Cell, Vol. 119, 831–845, December 17, 2004, Copyright ©2004 by Cell Press

Systematic Identification and Analysis of Exonic Splicing Silencers

Zefeng Wang,[1] Michael E. Rolish,[1,2] Gene Yeo,[1,3] Vivian Tung,[1] Matthew Mawson,[1] and Christopher B. Burge[1,*]
[1]Department of Biology
[2]Department of Electrical Engineering and Computer Science
[3]Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

Summary

Exonic splicing silencers (ESSs) are *cis*-regulatory elements that inhibit the use of adjacent splice sites, often contributing to alternative splicing (AS). To systematically identify ESSs, an in vivo splicing reporter system was developed to screen a library of random decanucleotides. The screen yielded 141 ESS decamers, 133 of which were unique. The silencer activity of over a dozen of these sequences was also confirmed in a heterologous exon/intron context and in a second cell type. Of the unique ESS decamers, most could be clustered into groups to yield seven putative ESS motifs, some resembling known motifs bound by hnRNPs H and A1. Potential roles of ESSs in constitutive splicing were explored using an algorithm, ExonScan, which simulates splicing based on known or putative splicing-related motifs. ExonScan and related bioinformatic analyses suggest that these ESS motifs play important roles in suppression of pseudoexons, in splice site definition, and in AS.

Introduction

Most human genes are transcribed as precursors containing long intervening segments that are removed in the process of pre-mRNA splicing. The specificity of splicing is defined in part by splice site and branch site sequences located near the 5′ and 3′ ends of introns. However, even considering transcripts with only very short introns, these sequences contain only about half of the information required for accurate recognition of exons and introns in human transcripts (Lim and Burge, 2001). Indeed, it is well known that human introns contain many sequences that match the consensus splice site motifs as well as authentic sites, yet are virtually never used in splicing; such sequences are often called "decoy" splice sites (Cote et al., 2001). Some of these decoy splice sites occur in pairs with a potential 3′ splice site (3′ss) followed by a potential 5′ splice site (5′ss) with spacing typical of exons and yet are still rarely if ever recognized by the splicing machinery; such splice site pairs are commonly referred to as "pseudoexons" (Senapathy et al., 1990; Sun and Chasin, 2000). The ability of the splicing machinery to reliably distinguish authentic exons and splice sites from a large excess of these imposters implies that sequence features outside of the canonical splice site/branch site elements must play important roles in splicing of most or all transcripts. Prime candidates for these features are exonic or intronic *cis*-elements that either enhance or silence the usage of adjacent splice sites.

Two major classes of *cis*-regulators of splicing are the exonic splicing enhancers (ESEs) and exonic splicing silencers (ESSs). These elements generally function by recruiting protein factors that interact favorably or unfavorably with components of the core splicing machinery such as U1 and U2 snRNPs (Kohtz et al., 1994; Wu and Maniatis, 1993). Several groups of ESEs are known, including purine-rich and AC-rich elements, as well as some with more complex composition (Graveley, 2000; Zheng, 2004). Most known ESEs function by recruiting members of the serine-arginine-rich (SR) protein family, which interact favorably with each other, with the pre-mRNA, and/or with snRNPs to enhance recognition of adjacent splice sites. By contrast, ESSs inhibit the use of adjacent splice sites, often acting through interactions with members of the heterogeneous nuclear ribonucleoprotein (hnRNP) family (Amendt et al., 1995; Caputi et al., 1999; Chen et al., 1999; Del Gatto-Konczak et al., 1999; Domsic et al., 2003; Kashima and Manley, 2003; Si et al., 1997; Staffa and Cochrane, 1995; Zahler et al., 2004; Zheng et al., 1998; Zhu et al., 2001). However, the molecular mechanisms by which ESSs inhibit splicing are just beginning to be understood (Wagner and Garcia-Blanco, 2001). Presence of a higher density of ESEs in authentic exons than in pseudoexons may contribute to recognition of the correct exons, while presence of ESSs in pseudoexons may suppress their splicing (Sironi et al., 2004; Zhang and Chasin, 2004). Thus, both of these classes of elements may contribute significantly to the specificity of pre-mRNA splicing (Smith and Valcarcel, 2000).

In addition to their known or presumed roles in constitutive splicing, both ESEs and ESSs contribute significantly to the regulation of alternative splicing (AS) (Black, 2003; Cartegni et al., 2002), a phenomenon observed in more than half of all human genes (Johnson et al., 2003). The process of AS is often precisely regulated in different developmental stages and different tissues (Black, 2003), and selection of the correct splicing variants in a given cell type is believed to be coordinated by multiple (sometimes overlapping) exonic and/or intronic splicing enhancers and silencers (Cartegni et al., 2002; Ladd and Cooper, 2002). ESSs can regulate AS either by antagonizing the function of ESEs or by recruiting factors that interfere with the splicing machinery directly (reviewed by Cartegni et al. [2002]; Ladd and Cooper [2002]; Zheng [2004]). For example, hnRNP I/PTB binds many intronic as well as exonic splicing silencers and appears to block access to the splicing machinery through protein multimerization (reviewed by Wagner and Garcia-Blanco [2001]).

Intriguingly, one third of randomly chosen human genomic DNAs of ~100 bases in length were found to have ESS activity, suggesting that ESSs are very prevalent

*Correspondence: cburge@mit.edu

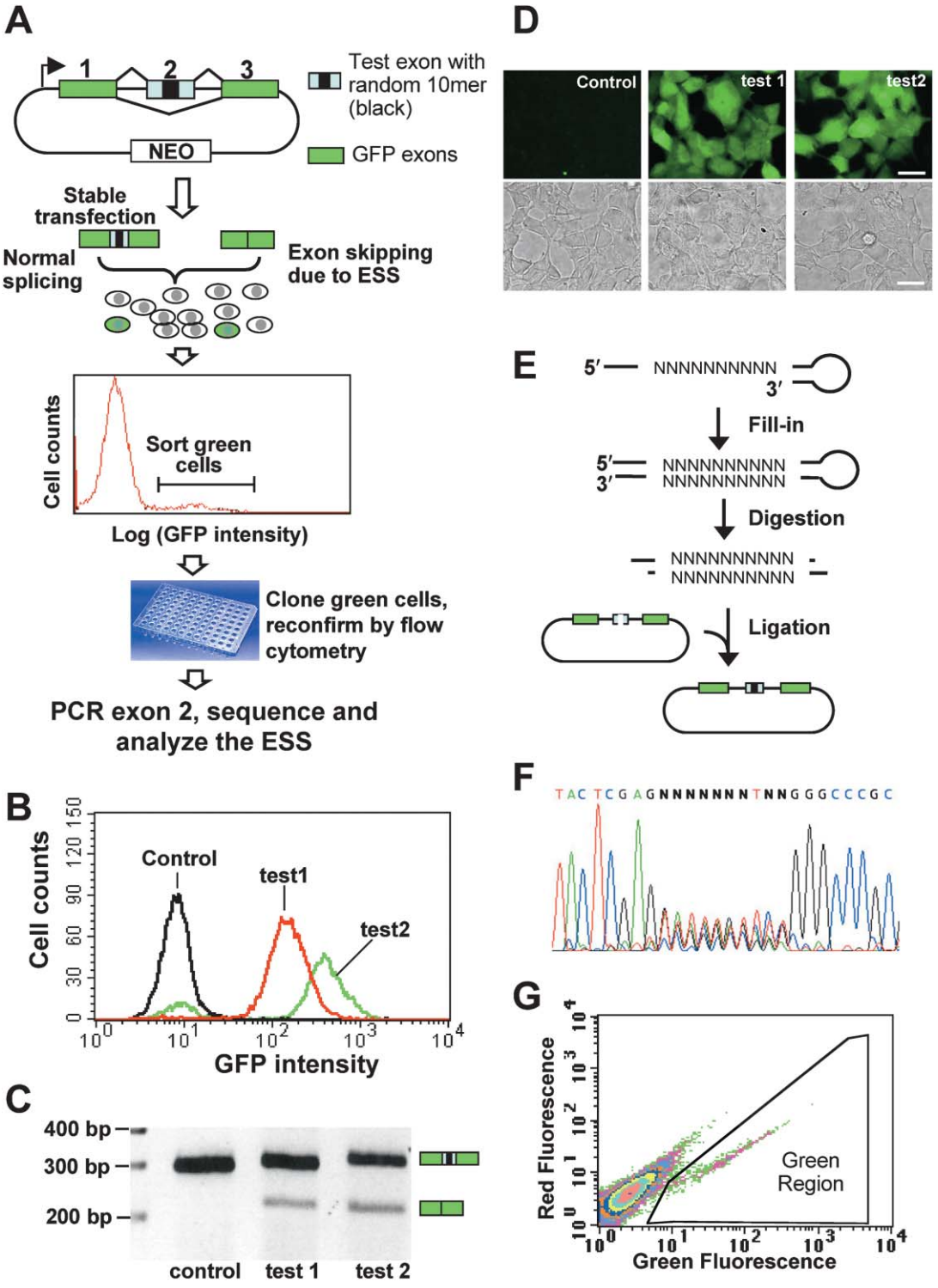


Figure 1. The Fluorescence-Activated Screen for Exonic Splicing Silencers

(A) Diagram of strategy used to screen for ESS.

(B) Test of the reporter system with two known ESS sequences. Test 1 (hnRNP A1 binding site, TATGATAGGGACTTAGGGT [Burd and Dreyfuss, 1994]) and test 2 (U2AF65 binding site, TTTTTTTTCCTTTTTTTTCCTTTT [Singh et al., 1995]) were inserted into the pZW4 reporter construct and transfected into 293 Flp-In cells, and positive transfectants were pooled for flow cytometry. The "Control" was a randomly chosen 10-mer sequence (ACCTCAGGCG) inserted into the same vector.

(C) RT-PCR results using RNA purified from the transfected cells as template, with primers targeted to exons 1 and 3 of pZW4.

(D) Microscopic images of transfected cells. Upper panel, GFP fluorescence. Lower panel, phase images. Scale bar, 50 μm.

(E) Construction of random decamer library. The foldback primer was synthesized with a random sequence of 10 bp, then extended with Klenow fragment, digested, ligated into pZW4, and transformed into *E. coli*.

