Nucleic Acids Research, 2005, Vol. 33, No. 16 ***5053–5062***
doi:10.1093/nar/gki810

Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes

Jinhua Wang†, Philip J. Smith†, Adrian R. Krainer and Michael Q. Zhang*

Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA

Received April 12, 2005; Revised July 22, 2005; Accepted August 16, 2005

ABSTRACT

Exonic splicing enhancers (ESEs) are pre-mRNA *cis*-acting elements required for splice-site recognition. We previously developed a web-based program called ESEfinder that scores any sequence for the presence of ESE motifs recognized by the human SR proteins SF2/ASF, SRp40, SRp55 and SC35 (http://rulai.cshl.edu/tools/ESE/). Using ESEfinder, we have undertaken a large-scale analysis of ESE motif distribution in human protein-coding genes. Significantly higher frequencies of ESE motifs were observed in constitutive internal protein-coding exons, compared with both their flanking intronic regions and with pseudo exons. Statistical analysis of ESE motif frequency distributions revealed a complex relationship between splice-site strength and increased or decreased frequencies of particular SR protein motifs. Comparison of constitutively and alternatively spliced exons demonstrated slightly weaker splice-site scores, as well as significantly fewer ESE motifs, in the alternatively spliced group. Our results underline the importance of ESE-mediated SR protein function in the process of exon definition, in the context of both constitutive splicing and regulated alternative splicing.

INTRODUCTION

Processing of pre-mRNA is a fundamental aspect of gene regulation. Most eukaryotic genes comprise multiple relatively short exons that are separated by much longer introns. The basic mechanism of splicing involves exon recognition via the 5′ and 3′ splice sites and branch site at or near the intron ends, and the precise removal of intronic sequences and ligation of exons, generating mature mRNA (1). However, accurate exon definition by the spliceosome is complicated by the presence of numerous intronic pseudo exons flanked by sequences that conform to the splice-site consensus motifs at least as well as those utilized by many true exons (2). The additional information required for exon definition is contained at least partly in *cis*-acting regulatory enhancer and silencer sequences (3).

Exonic splicing enhancers (ESEs) participate in both alternative and constitutive splicing, and many of them act as binding sites for members of the SR protein family (4,5). The SR proteins are a family of related proteins that share a conserved domain structure. They have one or two copies of an RNA-recognition motif (RRM) followed by a C-terminal domain that is highly enriched in arginine/serine dipeptides (RS domain) (6). The RRMs mediate substrate recognition via sequence-specific RNA binding, whereas the RS domain is thought to be involved mainly in protein–protein interactions, but apparently also in protein–RNA interactions (7,8). Exon definition may occur through ESE-bound SR proteins recruiting components of the splicing machinery through their RS domains (9,10), and/or by antagonizing the action of nearby splicing silencer elements (11).

It has been estimated that at least 15% of point mutations that give rise to human genetic diseases cause RNA splicing defects (12). These mutations exert their effects upon the standard consensus intronic splice sites, and normally result in exon skipping, or less commonly in the creation of an ectopic splice site or activation of a cryptic splice site (12). The effects of exonic point mutations are less well understood. Until recently, it was normally assumed that nonsense mutations produce truncated protein isoforms or in some cases target the mRNA for destruction, whereas missense mutations were thought to identify amino acids that are important for protein structure or function. Translationally silent mutations were normally classified as polymorphisms and considered neutral. The generality of these assumptions is now being challenged, in part through the analysis of the mRNAs produced from mutant alleles, and this analysis is leading to the re-classification of a number of exonic mutations and to the realization that an even higher proportion of mutations affect splicing (3). One possible explanation for the effects of such

*To whom correspondence should be addressed. Tel: +1 516 367 8393; Fax: +1 516 367 8461; Email: Mzhang@cshl.edu

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

mutations is that they interfere with the function of exonic regulatory sequences. Indeed, recent data implicate ESE inactivation by point mutations as a significant cause of genetic disease (13–26).

Several groups have employed functional systematic evolution of ligands by exponential enrichment (SELEX) for the purpose of identifying sequences that can function as ESEs. Functional SELEX, both *in vivo* (27) and *in vitro* (28–30), has led to the discovery that a diverse array of both purine-rich and non-purine-rich sequences can act as ESEs. A further refinement of functional SELEX allowed the identification of sequence motifs that can act as ESEs in response to specific SR proteins (31,32). The motifs identified are short (6–8 nt), degenerate and sometimes partially overlap. The frequencies of the individual nucleotides at each position were used to derive score matrices that can be used to predict the location of SR protein-specific putative ESEs (31,32). The nucleotide-frequency matrices are available in a web-based program called ESEfinder (33). Previously, the matrices were used to examine a limited set of exon sequences for the presence of ESE motifs. Exonic high-score motifs were often found to be clustered and also to be enriched in regions with known natural enhancers (31,32). In addition, the motifs were found to be present at a higher density within exons, compared with introns. The predictive power of ESEfinder has been demonstrated through the observation that a number of disease-associated point mutations that result in exon skipping reduce high-score motifs to below threshold values (13,14,17, 20,22,24–26). Conversely, a mutation that results in activation of a cryptic 5′ splice site due to increased SC35 binding to an ESE, is consistent with the ESE scores predicted by ESEfinder (34).

Ab initio computational approaches to identify ESE motifs have recently been developed. RESCUE-ESE (35,36) identified putative ESE motifs by comparing hexanucleotide frequencies in constitutive exons with weak versus strong splice sites. Sequences preferentially associated with weak splice sites were clustered into several families and demonstrated to possess enhancer activity when functionally tested. A similar approach compared octamer frequencies from internal non-coding exons versus unspliced pseudo exons and the 5′-untranslated regions (5′-UTRs) of intronless genes, to identify putative regulatory sequences involved in splicing (37). This approach led to the discovery of both functional enhancer and silencer sequences.

We have undertaken a large-scale analysis of SR-protein-dependent ESE motif frequencies in the human genome using ESEfinder. A thorough survey of ESE prevalence was warranted, in light of the high percentage of mutations that cause genetic diseases through aberrant splicing. In addition, a genome-wide survey of ESE motifs in protein-coding genes can give an indication of their importance in constitutive and alternative splicing, and their overall contribution to exon definition and splice-site selection.

MATERIALS AND METHODS

Database creation

The EnsMart search engine (38) was used to retrieve human genomic sequence from Ensembl (version 24) (39). A set of 63 218 constitutively spliced internal protein-coding exons plus 100 nt each of flanking upstream and downstream intronic sequence, was derived from a total of 12 216 genes. Constitutive exons were defined from genes having definitive annotation in the NCBI Reference Sequence (RefSeq) collection, whose transcripts demonstrated no evidence of alternative splicing. Protein-coding exons were derived by BLAST searching of exons with cDNA sequences, allowing the elimination of non-coding and partially coding exons. We also created a database of 2620 alternatively spliced (cassette) exons from RefSeq genes with multiple transcripts, by mapping exons from these genes to their respective genomic coordinates. For comparison with the alternative exons, we created a set of 2880 constitutive exons selected to have a similar length distribution (same mean and standard deviation of exon lengths). A database of 20 580 repeat-free intronic pseudo exons was kindly provided by Dr Lawrence Chasin (37). Sequence databases are available upon request.

