and silencers? Presumably because enhancers and silencers are part of the overall signals and factors that determine exon selection for both constitutively and alternatively spliced exons, and those are precisely the same signals that regulate the inclusion level in alternative splicing. In most of the experiments, we did find an independent effect, namely that the outcome of the effect of two binding sites equals the sum of the effect of each one (a linear effect). However, in several experiments, such as positions Q and T for both SF2/ASF and SRp40, positions R and O for SF2/ASF (compare Figures 5B and 5C to Figure 5E), and the switching between the binding sites of SRp40 and SF2/ASF in Figure 6B, we did detect a nonlinear effect, which suggests a combinatorial effect in these cases (see also Smith and Valcarcel [2000]).

It is intriguing that ESRs are more conserved in alternative exons, despite their higher abundance in constitutive exons. This suggests that constitutive exons can afford changes in these elements because their proper splicing is ensured by multiple redundant signals (such as strong splice sites), whereas alternatively spliced exons are more sensitive because of the more delicate balance between signals leading to alternative patterns of processing. The average human exon size is about 129 bp long. This may explain why small alternatively spliced exons are far from being defined as ideal exons for the exon definition mechanism (Ast, 2004; Izquierdo et al., 2005). Therefore, these small exons are probably under high purifying selection to conserve the ESRs, especially those located at the ends of the exons, because they are crucial for exon recognition.

Our results indicate that the positional-dependent effect is a general mechanism in the regulation of splicing. This effect is consistent with previous studies indicating that the position of SR-protein binding sites in pre-mRNA determines whether these sites function as enhancers or silencers in splicing (Buratti et al., 2004; Han et al., 2005; Kanopka et al., 1996).

The transitions from intron to exon definition during evolution presumably led to the selection of proteins that bind exonic sequences and assist the splicing machinery to correctly recognize the exon/intron junction. The edges of the suboptimal alternatively spliced exons show the highest conservation levels, indicating strong selective pressures acting on those ESRs. The ESRs that are located in different positions along the exon may enhance or repress the binding of the basal splicing machinery to the appropriate splice site. Thus, we suggest that the definition between ESE and ESS, depending on the type of protein that binds it, should be changed to ESR sequence in which the function of those sequences as enhancers or silencers also depends on their relative positions within the exons, on spatial effects, and on other combinatorial effects.

Experimental Procedures

Search Engine

ESR-Search is a web-based resource to identify those and other ESRs (<http://ast.bioinfo.tau.ac.il/ESR.htm>).

Data Set Compilation

We compiled a unified data set of human-mouse orthologous exons (Carmel et al., 2004; Sorek et al., 2004b; Sugnet et al., 2004). We filtered out exons that are present more than once in the unified data set, exons that were longer than 250 nucleotides, exons that did not present the canonical AG-GT splice site, and human-mouse exon pairs in which the exons differ in size. This left us with 46,103 pairs of exons. As the database was compiled using EST sequences, we blasted all exons against RefSeq (Pruitt et al., 2005) to find the reading frame of each exon for further analysis. For the definition of conserved alternative skipped exon events, we demanded at least four EST/cDNAs that skip and four that include the internal exon for the human exon, and one EST/cDNA that contains and one that skips the exon for the mouse, which is the same definition for the exons obtained from Sugnet et al. (2004).

5' ss Strength and Subgroup Definitions

The criteria we used to divide exons into subgroups were as follows: (1) the strength of the 5' ss and (2) the length of the exon. The strength of the 5' ss was defined as the prediction of free energy (ΔG) released from U1/5' ss base pairing and was calculated using the Oligo-ArrayAux 1.9 (Rouillard et al., 2002). See Supplemental Data for more information. Exons were divided into groups according to their predicted ΔG scores. The 25% of exons with the lowest and highest ΔG scores were defined as strong and weak 5' ss, respectively (alternative low ΔG , namely, strong 5' ss, < -7.3 kcal/mol; alternative high ΔG , namely weak 5' ss, > -4.2 kcal/mol; constitutive low ΔG , ≤ -7.6 kcal/mol; constitutive high ΔG , ≥ -4.2 kcal/mol). To compare the differences between long and short exons for alternative and constitutive exons statistically, we subdivided the constitutive and alternative exons, this time based on their length: short exons are defined as ≤ 57 nucleotides; long exons are defined as > 126 nucleotides. These length thresholds were chosen because they correspond to the top and bottom quartiles of length in alternative exons. Defining the length thresholds based on constitutive exons yielded similar results (Figure S2).

ESR Insertions

The ADAR2 minigene was previously cloned (Lev-Maor et al., 2003) to PEGFP-C3 vector containing the CMV-IE promoter and two polyadenylation sites in positions 1546 and 1575 (AATAAA). The exon we used was shortened from the wt 120 bases to 81 (Supplemental Data).