(Fairbrother and Chasin, 2000). Relative to ESEs, only a limited number of ESSs have been characterized by mutational analysis; a comprehensive list has been compiled recently by Zheng (2004). Most of these examples share little similarity, suggesting that many more ESSs remain to be discovered. Thus, exonic silencers represent a relatively less studied aspect of splicing, and identification of additional ESSs is likely to contribute to understanding of both alternative and constitutive splicing. To systematically identify ESSs, we developed a cell-based splicing reporter system to screen a random sequence library for short sequences with splicing silencer activity in cultured human cells. The potential roles of derived ESS motifs in constitutive and alternative splicing were explored using a combination of experimental and bioinformatic approaches, including development of a splicing simulation algorithm and application to large sets of human transcripts.

Results and Discussion

Development of an Effective Reporter System for ESS Screening

We designed a three-exon minigene construct as a reporter for exon silencing (Figure 1A). Exons 1 and 3 of this construct form a complete mRNA encoding the enhanced GFP (eGFP) protein, whereas exon 2 is a test exon containing a cloning site into which oligonucleotides can be inserted. A small constitutively spliced exon—exon 2 of the Chinese hamster dihydrofolate reductase (*DHFR*) gene—was used as the test exon, together with its flanking introns. In this minigene, the test exon is normally included to form an mRNA that does not encode functional protein. However, insertion of an ESS sequence into the test exon can cause skipping of this exon, producing an mRNA encoding functional eGFP protein.

Our screen, which we call the fluorescence-activated screen for exonic splicing silencers (FAS-ESS, or FAS for short), is diagrammed in Figure 1A. Since the cores of ESSs are thought to be relatively short (~6–10 nt), we inserted a random pool of decanucleotides into the test exon. The resulting library was transfected into cultured human 293 cells, and stably transfected cells were combined and sorted for GFP-expressing cells by FACS. Total DNA was purified from positive clones, and the inserts were amplified by PCR and sequenced to identify the oligonucleotides responsible for exon skipping. To ensure that only a single minigene was stably inserted into each cell, we constructed our minigene reporter (which we call pZW4) using the pcDNA5/FRT vector that inserts into mammalian host cells by site-specific recombination (O'Gorman et al., 1991).

We first tested whether insertion of known ESS sequences can induce detectable exon skipping in this reporter system. Two known ESS sequences, together with an arbitrarily chosen 10-mer as a negative control, were inserted into the reporter construct pZW4. After transfection, hygromycin-resistant clones were selected and pooled for flow cytometry analysis. Most cells transfected with known ESS sequences were found to be GFP positive, as judged by flow cytometry and fluorescent microscopy, whereas cells transfected with the control construct had negligible green fluorescence (Figures 1B and 1D). RT-PCR analysis confirmed that the reporters containing the known ESSs produced mRNAs resulting from skipping of the test exon, whereas no skipping was detected in the control (Figure 1C).

To make the random 10-mer library, foldback DNA was synthesized with a 5′ overhang containing a random 10-mer region (Figure 1E). The unpaired random region was filled in by polymerase and then inserted into the pZW4 vector. Sufficient numbers of *E. coli* cells were transformed to obtain $\sim 2 \times 10^6$ colonies, providing ~2-fold coverage of the $4^{10} = \sim 10^6$ possible DNA decamers. To evaluate the quality of this library, 25 colonies were randomly picked for plasmid extraction and sequencing. We found that 24 out of 25 had 10 bp inserts, with little or no sequence bias at any position. The whole set of $\sim 2 \times 10^6$ colonies was pooled to extract a plasmid DNA library. To examine whether this randomness was maintained after transfection into 293 cells, total DNA from a pool of stably transfected cells was purified, and the insertion fragments were amplified and sequenced. The insertion fragments were essentially free of sequence biases at all positions (Figure 1F), suggesting that the starting pool represented an essentially random pool of decamers.

Each transfection was conducted with $\sim 2 \times 10^7$ 293-FlpIn cells grown in a 15 cm tissue culture dish. After hygromycin selection, ~ 2–6×10^3 positive clones were typically visible. All positive clones were pooled for FACS analysis, and typically about one in 5000 cells was found to be GFP positive (Figure 1G). Each GFP-positive cell was sorted into a single well of a 96-well plate. (Rarely, multiple cells would stick together and be sorted into same well; any cases in which multiple colonies were observed were discarded before sequencing.) Cells normally grew up in ~10% of the wells (~10 wells/plate). Each clone was replica plated. One duplicate from each clone was used for flow cytometry to reconfirm GFP expression, and the other was used to purify genomic DNA for PCR.

Our system has a number of desirable features. First, GFP was used as the reporter gene, so no growth advantage is expected between cells with the two splicing forms. Second, the three exons of the minigene share no homology to each other, minimizing the likelihood that DNA recombination would complicate the screen (Fairbrother and Chasin, 2000). Third, the test exon lacks known ESS or ESE sequences that might interfere with the screen. Fourth, we took advantage of the FLP recombinase system and a host cell line containing a single FRT integration site to generate a library of stably transfected cells, each inserted with a single minigene. Insertion into the same locus in every cell should also ensure

(F) Sequencing of the random decamer region. 293 cells stably transfected with the pZW4 library were pooled to purify total DNAs, from which minigene fragments were amplified by PCR and sequenced. Sequences around the insertion region are shown.
(G) Flow cytometry profile of single transfection using pZW4 random decamer library.

consistent expression of the reporter minigene. Fifth, PCR amplification was not used prior to the screening step, avoiding the sequence biases that can result from multiple rounds of PCR in SELEX. Finally, we recovered cells that skipped the test exon using FACS, providing very high sensitivity (one positive in >10,000 cells can be recovered).

Identification of ESS Decamers

We conducted 236 transfections in 17 batches for the screen, from which 141 ESS decamers were identified (see Supplemental Table S1 at http://www.cell.com/cgi/content/full/119/6/831/DC1/). Eight of these decamers were identified twice in independent transfections, and 21 pairs of decamers differed only by a single nucleotide. (In two cases, the ESSs identified were 9-mers rather than 10-mers, presumably due to imperfect synthesis of the randomized region in the foldback primer.) The ESS decamers were clustered based on sequence similarity and multiply aligned using CLUSTALW, as described previously (Fairbrother et al., 2002), to identify candidate silencer motifs (Figure 2A). At a cutoff dissimilarity score of 4.4, most ESS decamers fell into one of seven clusters of at least six sequences each, which were designated FAS-ESS groups A–G (Figure 2A). A total of 17 decamers were not included in these clusters. Weight matrices derived from these seven groups of aligned sequences were represented as pictograms, in which the heights of letters are proportional to the frequencies of the corresponding bases at each aligned position (Figure 2A). Comparing these motifs to a comprehensive collection of known ESSs (Zheng, 2004), we observed that two of the motifs, groups B and G, resemble known ESSs, which crosslink to hnRNP A1 and whose ESS activity can be disrupted by mutations that reduce A1 binding (Del Gatto-Konczak et al., 1999). A third, group C, resembles known ESSs that recruit hnRNP H to silence splicing (Chen et al., 1999). The group F and G motifs also bear a strong resemblance to positions +1 through +6 of the human 5′ss consensus sequence (/GT[AG]AGT, where the splice junction is indicated by / and [AG] indicates A or G; Lim and Burge, 2001). The remaining groups, A, D, and E, and a number of the decamers that did not fall into the seven large clusters appear to represent novel classes of ESS elements.

Although the initial decamer library was essentially random (~25% frequency of each base at each position), the ESS sequences identified from the screen had higher content of T (38%) and G (36%) and reduced levels of A and C (17% and 9%, respectively) (Figure 2B). We also calculated the frequencies of dinucleotides in the ESS decamer set and identified dinucleotides that are either overrepresented in these ESSs (e.g., CC, TA, GG, TC) or underrepresented (e.g., GA and AC). Many dinucleotides underrepresented in ESSs, including GA and AC, are overrepresented in the set of hexamers predicted to have ESE activity by the RESCUE-ESE approach (Fairbrother et al., 2002), and vice versa (data not shown). RESCUE-ESE hexamers were underrepresented in the recovered decamers by a factor of seven relative to random sets of 238 hexamers (10 versus 69).

To further validate the results of our screen, 21 of the recovered ESS decamers were inserted back into our reporter minigene vector to assess silencer activity in a transient transfection assay. The 21 candidates for reconfirmation (underlined in Figure 2A; sequences listed in Table 1) were selected so as to cover all FAS-ESS groups, but selection was otherwise arbitrary. After transient transfection with minigenes containing different decamers, skipping of the test exon caused 20%–45% of the cells to express GFP, whereas random decamer controls yielded <1% green cells (Figure 3A and see Supplemental Figure S1B on the *Cell* web site). The percentage of green cells was presumably limited by the efficiency of transient transfection, as stable lines containing the same ESS decamers yielded >90% green cells (Supplemental Figure S1A). The first seven of these decamers were tested also by stable transfection of the construct. As expected, the variation in GFP expression was lower in stably transfected cells, as judged by the fluorescence intensity (Supplemental Figure S1). However, because transient transfection yielded good signal with very low background, it was used for all the remaining experiments.

Skipping of the test exon in constructs containing the ESS decamers was also confirmed by RT-PCR (Figure 3B). For ESS3 and ESS5, the sizes of the major RT-PCR products (indicated by arrows) were intermediate between the normal splicing form and the exon skipping form. Both decamers contain sequences resembling the consensus 5′ss motif, and direct sequencing revealed that the intermediate-sized products resulted from use of these sequences as cryptic 5′ splice sites. However, in both cases, a significant level of exon skipping was also observed, both by FACS analysis and by RT-PCR (Figures 3A and 3B), confirming that these decamers contain sequences with ESS as well as 5′ss activity. A possible role for splice site-like sequences in exon silencing has been proposed previously (Fairbrother and Chasin 2000; Siebel et al., 1992).