Sequence analysis

ESE motif scores were calculated using the position weight matrices available in ESEfinder version 2.0 (http://rulai.cshl.edu/tools/ESE/) (33). The default threshold values from the program were used. For the purposes of this study, we considered only above-threshold (high-score) ESE motifs as being significant. These thresholds were defined previously as the median of the highest score for each sequence in a set of randomly chosen 20 nt sequences from the starting pool used for the functional SELEX experiments (33). Note that the motif scores for different SR proteins are not directly comparable (33). Shuffled exonic and intronic sequences were generated using the EMBOSS Shuffleseq program (http://emboss.sourceforge.net/apps/shuffleseq.html).

Splice-site scores were calculated using score matrices derived from the exon-finding program MZEF (40). The matrices are based on position-dependent triplet-frequency preferences for real versus pseudo splice sites in the window (−15, +3) for 3′ splice sites and (−3, +8) for 5′ splice sites.

Statistical analysis

Bootstrap sampling was used to determine the level of significance for the differences in average ESE motif frequencies between exons and their flanking introns, and exons and pseudo exons. The mean ESEs/nt from random selections of 10 000 sequences from the exon, intron and pseudo exon groups were sampled and compared 5000 times to derive *P*-values. ESE motif frequency distributions were compared by quantile–quantile analysis, and median values were compared by the two-sample *t*-test. The significance of the overlap between motifs recognized by ESEfinder and RESCUE-ESE or the putative ESEs of Zhang and Chasin was defined by Fisher's exact test. Statistical tests with *P*-values <0.01 were deemed significant.

RESULTS

ESE motif frequencies in constitutive exons

To date, most studies of ESE function have concentrated on their role in alternative splicing, although functional ESEs are

also present in constitutive exons (13,14,41,42). Important questions remain unanswered, including the extent to which ESEs participate in the process of constitutive splicing. A large-scale analysis of ESE motif distribution in both exons and introns would give some indication of their functional relevance to splicing events of this nature.

We created a database of 63 218 constitutively spliced internal protein-coding exons of lengths ⩾100 nt from 12 216 human genes. To standardize for differences in exon length, we created composite 100 nt exon sequences consisting of 25 nt from each end plus 50 nt from the center. To ensure that the exons were constitutively spliced, sequences were collected from single-transcript genes. Exonic sequences plus 100 nt each of flanking upstream and downstream intronic sequences were retrieved from Ensembl. For comparison, we calculated ESE motif frequencies from a database of 20 580 repeat-free intronic pseudo exons (37) also standardized to 100 nt, plus 100 nt each of 5′- and 3′-flanking sequences. ESEfinder scores sequences for the presence of motifs matching the SELEX-derived consensus for four SR proteins: SF2/ASF, SRp40, SRp55 and SC35 (33). We calculated high-score ESE motif frequencies occurring at each position in consecutive windows of 10 nt. For the purposes of this study, all above-threshold values for a given motif were considered to be equivalent.

The ESE motif frequency distributions (ESEs/10 nt) were plotted separately for each SR protein (Figure 1). Points were plotted at the central position of the high-score motif. ESE motif frequencies were higher within exons than in the flanking intronic sequences for all four SR proteins. Sharp peaks and troughs at the exon/intron borders are a consequence of the conserved splice-site sequences. To avoid the contribution of the splice-site consensus motifs, we calculated the mean ESE motif frequencies (ESEs/nt) at the exact center of the exons and each of the flanking intronic regions (50 nt upstream of the 3′ splice-site, and 50 nt downstream of the 5′ splice-site) (Table 1). A bootstrap sampling strategy of the mean ESE motif frequencies revealed that the higher density of ESE motifs in exons than in introns was statistically significant for all four SR proteins, and the *P*-values were all <0.001 for comparisons with both upstream and downstream flanks. ESE motif frequencies were approximately constant within exons. By comparison, ESE motif frequencies in pseudo exons were significantly lower than in authentic exons for

Table 1. Mean ESE motif frequencies in constitutive exons, introns, and pseudo exons (ESEs/nt)

SR protein	Upstream intron	Exon	Downstream intron	Upstream flank	Pseudo exon	Downstream flank
SF2/ASF	0.0376	0.0496	0.0402	0.0387	0.0425	0.0431
SC35	0.0412	0.0448	0.0420	0.0422	0.0435	0.0424
SRp40	0.0398	0.0436	0.0405	0.0410	0.0426	0.0415
SRp55	0.0401	0.0442	0.0387	0.0432	0.0427	0.0426