In Figures 3 and 5, the different SR or ESR binding site substitutions were inserted by a primer tail of the forward primer. The SLC35B3 minigene is the GenBank term with the accession number NM015948. It was cloned into PEGFP-C3 vector from exons 4–6. See the reporting exon 5 sequence in the Supplemental Data.

The SRp40 binding site is the wt sequence CTCAAG that is part of position γ . The SF2/ASF sequence was created by a point mutation from AGGAGGA to CGGAGGA in position δ . The SXN13 minigene was a kind gift from Thomas A. Cooper (Coulter et al., 1997). We selected ten ESRs that were most prevalent in our data set (referred to ESR A–J; see also Table S5). Each hexamer is flanked with 2 bp in each side (see also Wang et al. [2004]).

Supplemental Data

Supplemental Data include three figures, five tables, Supplemental Experimental Procedures, and Supplemental References and can be found with this article online at <http://www.molecule.org/cgi/content/full/22/6/769/DC1/>.

Acknowledgments

We thank Javier F. Caceres, Jeremy Sanford, and Eddo Kim for their critical reading. We thank Chuck W. Sugnet, David Haussler, Ido Carmel, and Rotem Sorek for sharing data sets. We thank Adi Doron-Faigenboim for the FDR test program. This work was supported by a grant from the Israel Science Foundation (1449/04 and

40/05), MOP Germany-Israel, GIF, ICA through the Ber-Lehmsdorf Memorial Fund, and DIP. O.R and A.G. are supported by EURASNET and Safra Fellowship, respectively.