Silencer Activity in a Second Cell Type and in a Heterologous Context

ESS sequences are thought to function through binding to specific *trans*-acting splicing factors, whose levels and activity may differ between cell types. Therefore, it was of interest to determine whether the ESS decamers identified could function in a second cell type. HeLa cells were chosen for these tests both because they are readily transfectable and because comparison to 293 cells might yield clues to potential differences in splicing between cancer cells and "normal" cells. Both cell types were transiently transfected with constructs containing 12 ESS decamers, and all led to significant exon skipping in both cell lines as judged by flow cytometry (data not shown) and by RT-PCR (Figure 3C). Considering all of the transient transfection results described above, all 21 tested ESS decamers caused significant and reproducible exon skipping (Figures 3B and 3C and Supplemental Figure S1B; RT-PCR results for two ESS decamers not shown), suggesting that the false positive rate of our screen was very low.

Since our screen was performed using a constant test exon (exon 2 of the Chinese hamster *DHFR* gene) and its flanking introns, it was a possibility that the ESS

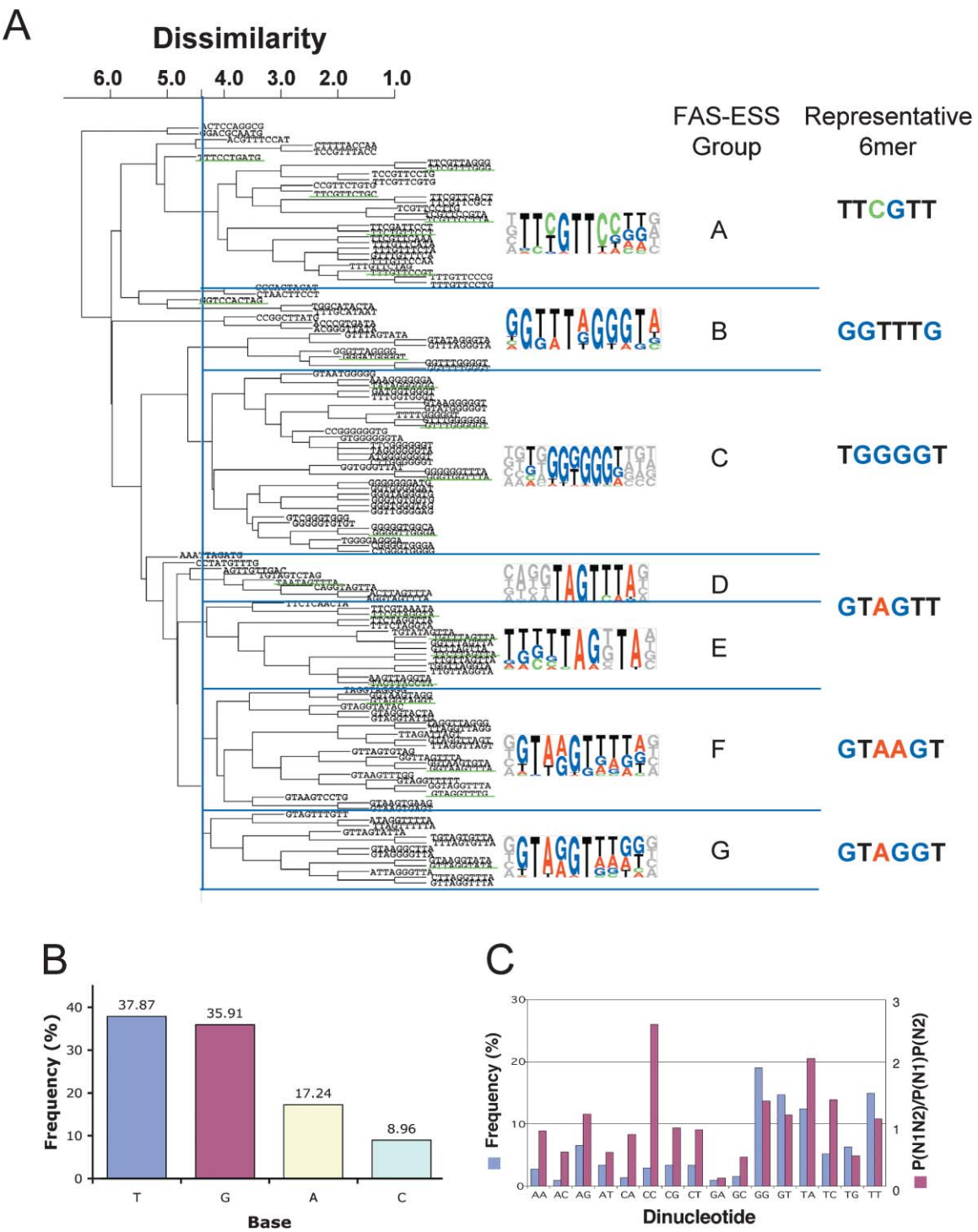


Figure 2. ESS Decamers, Clustering, and Sequence Composition

(A) Unique ESS decamers (133) were clustered and aligned as described (Fairbrother et al., 2002), using a dissimilarity cutoff of 4.4. Pictograms of the consensus motifs are shown beside each cluster. The clusters are designated FAS-ESS groups A–G. The decamers underlined are candidates that were reconfirmed by transient transfection (listed in Table 1). The hexamers listed at right are ESS candidates chosen for testing (see Figure 4 and text). The hexamer GTAGTT was chosen to represent both groups D and E.

(B) Base composition of ESS decamers.

(C) Frequencies and odds ratios of dinucleotides in ESS decamers.

Table 1. Summary of ESS Decamers that Were Reconfirmed

ESS Number	Sequence	FAS-ESS Group	Number of Times Recovered in Screen	Tested in Exon Context?	
				Original exon context	Heterologous exon context
ESS1	TTTGTTCCGT	A	1	Y	Y
ESS2	GGGTGGTTTA	C	1	Y	Y
ESS3	GTAGGTAGGT	F	1	Y	Y
ESS4	TTCGTTCTGC	A	2	Y	Y
ESS5	GGTAAGTAGG	F	1	Y	N
ESS6	GGTTAGTTTA	F	1	Y	N
ESS7	TTCGTAGGTA	E	2	Y	N
ESS8	GGTCCACTAG	–	1	Y	Y
ESS9	TTCTGTTCCT	A	1	Y	Y
ESS10	TCGTTCCTTA	A	1	Y	Y
ESS11	GGGATGGGGT	B	1	Y	Y
ESS12	GTTTGGGGGT	C	1	Y	Y
ESS13	TATAGGGGGG	C	1	Y	Y
ESS14	GGGGTTGGGA	C	1	Y	Y
ESS15	TTTCCTGATG	–	1	Y	N
ESS16	TGTTTAGTTA	E	1	Y	Y
ESS17	TTCTTAGTTA	E	1	Y	Y
ESS18	GTAGGTTTG	F	1	Y	N
ESS19	GTTAGGTATA	G	1	Y	Y
ESS20	TAATAGTTTA	D	1	Y	N
ESS21	TTCGTTTGGG	A	1	Y	N

The sequences of all the ESS decamers are listed in Supplemental Table S1.

sequences identified might require sequence context specific to this exon for function. For instance, the insertion of foreign sequences could induce some inhibitory RNA secondary structure or disrupt some positive-acting secondary structure involved in splicing (Buratti et al., 2004). To address this possibility, we first folded the test exon with and without the decamer insertions using RNAfold (Hofacker et al., 1994). However, we did not find significant differences in predicted RNA secondary structure. Selected decamers were also inserted into a similar minigene construct but with the *DHFR* test exon and flanking introns replaced by an unrelated test exon—the constitutively spliced exon 6 of the human *SIRT1* gene and its flanking introns. The insertion of all 14 ESS decamers tested (sequences listed in Table 1) gave rise to GFP-positive cells (data not shown), and exon skipping was further confirmed by RT-PCR (Figure 3D and data not shown). As controls, insertion of a random decamer or the vector alone with an eight-base multicloning site did not exhibit detectable exon skipping. These results demonstrate that the ESS decamers identified in our screen can generally function in a heterologous exon context.

Since most general splicing factors are ubiquitously expressed, use of 293 cells might be expected to allow identification of a large fraction of the ESSs that have activity in a broad range of cell types. Indeed, all tested ESSs identified in our screen functioned in another cell type (Figure 3C). However, it is likely that certain ESSs were not detected in our screen because the corresponding *trans*-factors are not expressed in 293 cells (or are expressed at very low levels). Since our screen used 10-mers in the context of a constitutive exon, it might be expected to miss those ESSs that are long ($>$10 bases), those whose silencer activity is weak or requires multiple copies, and those that generated overlapping ESEs when inserted into our test construct.

Estimation of the Total Number of ESS Decamers

Oligonucleotides of size 10 were chosen for our ESS screen because this size was large enough to contain the core sequences of most known ESSs and hnRNP binding motifs while remaining small enough that we could create a library providing in excess of 1-fold coverage of the $\sim 1 \times 10^6$ distinct decamers. The observation that the 141 decamers identified in our screen contained eight distinct sequences that were recovered twice in independent transfections allowed us to make maximum likelihood estimates of the size of the pool of ESS decamers from which this sample was drawn as between $\sim$1100 and 1500 (Supplemental Data). Thus, our screen identified $\sim$10% of all ESS decamers that could theoretically be identified in a completely exhaustive screen using this system. Of course, this level of recovery of individual ESS decamers implies a much higher coverage of shorter ESS motifs, which will be represented in many distinct decamers. For ESSs of length 6, our screen probably approached saturation. For example, in the last batch of 31 ESS decamers sequenced, all contained at least one hexamer that had appeared in one or more of the previously recovered decamers, motivating additional analyses of the hexanucleotide content of the recovered decamers.