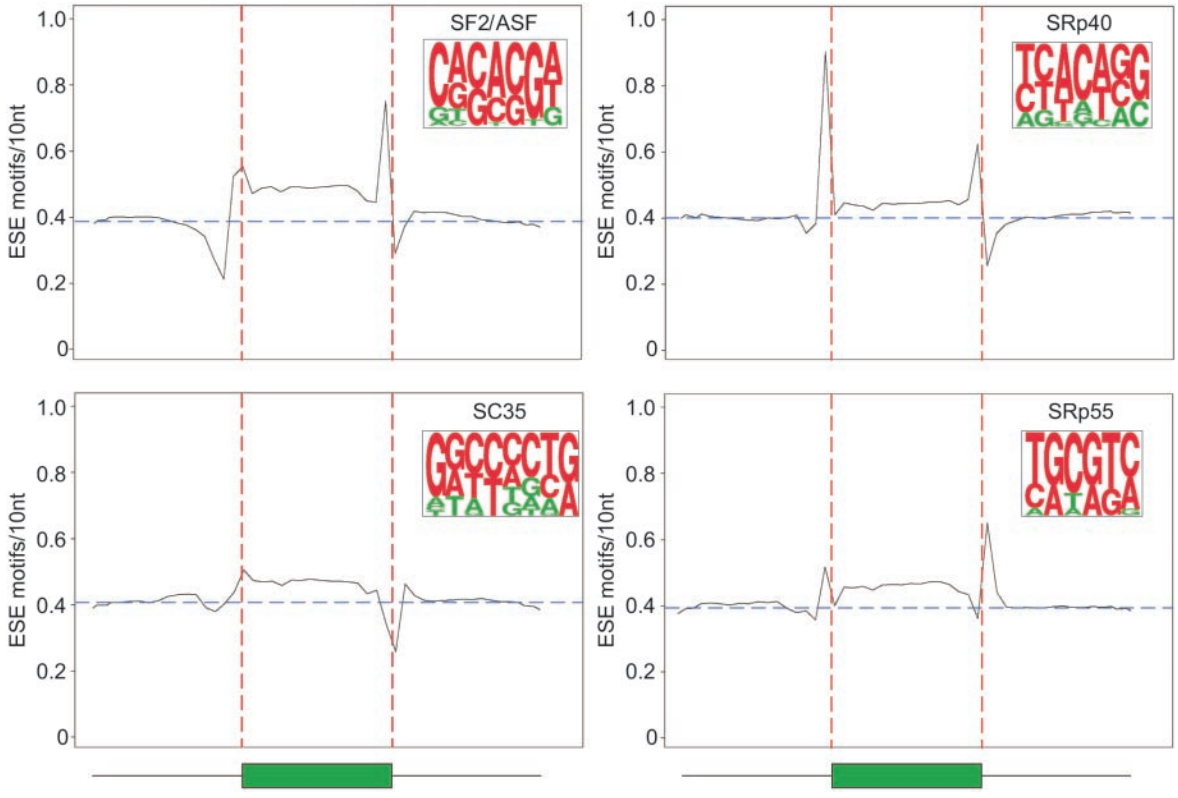

Figure 1. ESE motif frequency distributions in constitutive coding exons and flanking introns. ESEfinder was used to analyze 63 218 constitutive coding exons ⩾100 nt in length for the presence of high-score ESE motifs. The green box represents a composite exon standardized to 100 nt, as described in the text. The thin lines represent 100 nt each of flanking upstream and downstream intronic sequence. ESE motif scores were measured at each position in windows of 10 nt, and high-score motifs plotted at the central position of the motif. Exon/intron boundaries are indicated by the red vertical dashed lines. The blue horizontal dashed line represents the mean intronic ESE motif density. The consensus motif derived from functional SELEX is shown for each SR protein. Red letters indicate above-background nucleotide frequencies.

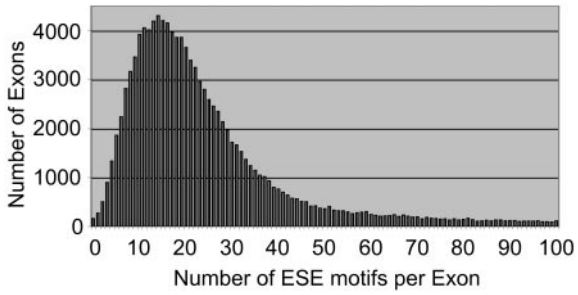

Figure 2. Frequency distribution of ESE motifs in constitutive exons. ESEfinder was used to score 63 218 internal protein-coding exons (≥100 nt) for the presence of high-score ESE motifs for SF2/ASF, SRp40, SRp55 and SC35.

three of the four SR proteins (Table 1) (*P*-values <0.002 for SF2/ASF, <0.01 for SC35, <0.006 for SRp40 and <0.02 for SRp55). The frequencies of ESE motifs in intronic pseudo exons were similar to the frequencies found in the other intronic regions analyzed. As a control, we shuffled the exonic and intronic sequences, maintaining the nucleotide composition, and scored the resulting sequences with ESEfinder. The frequency of ESE motifs in the shuffled exonic sequences decreased for all four of the SR proteins, whereas the frequencies in shuffled intronic sequences were higher than in the real intronic sequences (data not shown). This provides further evidence for the functionality of the ESEfinder motifs when present at exonic locations.

We observed a wide variation in the absolute numbers of ESE motifs per exon when we analyzed the complete exons in our constitutive exon database (Figure 2). The exons ranged in size from 100 nt to ~6 kb, and there was a modal frequency of 14 ESE motifs per exon. Interestingly, a small number of exons (158) contained no ESE motifs, although it should be emphasized that the current version of ESEfinder searches for high-score motifs for only 4 of the ~10 SR proteins.

Correlation of ESE motif frequencies with splice-site strength

It has been postulated that one function of ESEs is the recruitment of spliceosomal components to weak 5′ or 3′ splice sites (43). Therefore, it is possible that exons with weak splice sites will have elevated frequencies of ESEs. This property was one of the criteria used to identify ESE motifs by RESCUE-ESE (35). We chose to investigate this hypothesis in the context of constitutive splicing, to eliminate as far as possible any complications arising from mechanisms regulating alternative splicing.

We calculated the 5′ and 3′ splice-site values for each exon in our constitutive exon database. We then ranked the exons as strong (top 15%) or weak (bottom 15%) for 5′ and 3′ splice sites independently. ESEfinder was used to calculate high-score ESE motifs from the four groups of exons. The number of high-score ESE motifs was divided by exon length to give ESEs/nt, and the frequency distributions were plotted as number of exons versus ESEs/nt. The ESE motif frequency distributions of the exons with strong and weak 3′ splice sites were compared, as were the distributions of the exons with strong and weak 5′ splice sites, by quantile–quantile analysis (Supplementary Figure 1). This type of analysis determines if two datasets come from populations with a common distribution. If the strong and the weak splice-site score exons have the same distribution of ESE motifs, then the points will fall approximately on the 45° reference line. Departure from the 45° reference line, either below or above, indicates higher ESEs/nt in exons with strong or weak splice sites, respectively. Differences in the ESE motif frequency distributions were observed between exons with weak and strong splice sites, for both 5′ and 3′ splice sites, for some of the SR proteins. A summary of the data is shown in Table 2. The correlation of ESE frequencies with splice-site strength reveals a complicated relationship. For most of the comparisons, there are no significant differences between exons with strong versus weak splice sites. However, exons with weak 5′ splice sites and exons with weak 3′ splice sites have more SRp55 and SF2/ASF motifs, respectively. In contrast, exons with strong 5′ splice sites have significantly more SRp40 motifs than their weak splice-site counterparts, and exons with strong 3′ splice sites have significantly more SC35 and SRp40 motifs.

Table 2. Mean ESE motif frequencies in exons with strong or weak splice sites (ESEs/nt)

SR protein	Strong 5′	Weak 5′	Strong 3′	Weak 3′
SF2/ASF	0.0435	0.0434	0.0400	**0.0425**
SC35	0.0417	0.0424	**0.0427**	0.0398
SRp40	**0.0444**	0.0427	**0.0430**	0.0415
SRp55	0.0241	**0.0258**	0.0240	0.0244

Frequencies in boldface are significantly higher than in the other exon set in a pairwise comparison (strong versus weak) (two-sample *t*-test, $P < 0.01$)

We further classified our constitutive exon dataset into strong exons possessing both strong 5′ and 3′ splice sites (top 15%), or weak exons possessing both weak 5′ and 3′ splice sites (bottom 15%). Quantile–quantile analysis (Figure 3) of the ESE motif frequency distributions of these two groups of exons revealed significant differences in ESE motif prevalence for three of the SR proteins: exons with strong splice sites have more SC35 and SRp40 motifs, whereas exons with weak splice sites have more SRp55 motifs (Table 3). Therefore, there does not appear to be a simple correlation between ESE motif frequencies and splice-site strengths. When we combined the output of all four matrices for the exon datasets in Tables 2 and 3, the differences between strong and weak exons were averaged out (data not shown). Our observations with the individual matrices suggest a potential role for a subset of the motifs and corresponding SR proteins in the recognition of exons associated with weak splice sites.