Received: December 12, 2005

Revised: April 6, 2006

Accepted: May 3, 2006

Published: June 22, 2006

References

- Ast, G. (2004). How did alternative splicing evolve? *Nat. Rev. Genet.* **5**, 773–782.
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. 57*(Series A), 289–300.
- Black, D.L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* **72**, 291–336.
- Blanchette, M., and Tompa, M. (2002). Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.* **12**, 739–748.
- Blencowe, B.J. (2000). Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem. Sci.* **25**, 106–110.
- Buratti, E., Baralle, M., De Conti, L., Baralle, D., Romano, M., Ayala, Y.M., and Baralle, F.E. (2004). hnRNP H binding at the 5' splice site correlates with the pathological effect of two intronic mutations in the NF-1 and TSHbeta genes. *Nucleic Acids Res.* **32**, 4224–4236.
- Burd, C.G., and Dreyfuss, G. (1994). RNA binding specificity of hnRNP A1: significance of hnRNP A1 high-affinity binding sites in pre-mRNA splicing. *EMBO J.* **13**, 1197–1204.
- Carmel, I., Tal, S., Vig, I., and Ast, G. (2004). Comparative analysis detects dependencies among the 5' splice-site positions. *RNA* **10**, 828–840.
- Cartegni, L., and Krainer, A.R. (2002). Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. *Nat. Genet.* **30**, 377–384.
- 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.
- Cartegni, L., Wang, J., Zhu, Z., Zhang, M.Q., and Krainer, A.R. (2003). ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res.* **31**, 3568–3571.
- Chamary, J.V., Parmley, J.L., and Hurst, L.D. (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* **7**, 98–108.
- Coulter, L.R., Landree, M.A., and Cooper, T.A. (1997). Identification of a new class of exonic splicing enhancers by in vivo selection. *Mol. Cell Biol.* **17**, 2143–2150.
- Fairbrother, W.G., Yeh, R.F., Sharp, P.A., and Burge, C.B. (2002). Predictive identification of exonic splicing enhancers in human genes. *Science* **297**, 1007–1013.
- Fairbrother, W.G., Holste, D., Burge, C.B., and Sharp, P.A. (2004). Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol.* **2**, e268. 10.1371/journal.pbio.0020268.
- Fedorov, A., Saxonov, S., Fedorova, L., and Daizadeh, I. (2001). Comparison of intron-containing and intron-lacking human genes elucidates putative exonic splicing enhancers. *Nucleic Acids Res.* **29**, 1464–1469.
- Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., and Mercier, R. (1981). Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.* **9**, r43–r74.
- Graveley, B.R. (2000). Sorting out the complexity of SR protein functions. *RNA* **6**, 1197–1211.
- Graveley, B.R. (2001). Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* **17**, 100–107.
- Han, K., Yeo, G., An, P., Burge, C.B., and Grabowski, P.J. (2005). A combinatorial code for splicing silencing: UAGG and GGGG motifs. *PLoS Biol.* **3**, e158. 10.1371/journal.pbio.0030158.
- Hartmuth, K., Urlaub, H., Vormlocher, H.P., Will, C.L., Gentzel, M., Wilm, M., and Luhrmann, R. (2002). Protein composition of human pre-spliceosomes isolated by a tobramycin affinity-selection method. *Proc. Natl. Acad. Sci. USA* **99**, 16719–16724.
- Izquierdo, J.M., Majos, N., Bonnal, S., Martinez, C., Castelo, R., Guigo, R., Bilbao, D., and Valcarcel, J. (2005). Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. *Mol. Cell* **19**, 475–484.
- Jurica, M.S., and Moore, M.J. (2003). Pre-mRNA splicing: awash in a sea of proteins. *Mol. Cell* **12**, 5–14.
- Kanopka, A., Muhlemann, O., and Akusjarvi, G. (1996). Inhibition by SR proteins of splicing of a regulated adenovirus pre-mRNA. *Nature* **381**, 535–538.
- Kashima, T., and Manley, J.L. (2003). A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. *Nat. Genet.* **34**, 460–463.
- Kol, G., Lev-Maor, G., and Ast, G. (2005). Human-mouse comparative analysis reveals that branch-site plasticity contributes to splicing regulation. *Hum. Mol. Genet.* **14**, 1559–1568.
- 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.
- Lim, L.P., and Burge, C.B. (2001). A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl. Acad. Sci. USA* **98**, 11193–11198.
- Lu, J., and Wu, C.I. (2005). Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. *Proc. Natl. Acad. Sci. USA* **102**, 4063–4067.
- Maniatis, T., and Tasic, B. (2002). Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418**, 236–243.
- Nilsen, T.W. (2003). The spliceosome: the most complex macromolecular machine in the cell? *Bioessays* **25**, 1147–1149.
- Pagani, F., Raponi, M., and Baralle, F.E. (2005). Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proc. Natl. Acad. Sci. USA* **102**, 6368–6372.
- Parmley, J.L., Chamary, J.V., and Hurst, L.D. (2006). Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol. Biol. Evol.* **23**, 301–309. Published online October 12, 2005.
- Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**, D501–D504.
- Roca, X., Sachidanandam, R., and Krainer, A.R. (2005). Determinants of the inherent strength of human 5' splice sites. *RNA* **11**, 683–698.
- Rouillard, J.M., Herbert, C.J., and Zuker, M. (2002). OligoArray: genome-scale oligonucleotide design for microarrays. *Bioinformatics* **18**, 486–487.
- Sanford, J.R., Ellis, J., and Caceres, J.F. (2005). Multiple roles of arginine/serine-rich splicing factors in RNA processing. *Biochem. Soc. Trans.* **33**, 443–446.
- Singh, R., and Valcarcel, J. (2005). Building specificity with non-specific RNA-binding proteins. *Nat. Struct. Mol. Biol.* **12**, 645–653.
- Smith, C.W., and Valcarcel, J. (2000). Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci.* **25**, 381–388.
- Sorek, R., Lev-Maor, G., Reznik, M., Dagan, T., Belinky, F., Graur, D., and Ast, G. (2004a). Minimal conditions for exonization of intronic sequences: 5' splice site formation in alu exons. *Mol. Cell* **14**, 221–231.
- Sorek, R., Shamir, R., and Ast, G. (2004b). How prevalent is functional alternative splicing in the human genome? *Trends Genet.* **20**, 68–71.

- Sorek, R., Shemesh, R., Cohen, Y., Basechess, O., Ast, G., and Shamir, R. (2004c). A non-EST-based method for exon-skipping prediction. *Genome Res.* *14*, 1617–1623.
- Sugnet, C.W., Kent, W.J., Ares, M., Jr., and Haussler, D. (2004). Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac. Symp. Biocomput.*, 66–77.
- Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M., and Burge, C.B. (2004). Systematic identification and analysis of exonic splicing silencers. *Cell* *119*, 831–845.
- Wang, J., Smith, P.J., Krainer, A.R., and Zhang, M.Q. (2005). Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes. *Nucleic Acids Res.* *33*, 5053–5062.
- 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.
- Xing, Y., and Lee, C. (2005). Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc. Natl. Acad. Sci. USA* *102*, 13526–13531.
- Zhang, X.H., and Chasin, L.A. (2004). Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev.* *18*, 1241–1250.
- Zhang, X.H., Kangsamaksin, T., Chao, M.S., Banerjee, J.K., and Chasin, L.A. (2005). Exon inclusion is dependent on predictable exonic splicing enhancers. *Mol. Cell. Biol.* *25*, 7323–7332.
- Zheng, C., Fu, X., and Gribkov, M. (2005). Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. *RNA* *11*, 1777–1787.
- Zhou, Z., Licklider, L.J., Gygi, S.P., and Reed, R. (2002). Comprehensive proteomic analysis of the human spliceosome. *Nature* *419*, 182–185.

Accession Numbers

The GenBank accession number for the SLC35B3 sequence reported in this paper is NM015948.