Identification of Overrepresented Motifs in ESS Decamers

Statistical overrepresentation in the set of recovered decamers was used as a criterion to identify specific hexanucleotides likely to possess intrinsic ESS activity. For this analysis, each ESS decamer was extended into a 14-mer by appending 2 nt of the vector sequence at each end to allow for cases in which silencer activity derived from sequences overlapping the vector. These extended sequences contained a total of 1195 overlapping hexamers (nine hexamers from each of the 131

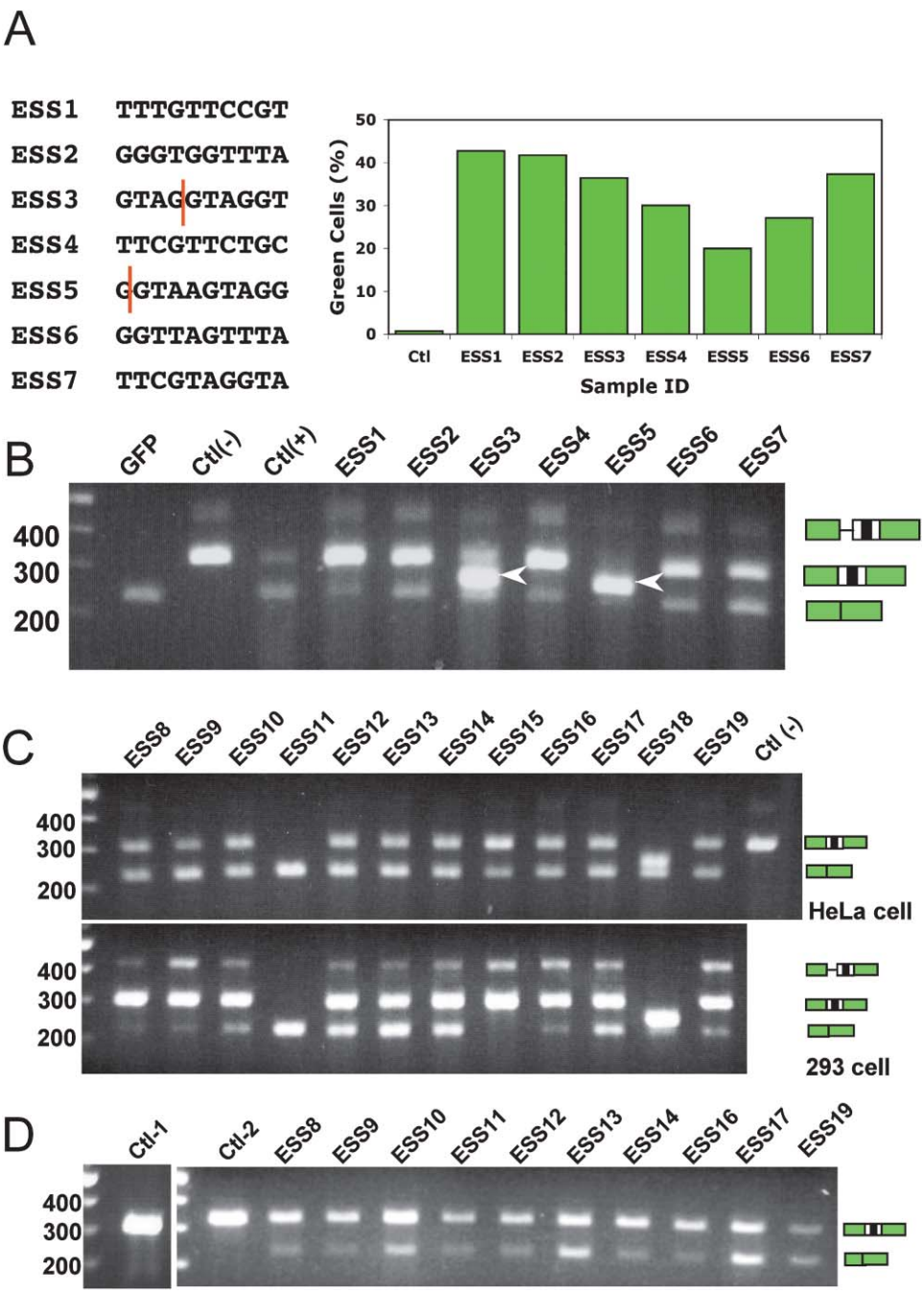


Figure 3. Test of ESS Activity by Transient Transfection, in a Second Cell Type, and in a Heterologous Exon Context

(A) The sequences of selected ESS decamers (vertical lines indicate cryptic 5′ss) and flow cytometry profiles of cells transfected with ESS. All ESS decamers were inserted into the pZW2 vector, transfected into 293 cells, and assayed by flow cytometry 1 day after transfection. The percentage of GFP-positive cells was used to assess exon skipping.

(B) RT-PCR results using RNA purified from cells transfected with the ESS constructs listed in (A). The white arrowheads indicate use of the cryptic 5′ss shown in (A). Exon skipping caused by ESS5 was additionally confirmed by flow cytometry (Supplemental Figure S1) and by RT-PCR with a primer complementary to the junction between exons 1 and 3 (data not shown).

(C) Comparison of ESS activity in two cell types. pZW2 constructs with different ESS decamer insertions were transfected into HeLa (upper panel) and 293 cells (lower panel). The decamers used are listed in Table 1. As a control, a random decamer (GATCATTCAT) gave total exon inclusion in HeLa cells. For ESS15 in 293 cells, exon skipping was also confirmed by flow cytometry (Supplemental Figure S1) and by RT-PCR with a primer spaning the junction between exons 1 and 3 (data not shown).

(D) Test of ESS activity in a heterologous exon context (human *SIRT1* exon 6 and flanking introns; construct pZW8). Selected ESS decamers were inserted into the *SIRT1* exon of pZW8, transfected into HeLa cells, and analyzed by RT-PCR. "Ctl-1" is the pZW8 construct with a random decamer insertion (ACCGAAGAGC). "Ctl-2" is this construct with a cloning site (octamer) in place of the decamer. Constructs pZW8-ESS15 and pZW8-ESS18 were attempted but later found to have no insert (data not shown).

unique extended decamers, plus eight hexamers from each of the two extended nonamers). The number of hexamers occurring at least three times in the extended ESS decamers was 103, more than 3-fold higher than expected for sets of randomly chosen decamers ($p < 10^{-4}$, based on 10,000 samplings of 133 random decamers inserted into the same flanking sequence context; data not shown). Thus, there are substantial biases in the hexamer composition of the ESS decamers, and this set of 103 hexamers (which we refer to as the FAS-hex3 set) is likely to be highly enriched for hexamers that represent core portions of ESS motifs. We also considered the set of 176 hexamers that occurred at least twice, a set we refer to as FAS-hex2. Although this set is not much larger than was observed in our random samplings, intuitively, hexamers that occur twice in a set of ESS decamers are more likely to have ESS activity than randomly chosen hexamers. Therefore, the FAS-hex-2 set may have somewhat higher sensitivity for ESS detection. For reference, both sets of hexamers and their counts are listed in Supplemental Table S2.

These observations raised the question of whether significantly overrepresented hexanucleotides such as those of the FAS-hex3 set have intrinsic ESS activity. To explore this question, we chose one hexamer from the FAS-hex3 set that resembled the consensus of each group of decamers (Figure 2A) to test for silencer activity in our splicing reporter construct. (In one case, the same hexamer, GTAGTT, was used to represent two groups—D and E.) For group C, in order to avoid the technical difficulties associated with synthesis of poly-G-containing oligonucleotides, a sequence containing fewer Gs (TGGGGT, from the FAS-hex2 set) was used.

Two strategies were used to test the silencer activity of overrepresented hexamers. In an initial experiment, the six hexamers representing the ESS motifs of Figure 2A were inserted by themselves (as well as a single or double point mutant control for each) into the reporter minigene, and these constructs were transfected into 293 cells to assay for exon skipping. In this experiment, four out of the six hexamers tested had silencer activity, producing a significant fraction of GFP-positive cells, while the mutant control sequences all had low background levels of green cells (Figure 4A). In a second experiment, these hexamers were extended to create decamers that contained two or more overlapping hexamers from the FAS-hex3 set but which were themselves not recovered as ESS decamers in our screen. Again, control decamer sequences were designed containing point mutations at one or two positions in each contained hexamer, and each decamer was inserted into the reporter minigene and assayed for silencer activity as above. In this second experiment, all six putative ESS decamers, including the two that were based on hexamers that had failed the initial test, produced a significant fraction of GFP-positive cells (Figure 4B). Again, low background levels of green cells were obtained for all of the mutant controls. The observation that a majority of tested hexamers had ESS activity alone and that new ESS decamers could be reliably designed using these hexamers supports the idea that most of the FAS-hex3 hexamers represent core ESS motifs.

No Evidence for Reading Frame Effects on ESS Activity

In certain cases, presence of a premature termination codon (PTC) can lead to skipping of the exon containing the PTC, a process called nonsense-associated altered splicing (NAS). A hypothetical process, nuclear scanning, has been proposed to recognize PTCs in the nucleus and alter splicing to skip the nonsense-containing exon (Wilkinson and Shyu, 2002). Alternative rationales for observed NAS events include explanations involving RT-PCR artifacts and nonsense-mediated mRNA decay (NMD) (Valentine and Heflich, 1997), or explanations involving alterations to splicing regulatory sequences, e.g., disruption of ESEs (Liu et al., 2001), or both (Caputi et al., 2002).

In the set of 133 unique ESS decamers recovered, 59 ($\sim$44%) contained one or more PTCs, higher than expected by chance ($\sim$16%). To address the possibility that these sequences function as silencers through the process of NAS, we constructed three different vectors for each of three PTC-containing ESS decamers (ESS2, ESS6, and ESS7; Supplemental Figure S2) by inserting one to three bases before the decamer insertion site. All nine constructs were observed to cause exon skipping in transient transfection assays, regardless of whether they generated PTCs (Supplemental Figure S2), consistent with direct ESS activity for these decamers and inconsistent with models involving NAS. Of the 62 PTCs in our ESS decamers, 55 were TAG, compared to only four occurrences of TGA and three of TAA. These data can be simply explained by the presence of TAG triplets in several of the candidate ESS motifs of Figure 2A (e.g., groups B, D, and E). Consistent with this idea, the counts of out-of-frame triplets were also far higher for TAG than for TGA or TAA in both alternate reading frames (data not shown). Thus, our experimental and statistical analyses found no evidence for effects of reading frame on splicing in our reporter system. These observations suggest that, in addition to frequent disruption of ESEs (Liu et al., 2001), some nonsense mutations (especially amber mutations) could create ESSs, potentially providing an alternative explanation for some apparent cases of NAS.