Comparison of ESE motif frequencies in constitutive versus alternatively spliced exons

Alternative splicing events have previously been documented to be associated with weak splice sites (44), traditionally on a single transcript basis. Such a correlation is limited by the lack of large-scale analyses. One recent report analyzed relatively large datasets of both 5′ and 3′ splice site scores from constitutive and alternative exons from a number of different species, and found consistently higher scores for the constitutive exons (45). However, the link between splice-site score and alternative splicing remains unclear, and may not reflect a

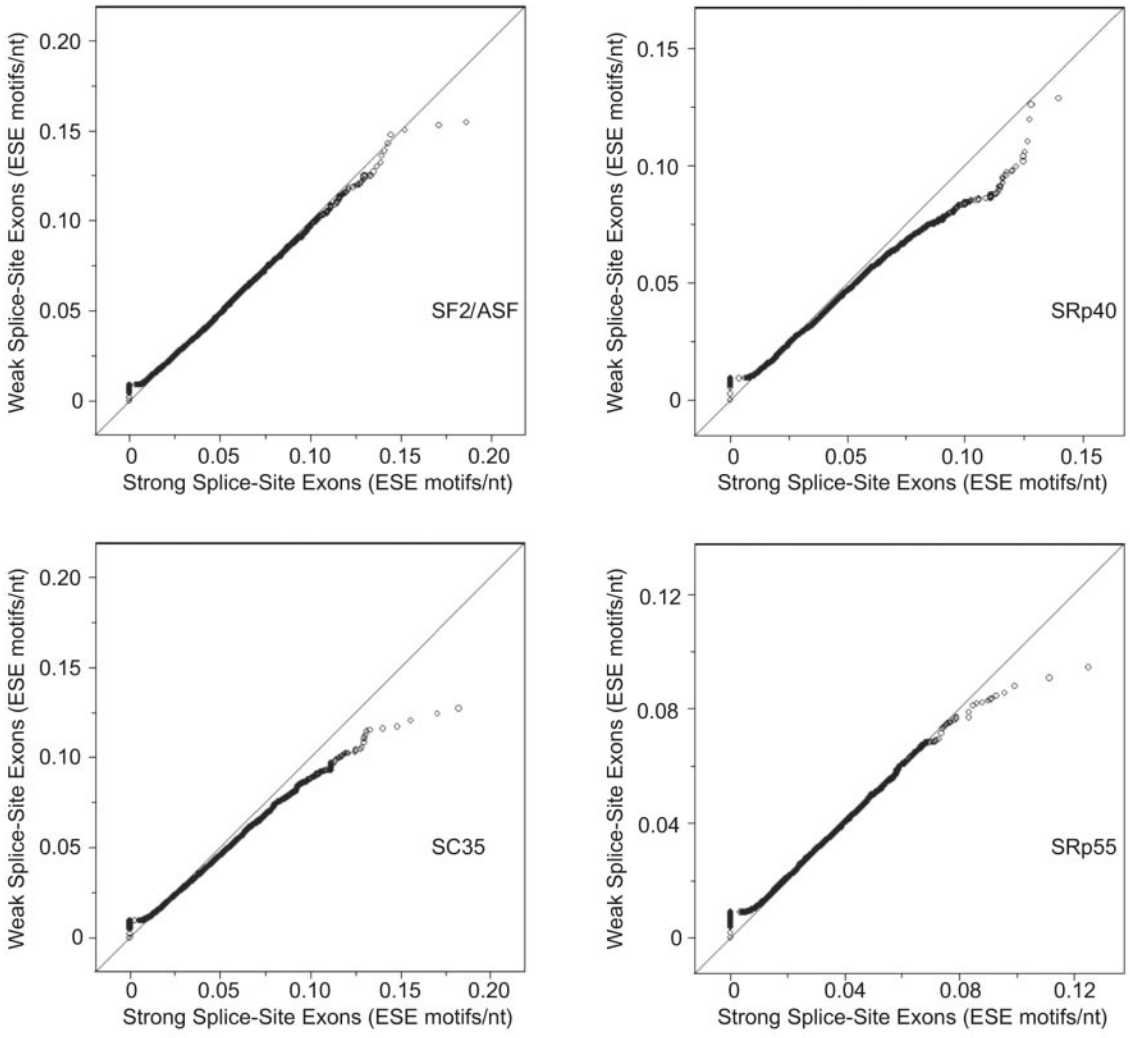

Figure 3. Correlation of ESE motif frequencies with splice-site strength. Constitutive exons were classified as weak, if both their 3′ and 5′ splice-site scores were in the bottom 15%, or strong, if both their 3′ and 5′ splice-site scores were in the top 15%. ESE motif frequency distributions from the two exon groups were compared by quantile–quantile analysis.

Table 3. Mean ESE motif frequencies in exons with both strong 5′ and 3′ splice sites and exons with both weak 5′ and 3′ splice sites (ESEs/nt)

SR protein	Strong	Weak
SF2/ASF	0.0429	0.0421
SC35	**0.0440**	0.0404
SRp40	**0.0447**	0.0416
SRp55	0.0238	**0.0252**

Frequencies in boldface are significantly higher than in the other exon set in a pairwise comparison (strong versus weak) (two-sample *t*-test, $P < 0.01$).

simple relationship. The results of our studies of splice-site score and ESE motif frequencies in constitutive exons led us to investigate the corresponding frequencies in alternative exons, and their correlation with alternative splicing events. There are several forms of alternative splicing [reviewed in (46)], and for simplicity we chose to investigate the most common one, namely exon skipping/inclusion.

We created a database of 2620 skipped internal protein-coding exons from RefSeq genes with multiple transcripts, and scored them with ESEfinder. High-score ESE motifs were divided by exon length to give ESEs/nt. This analysis was repeated on a set of 2880 constitutive exons selected to have a similar length distribution (same mean and standard deviation of exon lengths). ESE motif frequency distributions were derived and compared by quantile–quantile analysis (Figure 4). Departure from the 45° reference line, either below or above, indicates higher ESEs/nt in constitutive or skipped exons, respectively. Scoring for all four SR proteins combined revealed that ESE motif frequencies were significantly lower in skipped compared with constitutive exons, with median values of 0.1466 and 0.1605 ESEs/nt, respectively (two-sample *t*-test, $P < 0.00001$). The same result was obtained when the exons were scored for individual SR proteins. For example, skipped and constitutive exons scored for SF2/ASF motifs had median values of 0.0384 and 0.0421 ESEs/nt, respectively ($P < 0.0001$).