Enrichment of ESS Hexamers in Pseudoexons, Strong Exons, and Alternative Exons

In an initial exploration of the function of FAS-ESSs in endogenous human gene loci, the relative frequencies of ESS sequences were analyzed in different functional categories of exons. This analysis used large data sets of genomically aligned human transcripts to generate five categories of sequences (see Experimental Procedures): constitutively spliced exons (CEs)—internal exons for which alternative splicing was not observed in available cDNA and EST sequences; pseudoexons (PEs)—intronic pairs of high-scoring potential 3′ss and 5′ss spaced 50–250 bp apart; "strong" exons—CEs with both 3′ss and 5′ss in the top quartile of scores; "weak" exons—CEs with both splice sites in the bottom quartile of scores; and skipped exons (SEs) —exons for which transcripts excluding and including the exon were identified. The decamer and hexamer content of these sequences was then tallied, and χ^2 statistics were used to

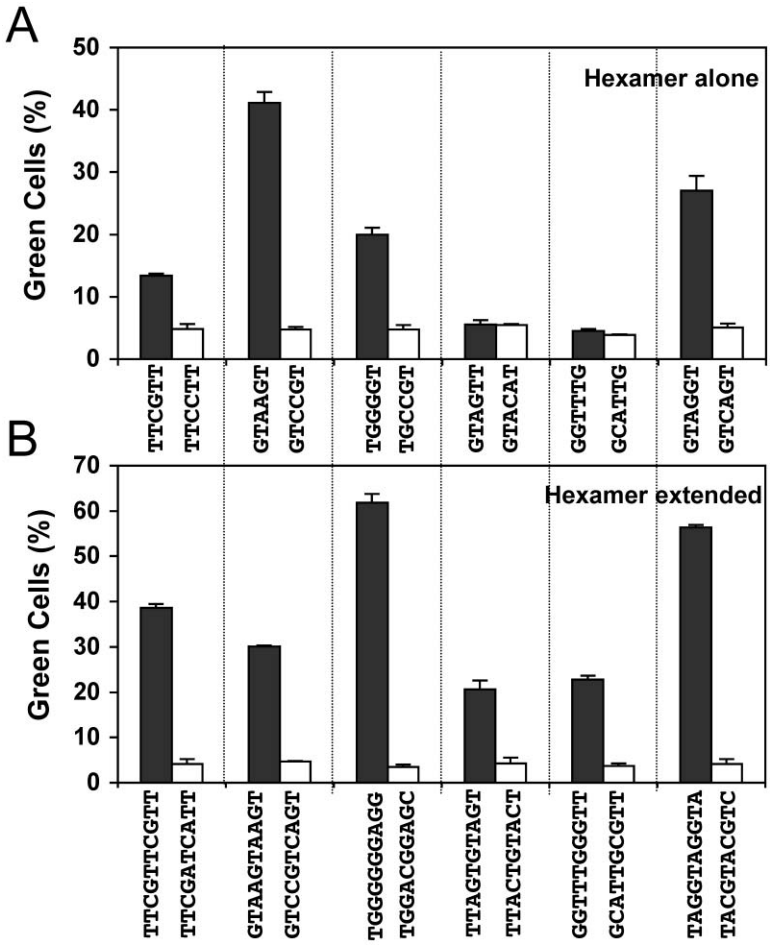


Figure 4. Silencer Activity of Putative ESS Hexamers and Extended Hexamers

(A) The hexamers listed in Figure 2 and corresponding mutants were inserted into pZW2 and transfected into 293 cells. Exon skipping was assessed by GFP fluorescence.

(B) These or closely related hexamers were extended into 10-mers by appending overlapping hexamers from the FAS-hex3 set. ESS activity of these extended sequences and mutants was tested as in (A). All transfections were conducted at least twice, and the mean and standard deviation are plotted. Black bars, putative ESS sequences; white bars, mutant sequences.

assess the significance of the observed biases. These data sets were sufficiently large to enable analysis of individual hexamers but not of individual decamers. Therefore, the 133 unique FAS-ESS decamers were analyzed as a group.

If the ESS decamers recovered in our screen function commonly as splicing silencers in endogenous human gene loci as they do in our reporter minigene, then selection should tend to eliminate them from CEs, where ESSs would be expected to interfere with efficient splicing but might favor them in PEs, where splicing is undesirable. Consistent with this expectation, the set of ESS decamers was substantially depleted in CEs relative to PEs (χ^2 test, $p << 2.2 \times 10^{-16}$). Furthermore, since exons with weaker splice sites are likely to be more prone to being silenced by ESSs, one might expect increased selection against ESSs in CEs with weak splice sites. Consistent with this idea, the set of ESS decamers was also found to be depleted in weak exons relative to strong exons, as defined above (χ^2 test, $p << 2.2 \times 10^{-16}$). Thus, both the direction and strength of these observed biases support the idea that the decamers obtained in our screen function commonly as ESSs in endogenous human gene loci. These ESS decamers also had marginally higher frequency in SEs than in CEs (χ^2 test, $p = \sim 0.05$), suggesting that some might play a role in control of AS.

Analysis of distributional biases for the FAS-hex3 hexamers supported the trends observed for ESS decamers. Differences in frequency between the three pairs of sequence categories analyzed above (PEs versus CEs, strong exons versus weak exons, and SEs versus CEs) were plotted using "χ" statistics (square root of the χ^2 statistic with sign indicating direction of bias) for individual hexamers in a three-dimensional scatter plot format (Figure 5A). Two-dimensional projections of this plot across the three pairs of axes are also displayed (Figures 5B, 5C, and 5D). These data show that most FAS-hex3 hexamers are significantly enriched in PEs relative to CEs and that many are also enriched in strong exons relative to weak exons and in SEs relative to CEs. Only a handful of these hexamers showed significant enrichment in CEs versus PEs or in weak exons versus strong exons, and none showed significant enrichment in CEs relative to SEs. Thus, taken as a whole, these data support the idea that many but not necessarily all of the FAS-hex3 hexamers commonly play a role as ESSs (or portions thereof) in endogenous human exons and that some may play a role in control of AS. For reference, the χ^2 values for individual ESS hexamers for these three comparisons are listed in Supplemental Table S3. To further explore the possible roles of FAS-hex3 hexamers in AS, we compared the frequency of the FAS-hex3 set of hexamers in the "extended" portions of a database of alternative 5′ splice site exons (i.e., the portion between the two alternative 5′ splice sites) relative to similarly sized exonic sequences adjacent to constitutive 5′

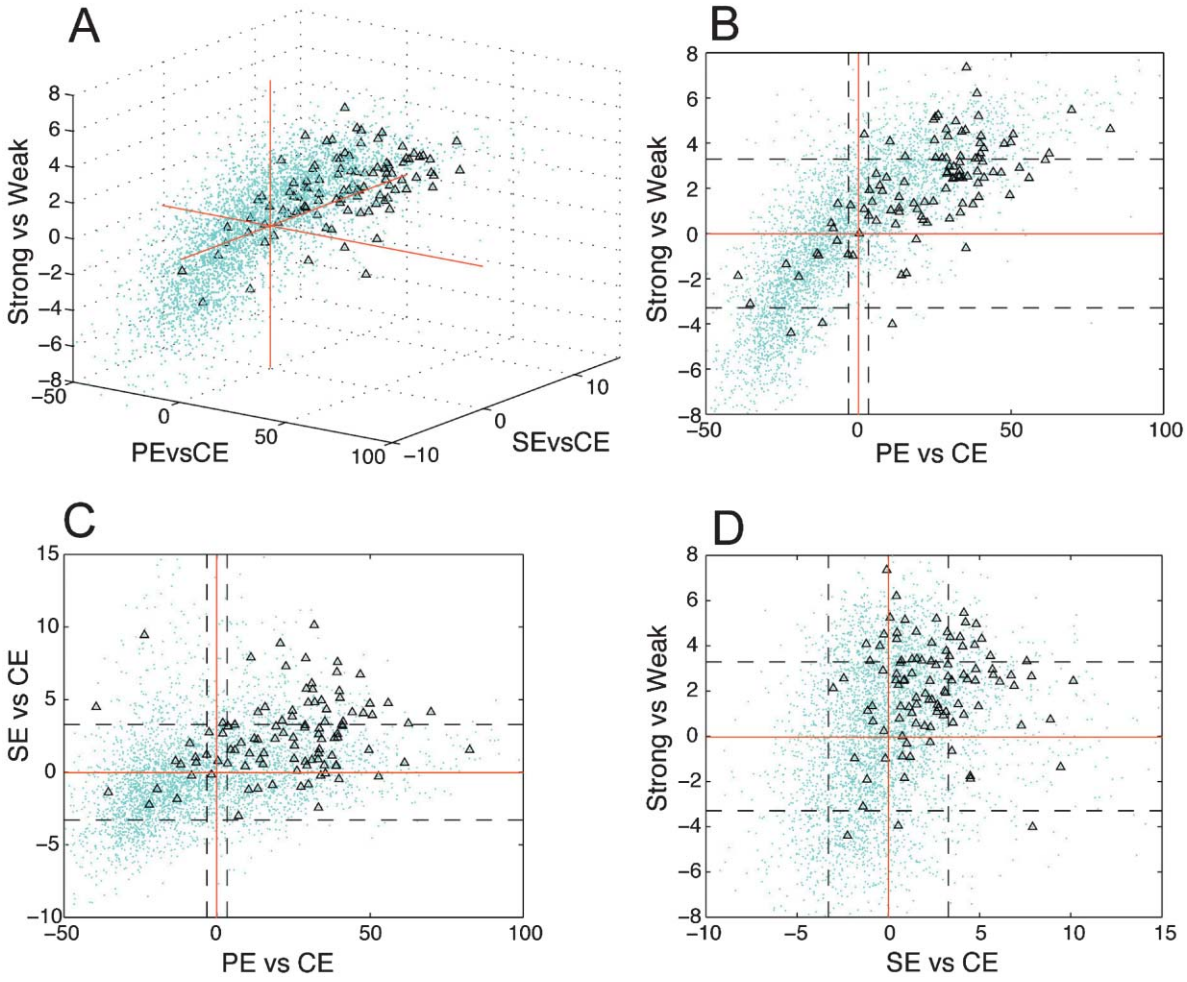


Figure 5. Association of ESS Hexamers with Pseudoexons, Skipped Exons, and Exons with Strong Splice Sites

(A) Three-dimensional scatter plot of the χ statistic (see Experimental Procedures) for all hexamers (blue dots) and for FAS-hex3 hexamers (black triangles) across three axes (PE versus CE, SE versus CE, and strong exon versus weak exons; see text for definitions).

(B–D) Two-dimensional projections of the data in (A). Dashed lines indicate the χ statistic cutoff of 3.29 (corresponding to $\chi^2 = 10.8$, $p < 0.001$).

splice sites (Experimental Procedures) and found that they are enriched in the alternative 5′ exons ($\chi^2 = 46.6$, $p < 10^{-11}$). A similar result was obtained for alternative 3′ splice sites ($\chi^2 = 23.8$, $p < 10^{-6}$). These data suggest that the FAS-hex3 hexamers may also be involved in regulation of exons with alternative 5′ or 3′ splice sites.

ESS Hexamers Are Predictive of Exon Skipping in the *HPRT* Locus

To evaluate the potential utility of ESS hexamers in predicting the splicing phenotypes of mutations in endogenous genes, we adapted the approach used previously (Fairbrother et al., 2002). From published mutation data (Tu et al., 2000; Valentine, 1998), a set of exon mutations that cause exon skipping in the *HPRT* gene was collected. The *HPRT* exons were considered as sets of overlapping hexamers, so a point mutation is considered to disrupt six overlapping hexamers in the wild-type sequence, creating six new hexamers that are point mutants of the wild-type hexamers. Of the 58 *HPRT* mutations that alter hexamer composition and cause exon skipping, 14 of them created FAS-hex2 hexamers that were not present in the wild-type sequence, as compared to only three mutations that disrupted FAS-hex2 hexamers (Supplemental Table S4). This ratio (14/3 = ~4.7) of ESS creation to ESS disruption was statistically significant compared to random sets of 176 hexamers ($p < 0.01$). Similar results were obtained using the FAS-hex3 hexamers (ratio: 9/3 = 3.0, $p < 0.03$). The ratios obtained were also similar to the ratio of the number of mutations that disrupted RESCUE-ESE hexamers to the number that created RESCUE-ESE hexamers (~4.7) observed previously in a smaller set of *HPRT* exon-skipping mutants (Fairbrother et al., 2002). These observations suggest that the putative ESS hexamers identified here should prove useful as a predictive tool for analyzing the splicing phenotypes of exonic mutations or polymorphisms. Interestingly, five of these mutations both create putative ESS hexamers and disrupt RESCUE-ESE hxamers, suggesting that some single base changes may be capable of switching splicing regulation from positive to negative or vice versa (Supplemental Table S4). Consistent with this possibility, a

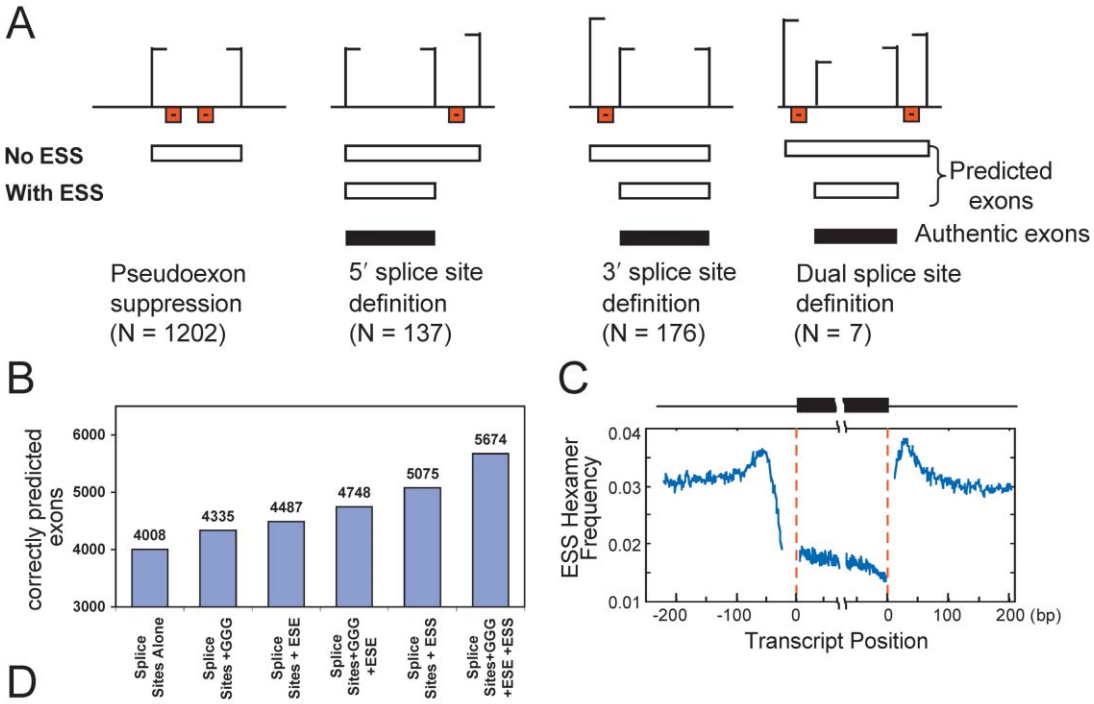
point mutation in the human *SMN2* locus appears to have the potential to both disrupt an ESE (Cartegni and Krainer, 2002) and create an ESS (Kashima and Manley, 2003).

Splicing Simulations Suggest ESS Roles in Pseudoexon Suppression and Splice Site Definition

To explore the potential contributions of ESSs and other splicing regulatory elements to the specificity of constitutive splicing, we developed a first-generation splicing simulation algorithm called ExonScan. This algorithm integrates scoring of potential splice sites with scoring of known or putative splicing regulatory sequences to predict the locations of internal exons in primary transcript sequences. Each of the candidate exons (nearby potential 3′ss and 5′ss pairs) in a transcript is scored for splice site strength and for the occurrence of splicing regulatory motifs in the candidate exon (for ESE and ESS motifs) or in nearby intronic regions (for intronic splicing enhancers [ISEs]). Nonoverlapping candidate exons that score above a score cutoff are predicted as exons (see Figure 6A for examples and Experimental Procedures for details). The score cutoff is chosen as the score for which the number of predicted exons most closely matches the number of authentic exons present in a training set of transcripts. At this cutoff, the numbers of false positive predictions and false negative predictions are essentially equal. ExonScan predictions using different combinations of regulatory sequences can then be compared to the authentic exons in a set of transcripts to assess the potential contributions of different types of splicing regulatory elements to constitutive splicing. Like its predecessor IntronScan (Lim and Burge, 2001), which only simulated splicing of short introns, ExonScan is a splicing simulation algorithm rather than an exon or gene finder, because the model uses only information known or predicted to be used by the nuclear splicing machinery and ignores features such as sequence conservation in other species that are inaccessible to the splicing machinery. We refer to ExonScan as a first-generation model because the algorithm simply adds the contributions of discrete sequence elements to derive a composite score for each potential exon, ignoring all the complexities of *trans*-acting factors and their interactions, potential higher-order effects such as cooperativity or interference between *cis*-elements, and potential effects of RNA structure.

Versions of ExonScan using different combinations of splicing-regulatory elements were used to simulate splicing for a data set of 1820 human primary transcripts containing a total of 10,891 internal exons. This set of transcripts was derived by spliced alignment of full-length cDNAs to genomic sequences and excludes cases where transcript evidence of potential alternative splicing was observed (see Experimental Procedures). The results of these large-scale splicing simulations were assessed by comparing the set of predicted internal exons in each transcript to the set of authentic internal exons (Figure 6). Counting the number of exons predicted exactly (both splice sites correct), the version of ExonScan using models of the 3′ss and 5′ss motifs alone, which we call ExonScan-SS, identified 4008 exons correctly (36.8%). This result is consistent with the established notion that the splice site motifs contain enough information to identify some but not most exons in human primary transcripts (Lim and Burge, 2001). Regulatory element scores were then determined for each of three different types of splicing regulatory elements: G triplets, a known mammalian ISE (McCullough and Berget, 1997); RESCUE-predicted ESE hexamers (Fairbrother et al., 2002); and the FAS-hex3 putative ESS hexamers described above. Scores for ESE and ESS hexamers were determined using simple log-odds scoring, with scores proportional to the logarithm of the ratio of the frequency (number of occurrences/base) of the given hexamer in human exons to the frequency in introns, and reciprocally for the ISE sequence GGG. Use of log-odds scoring represents a reasonable default choice for this initial application of ExonScan but may not be optimal. With this scoring system, all RESCUE-ESE hexamers and GGG received positive scores, while almost all FAS-hex3 hexamers received negative scores. Regulatory elements were incorporated into ExonScan singly and in combination, and simulation accuracy was assessed as the number of exactly predicted internal exons (i.e., both splice sites correct) for the set of 1820 transcripts described above. The results showed that the ExonScan-GGG and ExonScan-ESE algorithms (incorporating GGG as an ISE and the RESCUE-ESE hexamers as ESEs, respectively) both gave improved accuracy, but the greatest improvement was observed for ExonScan-ESS, scoring the FAS-hex3 hexamers as ESSs (Figure 6B). These results suggest that ESSs may play at least as important a role in constitutive splicing as known positive-acting elements.