The observation that skipped exons had significantly fewer ESE motifs than constitutive exons led us to examine the ESE motif frequency distribution in the flanking intronic regions of skipped exons. We used ESEfinder to score 100 nt each of

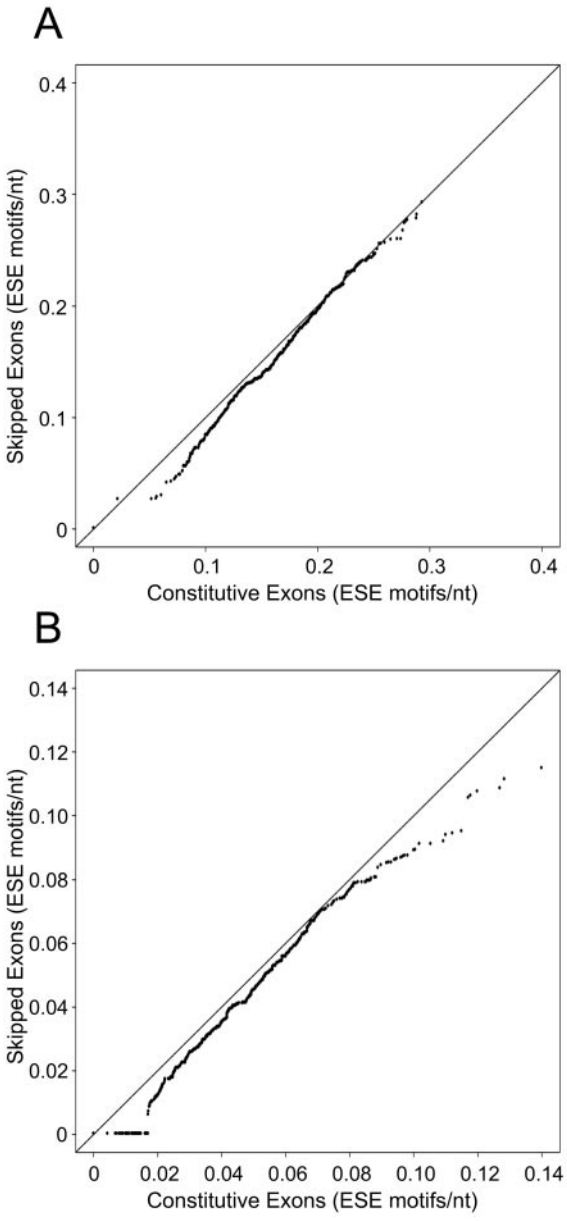

Figure 4. ESE motif frequencies in constitutive versus alternative exons. ESEfinder was used to calculate ESE motif frequencies in 2620 skipped exons and 2880 constitutive exons of comparable length. ESE motif frequency distributions (ESEs/nt) were compared by quantile–quantile analysis for all four SR proteins together (**A**), and for SF2/ASF alone (**B**).

flanking upstream and downstream intronic sequence. Mean ESE motif frequencies (ESEs/nt) at the exact center of the skipped exons and each of their flanking introns (50 nt upstream of the 3′ splice-site, and 50 nt downstream of the 5′ splice-site) were calculated (Table 4). Bootstrap resampling of the mean frequencies demonstrated that only the SF2/ASF motifs were significantly higher in the skipped exons compared with their flanking introns ($P < 0.001$ for comparison with upstream intron, $P < 0.003$ for comparison with downstream intron).

Calculation of the splice-site scores using position weight matrices (40) revealed that the skipped exons had significantly weaker splice sites than the constitutively spliced exons.

Table 4. Mean ESE motif frequencies in skipped exons and their flanking introns (ESEs/nt)

SR protein	Upstream intron	Exon	Downstream intron
SF2/ASF	0.0387	0.0418	0.0403
SC35	0.0399	0.0402	0.0406
SRp40	0.0402	0.0398	0.0395
SRp55	0.0407	0.0409	0.0404

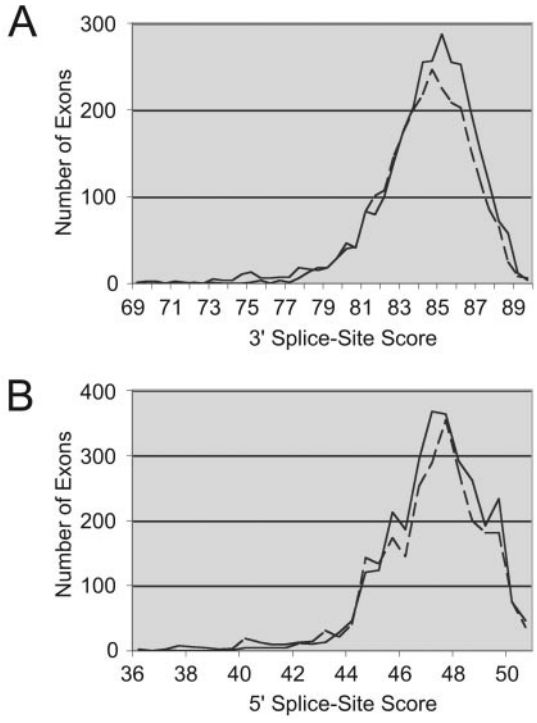

Figure 5. Splice-site score distributions of constitutively spliced and skipped exons. Splice-site scores of 2620 skipped exons (dotted lines) and 2880 constitutive exons (solid lines) were calculated and their frequency distributions plotted separately for 3′ (**A**) and 5′ (**B**) splice-site scores.

The mean values with standard deviations were 84.2 ± 2.25 and 83.7 ± 2.7 for constitutive and skipped 3′ splice sites, respectively, and 46.93 ± 1.7 and 46.67 ± 2.09 for constitutive and skipped 5′ splice sites, respectively. These values were significantly different (all P-values <0.01) when analyzed by both parametric (one sample t-test) and nonparametric (Wilcoxon rank test) statistical methods. It should be noted that although the mean splice-site scores are significantly different, the distributions of splice-site scores for both exon types are very similar (Figure 5), and that splice-site scores alone are insufficient to identify an exon as one that is alternatively spliced.

ESE motif recognition by ESEfinder and *ab initio* ESE-prediction methods

Two recent reports employed *ab initio* computational methods to predict sequences that have ESE activity. RESCUE-ESE, developed by the Burge laboratory (35), identified 238 hexamers preferentially associated with constitutive exons with weak splice sites, whereas the methodology of Zhang and Chasin (37) identified octamers overrepresented in

Table 5. Comparison of ESE motif recognition by ESEfinder and RESCUE-ESE

SR protein	Number of sequences	ESEfinder high scores	Expected number of high scores
SF2/ASF	1904	79	79
SC35	11 424	179	453
SRp40	1904	64	78
SRp55	238	4	8
Total	15 470	326	618

Table 6. Comparison of ESE motif recognition by ESEfinder and the putative ESEs of Zhang and Chasin

SR protein	Number of sequences	ESEfinder high scores	Expected number of high scores
SF2/ASF	4138	263	171
SC35	2069	80	82
SRp40	4138	166	169
SRp55	6207	185	202
Total	16 552	694	624

non-protein-coding exons compared with the 5′-UTR of intronless genes and pseudo exons. Both groups tested a number of candidate motifs and demonstrated enhancer function in transfected cells. Although these two methods and ESEfinder differ substantially, there may be some overlap in the sequences they recognize as putative ESEs.