Recently, two groups have used computational methods to predict ESSs (Sironi et al., 2004; Zhang and Chasin, 2004). Both methods used the assumption that ESSs are enriched in pseudoexons relative to authentic exons. Sironi et al. (2004) predicted three ESS motifs. One of these, which resembled the binding motif for hnRNP H, was confirmed experimentally. Zhang and Chasin (2004) predicted 974 putative ESS (PESS) 8-mers using two criteria: enrichment in pseudoexons relative to noncoding exons and enrichment in intronless 5′ UTRs relative to noncoding exons. They also identified 2069 putative ESE (PESE) 8-mers as sequences exhibiting the opposite enrichment patterns (enriched in noncoding exons relative to pseudoexons and relative to intronless 5′ UTRs). Of the identified PESS sequences, 11 were shown to increase exon skipping to between 50% and 80% from a background level of 10% in a splicing reporter construct. The 974 PESS 8-mers were clustered into 69 families. Although some individual PESS 8-mers resembled known ESS motifs, the consensus sequences of these families generally did not resemble known ESSs. These PESS 8-mers partially overlapped the ESS hexamers identified in our screen: 53% of the FAS-hex3 hexamers were contained in 8-mers of the PESS set, compared to an average of 22% for random sets of 103 hexamers. Scoring of the PESS 8-mers as ESSs in ExonScan improved the accuracy of splicing simulation relative to ExonScan-SS, supporting a role for these sequences in constitutive splicing, but yielded ~3%–4% fewer correct exons than when using the FAS-hex3 hexamers (data not shown). Deriving overrepre-



Features used in ExonScan	Improvement in accuracy relative to SS only (+=2%) <10kbp Exact	<10kbp Partial	10-30kbp Exact	10-30kbp Partial	>30kbp Exact	>30kbp Partial
SS only	0	0	0	0	0	0
SS+GGG	++	++	++	++	+	+
SS+ESE	++	++	++	+++	+++	+++++
SS+ESS	++++	++++	+++++	++++	+++++	++++
SS+ESE+ESS	+++++	+++++	+++++++	+++++++	++++++++	++++++++
SS+ESE+GGG+ESS	++++++	++++++	+++++++	+++++++	++++++++	++++++++

Figure 6. Results of ExonScan Analysis

(A) Diagram of typical cases in which scoring of FAS-hex3 hexamers by ExonScan-ESS improved prediction relative to scoring by ExonScan-SS. Vertical lines represent (positive) scores of potential splice sites (horizontal segment at top of line faces right for 3′ss, left for 5′ss); red boxes represent (negative) scores for ESS hexamers. Predicted exons without and with scoring of ESSs are represented by white boxes and authentic exons by filled boxes. Four representative cases are labeled, and the number of each case in the test set of 10,891 internal exons is indicated below in parentheses.

(B) The number of correctly predicted exons in the test set of transcripts for ExonScan versions using different combinations of regulatory elements. See text for details of scoring and score cutoffs used.

(C) The positional frequency of FAS-hex3 hexamers in the vicinity of 3′ss and 5′ss is plotted as the number of FAS-hex3 hexamers divided by total number of hexamers present at each position in and flanking CEs. The first and last 70 bases of exons and the first and last ~200 bases of introns are shown, excluding regions of 28 bases and 13 bases at the 3′ss and 5′ss, respectively.

(D) The relative improvement in simulation accuracy using additional information compared to using splice sites (SS) only is plotted for three sets of human transcripts, grouped by transcript length. "+" indicates an increase of 0%–2% in fraction of exons correct, "++" indicates a 2%–4% increase, etc. Exact and partial accuracy are defined in text.

sented hexamers from the PESS 8-mers and scoring these hexamers with ExonScan gave improved results relative to scoring the 8-mers directly. A version of ExonScan-ESS using these PESS-derived hexamers yielded 4900 correct exons in our test set of 1820 transcripts, close to the 5075 correct exons obtained for ExonScan using the FAS-hex3 hexamers (Figure 6B). Deriving over-represented hexamers (total of 248 hexamers) from the PESE 8-mers and scoring these with ExonScan gave roughly similar results (4378 correct exons) to those obtained using the set of 238 RESCUE-ESE hexamers (4487 correct exons), again demonstrating the ability of ExonScan to incorporate data from diverse sources. Appropriate ways of combining the sequences recovered by these different methods may be worth exploring.

To explore the ways in which the FAS-hex3 ESS hexamers contribute to splicing simulation accuracy, we classified predictions into four categories (Figure 6A): (1) "pseudoexon suppression," in which a potential exon predicted by ExonScan-SS is no longer predicted by ExonScan-ESS because of presence of internal ESS sequence(s); (2) "5′ splice site definition," in which a 5′ss

predicted incorrectly by ExonScan-SS is predicted correctly by ExonScan-ESS; (3) "3′ splice site definition," in which a 3′ss predicted incorrectly by ExonScan-SS is predicted correctly by ExonScan-ESS; and (4) "dual splice site definition," in which both splice sites predicted incorrectly by ExonScan-SS are predicted correctly by ExonScan-ESS. Comparing the frequencies of these events (listed in Figure 6A) suggests that the ESS hexamers studied play a predominant role in suppression of pseudoexons (see also Sironi et al. [2004]; Zhang and Chasin [2004]) but also play significant roles in definition of both 3′ and 5′ splice sites by suppressing upstream and downstream decoy splice sites. Analysis of the positional frequencies of the set of FAS-hex3 hexamers in the vicinity of exons (Figure 6C) showed substantially higher frequency in introns relative to exons, consistent with their inferred role in pseudoexon suppression. This distribution also revealed peaks of ESS hexamer frequency ~60 bp upstream of the 3′ss and ~40 bp downstream of 5′ss, consistent with their inferred roles in definition of both 5′ and 3′ splice sites.

Examination of ExonScan results for specific transcripts suggested that simulation accuracy might vary as a function of transcript length. To explore this phenomenon in more detail, we analyzed the results of different versions of ExonScan on three subsets of human transcripts grouped by transcript size (Figure 6D). For this analysis, simulation accuracy was measured in terms of the fraction of exons predicted exactly (both splice sites correct, as above) and also in terms of the fraction of exons predicted partially (at least one splice site correct). The results revealed several patterns. First, G triples contributed more to the accuracy of splicing simulation for short (<10 kbp) and medium-sized (10–30 kbp) transcripts than for long transcripts (>30 kbp), consistent with previous observations that GGG triplets are enriched in transcripts with short introns (McCullough and Berget, 1997; Yeo et al., 2004b). The opposite pattern was observed for ESE hexamers, which improved simulation accuracy more in long transcripts than in medium or short transcripts. These observations are consistent with the established principle that the intron definition mode of splicing, in which the unit of recognition is the intron and ISEs are often involved, is more prevalent in shorter transcripts and that the exon definition mode, in which ESEs often help to define exons, is more prevalent in long transcripts (Sterner et al., 1996). The results for ExonScan-ESS also showed greater improvement for longer transcripts than for shorter transcripts, consistent with the idea that ESSs play a predominant role in suppression of pseudoexons (Figure 6A), which are present at an increased ratio relative to authentic exons in longer transcripts because of the increased intronic content of these transcripts. The overall accuracy of ExonScan incorporating all three types of regulatory motifs (Supplemental Table S5) was higher in short- and medium-sized transcripts (76.5% and 69.2% partial accuracy, respectively) than in long transcripts (56.5%), likely reflecting the increased ratio of pseudoexons to authentic exons in longer transcripts.

Finally, we observed consistent differences in the improvement in exact versus partial accuracy for the three types of regulatory elements studied. While RESCUE-ESE hexamers generally improved partial accuracy more than exact accuracy, the ESS hexamers contributed more dramatically to exact accuracy. Although definitive interpretation of these data will require further studies, these observations are consistent with a model in which ESEs assist in recognition of the approximate locations of exons, while ESSs play a more prominent role in splice site definition by silencing distal decoy splice sites in the vicinity of exons (as diagrammed in Figure 6A). To further explore the latter idea, we generated two data sets of 3′ss-proximal intronic sequences: one set that contained upstream decoy 3′ss of strength comparable to the authentic 3′ss and another set that lacked such decoys. Comparing the density of ESSs in the two sets, we observed a significantly higher frequency of FAS-hex3 hexamers in the set of sequences with upstream decoy 3′ss ($\chi^2 = 146$, $p < 2.2 \times 10^{-16}$; Supplemental Table S6). These ESS hexamers were also enriched in analogous sets of 5′ss-proximal intronic sequences with downstream decoy 5′ss relative to those without such decoys ($\chi^2 = 229$, $p < 2.2 \times 10^{-16}$; Supplemental Table S6). These data provide additional support for a proposed role for ESSs in suppression of upstream decoy 3′ss and downstream decoy 5′ss in human transcripts.

Toward an RNA Splicing Code

A comprehensive description of the sequence specificity of pre-mRNA splicing—an "RNA splicing code"—will require precise knowledge of all of the types of splicing regulatory elements and their functions and interactions. Here, we have developed a cell-based screening protocol and applied it to systematically identify ESSs. With appropriate modifications, this protocol could be used to screen for other types of splicing regulatory elements. Improved knowledge of such elements should facilitate the development of increasingly effective splicing simulation algorithms, offering a possible route toward a more integrated understanding of splicing decisions.

Experimental Procedures

Constructs

The GFP reporter construct was amplified by PCR using pEGFP-C1 (Clontech) as a template with primers that contain splice sites. Candidate ESS sequences and controls were inserted into reporter vectors pZW2 or pZW4 using forward and reverse primers that contained the candidate sequences flanked by XhoI and ApaI sites. The two primers were annealed, digested, and ligated into the vectors. To make the random sequence library, the foldback primer was extended, then cut with XhoI and ApaI and ligated into pZW4. The heterologous reporter vector pZW8 contained exon 6 of the human *SIRT1* gene and portions of its flanking introns in place of the *DHFR* exon. The ESS sequences tested were inserted into the *SIRT1* exon by HindIII/KpnI digestion and ligation.

Cell Culture and Transfection of the Library

For stable transfection, cells were cotransfected with pOG44 that encodes the recombinase Flp (O'Gorman et al., 1991). To select stable transfectants, the cells were expanded by a 1 to 4 dilution 1 day after transfection and grown for 1 more day, and then hygromycin was added to a final concentration of 100 μg/ml. Resistant clones were trypsin digested, pooled, and analyzed by FACS using a Cytomation MoFlo high-speed sorter.

Exon and Intron Data Sets

The data sets of constitutively spliced or alternatively included/excluded (skipped) human exons were constructed using the ge-

nome annotation script GENOA (http://genes.mit.edu/genoa), which uses spliced alignment of cDNA and EST sequences to the human genome to annotate the locations of exons, as described (Yeo et al. 2004a). Intron and pseudoexon sequences were identified from a set of human introns downloaded from Ensembl (release 16.33), and filtered by BLASTN searches against the downloaded RepBase database (Jurka, 2000). Further details of these data sets are provided in Supplemental Data. All exon and intron data sets used are available upon request.

Statistical Analysis

The over- and underrepresentation of k-mers in sequences belonging to a set A relative to those of set B were analyzed using χ^2 statistics, with Yates' correction for small sample sizes. For convenience, we defined a χ statistic as the square root of the χ^2 statistic with a sign that is positive or negative according to whether the given k-mer was over- or underrepresented in set A versus set B. Thus the value of χ statistic indicates the degree enrichment of a k-mer in set A versus B (or B versus A, for negative values). To test the statistical significance for the ratio of *HPRT* exon-skipping mutations that create or disrupt ESSs, we randomly generated 10,000 sets of 176 hexamers and counted the numbers of *HPRT* mutations that created or disrupted hexamers from each of these sets. The p value was defined as the fraction of random sets of 176 hexamers that had a creation/disruption ratio greater than that observed for the FAS-hex2 hexamers ($14/3 \approx 4.67$), and similarly for the FAS-hex3 set.

ExonScan

The ExonScan splicing simulation algorithm takes as input a primary transcript sequence (excluding the initial and terminal exons) and predicts the locations of internal exons using a set of input scores for splice sites and (optionally) one or more sets of known or putative splicing *cis*-regulatory sequences. ExonScan first searches the transcript sequence for potential 5′ and 3′ splice sites (GT and AG dinucleotides, respectively) and scores all such potential sites using maximum entropy splice site models (Yeo and Burge, 2004). Candidate exons are defined as sequences 50–250 bases in length flanked by a potential 3′ss upstream and a potential 5′ss downstream. Each candidate exon is then scored for the presence of regulatory motifs (typically, enhancer scores are positive, and silencer scores are negative), and these scores are added to the splice site scores to obtain a final score for each candidate exon. A test set of 1820 human genes (containing 10,891 internal exons) was used to test ExonScan splicing simulations. The exact accuracy was defined in terms of the number of exons predicted exactly (both splice sites correctly predicted), and partial accuracy was defined in terms of the number of exons with at least one splice predicted correctly. Further details of the ExonScan algorithm are provided in Supplemental Data. An ExonScan web server is provided at http://genes.mit.edu/exonscan.

Acknowledgments

We thank P. Sharp and P. Grabowski for critical reading of our manuscript and D. Holste for help with sequence data sets. This material is based upon work supported by the National Science Foundation under grant no. 0218506 (C.B.B.), by the NIH (C.B.B.), and by a Damon Runyon Cancer Research Fellowship (Z.W.) and a Lee-Kuan-Yew graduate fellowship from Singapore (G.Y.).

Received: August 2, 2004
Revised: September 28, 2004
Accepted: November 2, 2004
Published: December 16, 2004

References

Amendt, B.A., Si, Z.H., and Stoltzfus, C.M. (1995). Presence of exon splicing silencers within human immunodeficiency virus type 1 tat exon 2 and tat-rev exon 3: evidence for inhibition mediated by cellular factors. Mol. Cell. Biol. *15*, 4606–4615.

Black, D.L. (2003). Mechanisms of alternative pre-messenger RNA splicing. Annu. Rev. Biochem. *72*, 291–336.

Buratti, E., Muro, A.F., Giombi, M., Gherbassi, D., Iaconcig, A., and Baralle, F.E. (2004). RNA folding affects the recruitment of SR proteins by mouse and human polypurinic enhancer elements in the fibronectin EDA exon. Mol. Cell. Biol. *24*, 1387–1400.

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.

Caputi, M., Mayeda, A., Krainer, A.R., and Zahler, A.M. (1999). hnRNP A/B proteins are required for inhibition of HIV-1 pre-mRNA splicing. EMBO J. *18*, 4060–4067.

Caputi, M., Kendzior, R.J., Jr., and Beemon, K.L. (2002). A nonsense mutation in the fibrillin-1 gene of a Marfan syndrome patient induces NMD and disrupts an exonic splicing enhancer. Genes Dev. *16*, 1754–1759.

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.

Chen, C.D., Kobayashi, R., and Helfman, D.M. (1999). Binding of hnRNP H to an exonic splicing silencer is involved in the regulation of alternative splicing of the rat beta-tropomyosin gene. Genes Dev. *13*, 593–606.

Cote, J., Dupuis, S., Jiang, Z., and Wu, J.Y. (2001). Caspase-2 pre-mRNA alternative splicing: identification of an intronic element containing a decoy 3′ acceptor site. Proc. Natl. Acad. Sci. USA *98*, 938–943.

Del Gatto-Konczak, F., Olive, M., Gesnel, M.C., and Breathnach, R. (1999). hnRNP A1 recruited to an exon in vivo can function as an exon splicing silencer. Mol. Cell. Biol. *19*, 251–260.

Domsic, J.K., Wang, Y., Mayeda, A., Krainer, A.R., and Stoltzfus, C.M. (2003). Human immunodeficiency virus type 1 hnRNP A/B-dependent exonic splicing silencer ESSV antagonizes binding of U2AF65 to viral polypyrimidine tracts. Mol. Cell. Biol. *23*, 8762–8772.

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

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.

Graveley, B.R. (2000). Sorting out the complexity of SR protein functions. RNA *6*, 1197–1211.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. Monatshefte f. Chemie *125*, 167–188.

Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., and Shoemaker, D.D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science *302*, 2141–2144.

Jurka, J. (2000). Repbase update: a database and an electronic journal of repetitive elements. Trends Genet. *16*, 418–420.

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.

Kohtz, J.D., Jamison, S.F., Will, C.L., Zuo, P., Luhrmann, R., Garcia-Blanco, M.A., and Manley, J.L. (1994). Protein-protein interactions and 5′-splice-site recognition in mammalian mRNA precursors. Nature *368*, 119–124.

Ladd, A.N., and Cooper, T.A. (2002). Finding signals that regulate alternative splicing in the post-genomic era. Genome Biol. *3*, reviews0008. Published online October 23, 2002. 10.1186/gb-2002-3-11-reviews0008

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.

Liu, H.X., Cartegni, L., Zhang, M.Q., and Krainer, A.R. (2001). A

mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. Nat. Genet. *27*, 55–58.

McCullough, A.J., and Berget, S.M. (1997). G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. Mol. Cell. Biol. *17*, 4562–4571.

O'Gorman, S., Fox, D.T., and Wahl, G.M. (1991). Recombinase-mediated gene activation and site-specific integration in mammalian cells. Science *251*, 1351–1355.

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

Si, Z., Amendt, B.A., and Stoltzfus, C.M. (1997). Splicing efficiency of human immunodeficiency virus type 1 tat RNA is determined by both a suboptimal 3′ splice site and a 10 nucleotide exon splicing silencer element located within tat exon 2. Nucleic Acids Res. *25*, 861–867.

Siebel, C.W., Fresco, L.D., and Rio, D.C. (1992). The mechanism of somatic inhibition of *Drosophila* P-element pre-mRNA splicing: multiprotein complexes at an exon pseudo-5′ splice site control U1 snRNP binding. Genes Dev. *6*, 1386–1401.

Singh, R., Valcarcel, J., and Green, M.R. (1995). Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. Science *268*, 1173–1176.

Sironi, M., Menozzi, G., Riva, L., Cagliani, R., Comi, G.P., Bresolin, N., Giorda, R., and Pozzoli, U. (2004). Silencer elements as possible inhibitors of pseudoexon splicing. Nucleic Acids Res. *32*, 1783–1791.

Smith, C.W., and Valcarcel, J. (2000). Alternative pre-mRNA splicing: the logic of combinatorial control. Trends Biochem. Sci. *25*, 381–388.

Staffa, A., and Cochrane, A. (1995). Identification of positive and negative splicing regulatory elements within the terminal tat-rev exon of human immunodeficiency virus type 1. Mol. Cell. Biol. *15*, 4597–4605.

Sterner, D.A., Carlo, T., and Berget, S.M. (1996). Architectural limits on split genes. Proc. Natl. Acad. Sci. USA *93*, 15081–15085.

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

Tu, M., Tong, W., Perkins, R., and Valentine, C.R. (2000). Predicted changes in pre-mRNA secondary structure vary in their association with exon skipping for mutations in exons 2, 4, and 8 of the Hprt gene and exon 51 of the fibrillin gene. Mutat. Res. *432*, 15–32.

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

Valentine, C.R., and Heflich, R.H. (1997). The association of nonsense mutation with exon-skipping in hprt mRNA of Chinese hamster ovary cells results from an artifact of RT-PCR. RNA *3*, 660–676.

Wagner, E.J., and Garcia-Blanco, M.A. (2001). Polypyrimidine tract binding protein antagonizes exon definition. Mol. Cell. Biol. *21*, 3281–3288.

Wilkinson, M.F., and Shyu, A.B. (2002). RNA surveillance by nuclear scanning? Nat. Cell Biol. *4*, E144–E147.

Wu, J.Y., and Maniatis, T. (1993). Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. Cell *75*, 1061–1070.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. *11*, 377–394.

Yeo, G., Holste, D., Kreiman, G., and Burge, C.B. (2004a). Variation in alternative splicing across human tissues. Genome Biol. *5*, R74.

Yeo, G., Hoon, S., Venkatesh, B., and Burge, C.B. (2004b). Variation in sequence and organization of splicing regulatory elements in vertebrate genes. Proc. Natl. Acad. Sci. USA, *101*, 15,700–15,705.

Zahler, A.M., Damgaard, C.K., Kjems, J., and Caputi, M. (2004). SC35 and heterogeneous nuclear ribonucleoprotein A/B proteins bind to a juxtaposed exonic splicing enhancer/exonic splicing silencer element to regulate HIV-1 tat exon 2 splicing. J. Biol. Chem. *279*, 10077–10084.

Zhang, X.H., and Chasin, L.A. (2004). Computational definition of sequence motifs governing constitutive exon splicing. Genes Dev. *18*, 1241–1250.

Zheng, Z.M. (2004). Regulation of alternative RNA splicing by exon definition and exon sequences in viral and mammalian gene expression. J. Biomed. Sci. *11*, 278–294.

Zheng, Z.M., Huynen, M., and Baker, C.C. (1998). A pyrimidine-rich exonic splicing suppressor binds multiple RNA splicing factors and inhibits spliceosome assembly. Proc. Natl. Acad. Sci. USA *95*, 14088–14093.

Zhu, J., Mayeda, A., and Krainer, A.R. (2001). Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. Mol. Cell *8*, 1351–1361.