The functional SELEX-derived consensus motifs are a hexamer for SRp55, heptamers for SF2/ASF and SRp40, and an octamer for SC35 (33). Because the sequences identified by RESCUE-ESE are hexamers, we expanded each RESCUE-ESE motif by the addition of either 1 (for SF2/ASF and SRp40) or 2 (for SC35) nt, and scored the resulting sequences with ESEfinder. As a control, we calculated the number of all possible ESEfinder high-score motifs for each SR protein. Of all 16 384 heptamers, 678 (4.1%) are high-score SF2/ASF motifs; 669 (4.1%) of all heptamers are high-score SRp40 motifs; 2599 (4.0%) of all 65 536 octamers are high-score SC35 motifs; and 133 (3.2%) of all 4096 hexamers are high-score SRp55 motifs. Using these percentages, we then calculated the expected number of ESEfinder high-score motifs from a complete random sample of all possible oligonucleotide sequences equal in length to the test set of RESCUE-ESE sequences. For example, for SF2/ASF and SRp40 (heptamer consensus motifs), the addition of 1 nt at either the beginning or the end of the RESCUE-ESE hexamers results in $(4 \times 238) \times 2 = 1904$ sequences. We then calculated the expected number of ESEfinder high-score motifs from 1904 random heptamers. The results of the comparison (Table 5) indicate that the sequences recognized as ESE motifs by RESCUE-ESE and ESEfinder do not overlap beyond what is expected by chance.

A similar strategy was employed to investigate the extent of overlap between the sequences recognized by ESEfinder and the 2069 putative ESEs (PESEs) identified by Zhang and Chasin (37). The PESEs were downloaded (http://www.columbia.edu/cu/biology/faculty/chasin/xz3/octamers.txt) and high-score ESE motifs calculated with ESEfinder. As a control, we calculated the expected number of high-score ESEfinder motifs from a random sample of all possible oligonucleotide sequences equal in length to the test set of sequences. For example, there are two possible heptamers contained within any one octamer; therefore, for SF2/ASF and SRp40, we calculated the expected number of high-score ESEfinder motifs from $2 \times 2069 = 4138$ random heptamers. The results for SC35, SRp40 and SRp55 (Table 6) reveal that high-score ESE motifs for these three proteins are not enriched within the PESE set. However, there are significantly more SF2/ASF motifs (Table 6) than would be expected by chance ($P < 0.00001$, Fisher's exact test), supporting the conclusion that there is some overlap between the sequences identified as ESE motifs by these two very different methods.

DISCUSSION

The importance of *cis*-regulatory sequences for accurate splice-site recognition and exon definition is well documented. However, most experimental studies to date have focused on the regulation of single splicing events. A more global understanding of pre-mRNA splicing requires some knowledge of the distribution of both splicing enhancers and silencers. Using ESEfinder (33), we have undertaken a large-scale genomic analysis in an attempt to uncover relationships between ESE motif frequencies and splicing regulation. Many of the experimental studies of ESE function have involved examination of their role in the regulation of alternative splicing, and as such little is known about their functional relevance to the process of constitutive splicing. Our studies implicate ESE participation in the regulation of both constitutive and alternative splicing.

Previously, the SR protein-specific matrices utilized by ESEfinder were used to search a limited set of genomic sequences for ESE motifs, which were found to occur more frequently in exons versus introns (31,32,47). We have greatly expanded these initial observations, and demonstrated a significant enrichment for ESE motifs in >60 000 internal constitutive protein-coding human exons. The motifs identified by the RESCUE-ESE technique (35) and the PESEs of Zhang and Chasin (37) also occur more frequently in exons versus introns. ESEfinder motif frequencies within exons were approximately constant, supporting the hypothesis that ESEs function to activate splicing from varying distances from the splice sites, an observation also made for the exonic distribution of PESEs (37). In addition, constant ESE motif frequencies along exons may be a consequence of the ability of single enhancer motifs to influence recognition of both 3′ and 5′ splice sites (43,48,49). The functional SELEX experiments used to derive the ESEfinder matrices were dependent upon the ability of sequences to enhance splicing of a 3′ terminal exon (31,32). However, numerous studies have implicated ESE motifs identified by ESEfinder in the splicing of internal exons (13–17,20–26,34) and our new data support the conclusion that these ESE motifs play a role in the splicing of internal exons, in addition to terminal exons.

ESE motif frequencies for three of the four SR proteins were significantly higher in exons versus pseudo exons, supporting a role for ESEs in exon definition, and consistent with previous studies of genomic ESE motif distributions (37,47). Zhang and Chasin (37) found fewer PESEs in the same set of pseudo exons that we analyzed with ESEfinder, but identification

of the PESE motifs was conditional on their overrepresentation in exons versus pseudo exons. Therefore, the observation that the PESE motifs were more frequent in a second test set of exons versus pseudo exons was a logical expectation (37). The functional SELEX experiments used to derive the ESEfinder motifs imposed no such a priori criteria; therefore, the fact that these motifs are present at significantly higher frequencies in exons versus pseudo exons supports the conclusion that they are involved in exon definition. In addition, there is evidence supporting a role for silencers in the suppression of pseudo exon splicing: a subset of pseudo exons with a relatively high frequency of ESEfinder motifs was found to have increased frequencies of elements capable of silencing splicing (47); and Zhang and Chasin (37) also observed overrepresentation of putative exonic splicing silencers in pseudo exons.

Experimental evidence demonstrated a role for ESEs in constitutive splicing (13,14,41,42), a function supported by our bioinformatic analysis. One ascribed function of ESEs is facilitating the recognition of suboptimal splice sites. Indeed, improving weak 3′ splice-site polypyrimidine tracts negates the enhancer requirement for a number of substrates (50,51). However, there is no evidence that all exons with weak splice sites have an increased dependence upon ESEs. Our comparison of ESE motif frequencies in constitutive exons with weak and strong splice sites implicates ESE involvement in splice-site recognition of all exons. We observed significant differences in some ESE motif frequencies when constitutive exons with strong and weak 3′ or 5′ splice sites were compared independently, or when exons with both strong 3′ and 5′ splice sites were compared with their counterparts with weak sites. However, there was not a simple relationship between splice-site score and ESE motif frequency, as in some instances exons with strong splice sites were found to contain more ESE motifs. In addition, when we repeated this analysis using Zhang and Chasin's PESEs, we observed no difference in the frequency of PESEs in exons with weak splice sites compared with those with strong splice sites (data not shown). It remains possible that weak splice sites tend to be associated with stronger ESEs, rather than with an increased number of ESEs, although it is known that multiple ESEs in the same exon act additively (52). This hypothesis remains to be tested, and will require a more quantitative version of ESEfinder.

A recent survey revealed an increase in the number of ESE motifs identified by RESCUE-ESE in the vicinity of the splice sites of constitutive exons (53). We only observed this trend with SF2/ASF and SRp55 motifs in exons with weak 3′ and 5′ splice sites, respectively. As described above, ESE motifs for some of the SR proteins are actually higher in exons with strong splice sites. These differences in ESE motif distributions may be a consequence of the very different methods used in their identification. The motifs identified as putative enhancers by RESCUE-ESE were constrained by the requirement to be enriched in constitutive exons with weak splice sites, whereas the sequences identified by functional SELEX were selected by their ability to activate exon inclusion in the presence of a particular SR protein. It is possible that RESCUE-ESE identified a set of enhancer sequences involved in the recognition of a restricted set of exons, and that ESEfinder recognizes enhancers involved in a more general aspect of exon definition.

Alternative splicing serves to greatly expand the proteome, with one recent report estimating that up to 74% of multiexon human genes are alternatively spliced (54). ESEs, and the SR proteins that bind them, have well defined roles in regulating the process of alternative splicing [reviewed in (1,4,5,44,55)]. A commonly held assumption states that exons that undergo alternative splicing have weaker splice sites, by comparison with those that are constitutively spliced. Our previous analysis of a limited set of alternatively spliced exons supported this assumption (56). In addition, a recent report found significantly higher splice-site scores for constitutive versus alternative exons in five species, including humans (45). We derived large datasets of constitutive and alternatively spliced (included or skipped) protein-coding human exons, and again demonstrated that alternatively spliced exons as a set have significantly weaker splice-site scores. However, the splice-site score distributions are surprisingly similar and largely overlapping, such that the splice-site scores alone are not sufficient to define a given exon as constitutive or alternative.

Intriguingly, we found that skipped exons have significantly fewer ESE motifs than constitutively spliced exons. In addition, skipped exons, unlike those that are constitutively spliced, do not have increased ESE motif frequencies in comparison with their flanking intronic regions, except for one of the four SR proteins tested, SF2/ASF. Zhang and Chasin (37) likewise reported finding fewer PESEs in alternative exons compared with constitutive exons, and a comparable number or slightly fewer RESCUE-ESE motifs were observed in skipped exons (35). One can speculate that fewer ESEs per exon may result in less efficient exon definition, and subsequently lead to exon skipping. However, this remains a hypothesis that will require appropriate experimental validation. Two recent publications (57,58) reported significant conservation of the flanking intronic regions of alternatively spliced exons, perhaps implying a function for intronic motifs in the control of alternative exon definition.

ESE motif identification by functional SELEX, and the computational methods of RESCUE-ESE or Zhang and Chasin's octamer analysis rely upon different methodologies. However, the motifs identified share some commonalities, namely overrepresentation in exons versus introns, and in constitutive versus alternatively skipped exons. Interestingly, our analysis revealed that the ESE motifs recognized by ESEfinder and RESCUE-ESE do not significantly overlap. Nevertheless, experimental data proved the ability of both methods to define functional enhancers (31,32,35), and as described above, these differences may arise at least in part from the constraint of association with weak splice sites inherent in RESCUE-ESE. Over 80% of the RESCUE-ESE hexamers are found in the collection of PESEs (37). However, in contrast to the analysis of RESCUE-ESE motif distribution (53), there was no increase in PESE frequency near the splice sites (37). This difference may be due to differences in the exonic databases analyzed, or it may be a consequence of a small subset of the RESCUE-ESE motifs accounting for the observed increase near splice sites. Our scoring of Zhang and Chasin's PESEs with ESEfinder revealed no enrichment for high-score SC35, SRp40 or SRp55 motifs. However, we did find an increase over the expected number of SF2/ASF motifs within the PESE group, indicating some overlap between the two methods. It should be noted that our analysis is limited to

four SR proteins, and it is highly probable that both the set of RESCUE-ESE hexamers and the PESE octamers contain enhancer sequences recognized by other SR and non-SR proteins, though these methods do not identify the factors responsible for motif recognition.

ESEfinder scores sequences for the presence of putative enhancers, and we emphasize that experimental validation is required for definitive proof that any given motif is a bona fide ESE in its natural context. Other factors may influence the ESE potential of any given motif. These include sequence context, e.g. the presence of nearby silencers, secondary structure effects and tissue-specific splicing factor concentrations. Experimental efforts are underway to refine the original matrices. Future improvements will include experimental refinement of threshold values, and additional SR protein-specific matrices.

SUPPLEMENTARY DATA

Supplementary Data is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Lawrence Chasin for kindly providing the pseudo exon database. This work was supported by NIH grants GM42699 to A.R.K. and HG01696/CA88351 to M.Q.Z, and by a postdoctoral fellowship from the US Army Medical Research and Matériel Command to P.J.S. The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Hastings,M.L. and Krainer,A.R. (2001) Pre-mRNA splicing in the new millennium. *Curr. Opin. Cell. Biol.*, **13**, 302–309.
2. Sun,H. and Chasin,L.A. (2000) Multiple splicing defects in an intronic false exon. *Mol. Cell. Biol.*, **20**, 6414–6425.
3. Cartegni,L., Chew,S.L. and Krainer,A.R. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Rev. Genet.*, **3**, 285–298.
4. Blencowe,B.J. (2000) Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem. Sci.*, **25**, 106–110.
5. Graveley,B.R. (2000) Sorting out the complexity of SR protein functions. *RNA*, **6**, 1197–1211.
6. Birney,E., Kumar,S. and Krainer,A.R. (1993) Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors. *Nucleic Acids Res.*, **21**, 5803–5816.
7. Shen,H., Kan,J.L. and Green,M.R. (2004) Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly. *Mol. Cell*, **13**, 367–376.
8. Shen,H. and Green,M.R. (2004) A pathway of sequential arginine-serine-rich domain-splicing signal interactions during mammalian spliceosome assembly. *Mol. Cell*, **16**, 363–373.
9. Zuo,P. and Maniatis,T. (1996) The splicing factor U2AF35 mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing. *Genes Dev.*, **10**, 1356–1368.
10. Graveley,B.R., Hertel,K.J. and Maniatis,T. (2001) The role of U2AF35 and U2AF65 in enhancer-dependent splicing. *RNA*, **7**, 806–818.
11. Kan,J.L. and Green,M.R. (1999) Pre-mRNA splicing of IgM exons M1 and M2 is directed by a juxtaposed splicing enhancer and inhibitor. *Genes Dev.*, **13**, 462–471.
12. Krawczak,M., Reiss,J. and Cooper,D.N. (1992) The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum. Genet.*, **90**, 41–54.
13. 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. *Nature Genet.*, **27**, 55–58.
14. 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. *Nature Genet.*, **30**, 377–384.
15. 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.
16. Aznarez,I., Chan,E.M., Zielenski,J., Blencowe,B.J. and Tsui,L.C. (2003) Characterization of disease-associated mutations affecting an exonic splicing enhancer and two cryptic splice sites in exon 13 of the cystic fibrosis transmembrane conductance regulator gene. *Hum. Mol. Genet.*, **12**, 2031–2040.
17. Colapietro,P., Gervasini,C., Natacci,F., Rossi,L., Riva,P. and Larizza,L. (2003) NF1 exon 7 skipping and sequence alterations in exonic splice enhancers (ESEs) in a neurofibromatosis 1 patient. *Hum. Genet.*, **113**, 551–554.
18. Pagani,F., Buratti,E., Stuani,C. and Baralle,F.E. (2003) Missense, nonsense, and neutral mutations define juxtaposed regulatory elements of splicing in cystic fibrosis transmembrane regulator exon 9. *J. Biol. Chem.*, **278**, 26580–26588.
19. Pagani,F., Stuani,C., Tzetis,M., Kanavakis,E., Efthymiadou,A., Doudounakis,S., Casals,T. and Baralle,F.E. (2003) New type of disease causing mutations: the example of the composite exonic regulatory elements of splicing in CFTR exon 12. *Hum. Mol. Genet.*, **12**, 1111–1120.
20. Ferrari,S., Giliani,S., Insalaco,A., Al-Ghonaium,A., Soresina,A.R., Loubser,M., Avanzini,M.A., Marconi,M., Badolato,R., Ugazio,A.G. *et al.* (2001) Mutations of CD40 gene cause an autosomal recessive form of immunodeficiency with hyper IgM. *Proc. Natl Acad. Sci. USA*, **98**, 12614–12619.
21. Moseley,C., Mullis,P., Prince,M. and Phillips,J.,III (2002) An exon splice enhancer mutation causes autosomal dominant GH deficiency. *J. Clin. Endocrinol. Metab.*, **87**, 847–852.
22. Fackenthal,J.D., Cartegni,L., Krainer,A.R. and Olopade,O.I. (2002) BRCA2 T2722R is a deleterious allele that causes exon skipping. *Am. J. Hum. Genet.*, **71**, 625–631.
23. James,P.D., O'Brien,L.A., Hegadorn,C.A., Notley,C.R.P., Sinclair,G.D., Hough,C., Poon,M.-C. and Lillicrap,D. (2004) A novel type 2A von Willebrand factor mutation located at the last nucleotide of exon 26 (3538G>A) causes skipping of 2 nonadjacent exons. *Blood*, **104**, 2739–2745.
24. Mas,C., Taske,N., Deutsch,S., Guipponi,M., Thomas,P., Covanis,A., Friis,M., Kjeldsen,M.J., Pizzolato,G.P., Villemure *et al.* (2004) Association of the connexin 36 gene with juvenile myocolonic epilepsy. *J. Med. Genet.*, **41**, e93.
25. Aretz,S., Uhlhaas,S., Sun,Y., Pagenstecher,C., Mangold,E., Caspari,R., Moslein,G., Schulmann,K., Propping,P. and Friedl,W. (2004) Familial adenomatous polyposis: aberrant splicing due to missense or silent mutations in the APC gene. *Hum. Mutat.*, **24**, 370–380.
26. Zatkova,A., Messiaen,L., Vandenbroucke,I., Wieser,R., Fonatsch,C., Krainer,A.R. and Wimmer,K. (2004) Disruption of exonic splicing enhancer elements is the principal cause of exon skipping associated with seven nonsense or missense alleles of NF1. *Hum. Mutat.*, **24**, 491–501.
27. 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.
28. Tian,H. and Kole,R. (1995) Selection of novel exon recognition elements from a pool of random sequences. *Mol. Cell. Biol.*, **15**, 6291–6298.
29. Boukis,L.A. and Bruzik,J.P. (2001) Functional selection of splicing enhancers that stimulate trans-splicing *in vitro*. *RNA*, **7**, 793–805.
30. Schaal,T.D. and Maniatis,T. (1999) Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences. *Mol. Cell. Biol.*, **19**, 1705–1719.
31. Liu,H.X., Zhang,M. and Krainer,A.R. (1998) Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev.*, **12**, 1998–2012.
32. Liu,H.X., Chew,S.L., Cartegni,L., Zhang,M.Q. and Krainer,A.R. (2000) Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. *Mol. Cell. Biol.*, **20**, 1063–1071.

33. 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.
34. Gabut,M., Mine,M., Marsac,C., Brivet,M., Tazi,J. and Soret,J. (2005) The SR protein SC35 is responsible for aberrant splicing of the E1 alpha pyruvate dehydrogenase mRNA in a case of mental retardation with lactic acidosis. *Mol. Cell. Biol*., **25**, 3286–3294.
35. 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.
36. Fairbrother,W.G., Yeo,G.W., Yeh,R., Goldstein,P., Mawson,M., Sharp,P.A. and Burge,C.B. (2004) *Nucleic Acids Res*., **32**, W187–W190.
37. Zhang,X.H. and Chasin,L.A. (2004) Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev*., **18**, 1241–1250.
38. Kasprzyk,A., Keefe,D., Smedley,D., London,D., Spooner,W., Melsopp,C., Hammond,M., Rocca-Serra,P., Cox,T. and Birney,E. (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res*., **14**, 160–169.
39. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al*. (2004) Ensembl 2005. *Nucleic Acids Res*., **33**, D447–D453.
40. Zhang,M.Q. (1997) Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl Acad. Sci. USA*, **94**, 565–568.
41. Schaal,T.D. and Maniatis,T. (1999) Multiple distinct splicing enhancers in the protein-coding sequences of a constitutively spliced pre-mRNA. *Mol. Cell. Biol*., **19**, 261–273.
42. Mayeda,A., Screaton,G.R., Chandler,S.D., Fu,X.D. and Krainer,A.R. (1999) Substrate specificities of SR proteins in constitutive splicing are determined by their RNA recognition motifs and composite pre-mRNA exonic elements. *Mol. Cell. Biol*., **19**, 1853–1863.
43. Lam,B.J. and Hertel,K.J. (2002) A general role for splicing enhancers in exon definition. *RNA*, **8**, 1233–1241.
44. Ladd,A.N. and Cooper,T.A. (2002) Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol*., **3**, reviews 0008.
45. Itoh,H., Washio,T. and Tomita,M. (2004) Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes. *RNA*, **10**, 1005–1018.
46. Black,D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem*., **72**, 291–336.
47. 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.
48. Selvakumar,M. and Helfman,D.M. (1999) Exonic splicing enhancers contribute to the use of both 3′ and 5′ splice site usage of rat beta-tropomyosin pre-mRNA. *RNA*, **5**, 378–394.
49. Bourgeois,C.F., Popielarz,M., Hildwein,G. and Stévenin,J. (1999) Identification of a bidirectional splicing enhancer: differential involvement of SR proteins in 5′ or 3′ splice site activation. *Mol. Cell. Biol*., **19**, 7347–7356.
50. Tian,M. and Maniatis,T. (1994) A splicing enhancer exhibits both constitutive and regulated activities. *Genes Dev*., **8**, 1703–1712.
51. Lorson,C.L. and Androphy,E.J. (2000) An exonic enhancer is required for inclusion of an essential exon in the SMA-determining gene SMN. *Hum. Mol. Genet*., **9**, 259–265.
52. Hertel K.J. and Maniatis T. (1998), The function of multisite splicing enhancers. *Mol. Cell*, **1**, 449–455.
53. 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.
54. 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.
55. Smith,C.W. and Valcárcel,J. (2000) Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci*., **25**, 381–388.
56. Stamm,S., Zhu,J., Nakai,K., Stoilov,P., Stoss,O. and Zhang,M.Q. (2000) An alternative-exon database and its statistical analysis. *DNA Cell Biol*., **19**, 739–756.
57. Sorek,R. and Ast,G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res*., **13**, 1631–1637.
58. Philipps,D.L., Park,J.W. and Graveley,B.R. (2004) A computational and experimental approach toward a priori identification of alternatively spliced exons. *RNA*, **10**, 1838–1844.