Human Molecular Genetics, 2006, Vol. 15, No. 16 **2490–2508**
doi:10.1093/hmg/ddl171
Advance Access published on July 6, 2006

An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers

Philip J. Smith[1], Chaolin Zhang[1,2], Jinhua Wang[3], Shern L. Chew[4], Michael Q. Zhang[1] and Adrian R. Krainer[1,*]

[1]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA, [2]Department of Biochemical Engineering, State University of New York at Stony Brook, NY 11794, USA, [3]St Jude Children's Research Hospital, Memphis, TN 38105, USA and [4]Department of Endocrinology, St Bartholomew's Hospital, Queen Mary, University of London, London EC1M 6BQ, UK

Received May 16, 2006; Revised and Accepted July 4, 2006

Numerous disease-associated point mutations exert their effects by disrupting the activity of exonic splicing enhancers (ESEs). We previously derived position weight matrices to predict putative ESEs specific for four human SR proteins. The score matrices are part of ESEfinder, an online resource to identify ESEs in query sequences. We have now carried out a refined functional SELEX screen for motifs that can act as ESEs in response to the human SR protein SF2/ASF. The test *BRCA1* exon under selection was internal, rather than the 3′-terminal *IGHM* exon used in our earlier studies. A naturally occurring heptameric ESE in *BRCA1* exon 18 was replaced with two libraries of random sequences, one seven nucleotides in length, the other 14. Following three rounds of selection for *in vitro* splicing via internal exon inclusion, new consensus motifs and score matrices were derived. Many winner sequences were demonstrated to be functional ESEs in S100-extract-complementation assays with recombinant SF2/ASF. Motif-score threshold values were derived from both experimental and statistical analyses. Motif scores were shown to correlate with levels of exon inclusion, both *in vitro* and *in vivo*. Our results confirm and extend our earlier data, as many of the same motifs are recognized as ESEs by both the original and our new score matrix, despite the different context used for selection. Finally, we have derived an increased specificity score matrix that incorporates information from both of our SF2/ASF-specific matrices and that accurately predicts the exon-skipping phenotypes of deleterious point mutations.

INTRODUCTION

Most eukaryotic transcripts comprise multiple short exons interrupted by much longer introns. Mature mRNA is generated through splicing of the pre-mRNA in a well-characterized process that involves exon recognition via intronic sequences at the 5′ and 3′ splice sites and branch site, removal of the intron and ligation of the exons (1). The information content provided by the splice sites is insufficient for accurate exon definition (2). Additional information is provided by *cis*-acting regulatory sequences that serve to enhance or repress splicing, and that may be exonic or intronic in nature (3).

Exonic splicing enhancers (ESEs) were first identified as regulators of alternative splicing (4) and later demonstrated to participate in the splicing of constitutive exons (5,6). Members of the SR protein family function both as general splicing factors and as regulators of alternative splicing, and they act in part through recognition of, and binding to, ESEs (4,7). SR proteins share a conserved domain structure consisting of one or two copies of an RNA-recognition motif (RRM) followed by a C-terminal domain highly enriched in argine/serine dipeptides (RS domain) (8). The RRMs provide substrate specificity via sequence-specific RNA binding, whereas the RS domain is involved mainly in protein–protein interactions, with recent data suggesting an additional role in protein–RNA interactions (9,10). ESE-bound SR proteins function in exon definition by recruitment of the splicing machinery through their RS domains (11,12), and/or by

*To whom correspondence should be addressed. Tel: +1 5163678417; Fax: +1 5163678453; Email: krainer@cshl.edu

antagonizing the action of nearby splicing silencer elements (13–15). In addition to their role in splicing, recent work points to a role for SR proteins in numerous steps of RNA metabolism. These include nuclear export, mRNA quality control and translation (reviewed in 16). Furthermore, a recent study provided evidence for the involvement of ESE-bound SR proteins in ensuring the correct linear order of exons in mature mRNA (17).

Many studies have demonstrated that a diverse array of sequences can function as ESEs. An important tool for ESE identification has been the employment of functional systematic evolution of ligands by exponential enrichment (SELEX) both *in vivo* (18) and *in vitro* (19–21). A refinement of functional SELEX experiments utilized the ability of recombinant SR proteins to complement SR protein-deficient S100 extracts, and thus derived SR-protein-specific sequence motifs (22,23). These consensus motifs are short (6–8 nt), degenerate, and were found to be frequently inactivated by disease-associated mutations that cause exon-skipping *in vivo* and *in vitro* (24,25). Nucleotide-frequency scoring matrices derived from the SELEX consensus motifs are available in a web-based program called ESEfinder (26) (http://rulai.cshl.edu/tools/ESE/). Use of ESEfinder has demonstrated that numerous disease-associated mutations that cause exon-skipping correlate with a reduction in high-score ESE motifs to below-threshold values (24,25,27–32).

The Burge and Chasin labs have employed *ab initio* computational approaches to identify ESE motifs (33,34). RESCUE-ESE identified hexamer motifs that were over-represented in exons with weak splice sites, and also in exons versus introns (33). Zhang and Chasin utilized a similar approach to identify octamers over-represented in internal non-coding exons versus unspliced pseudo exons and the 5′ untranslated regions of intronless genes (34). Both methodologies yielded motifs that functioned as ESEs, with Zhang and Chasin also identifying a number of exonic-splicing silencers.

Definitive classification of any given genomic variation as deleterious remains problematic (reviewed in 35) and it is likely that the current versions of ESE-prediction methodologies described earlier require further refinement. Our original SR-protein-specific functional SELEX was carried out in the context of an *IGHM* (IgM) two-exon minigene pre-mRNA, and involved the replacement of a naturally occurring segment with ESE activity (36) in the 3′-terminal M2 exon with a random 20 nt library (22,23). The natural enhancer, and by extension the sequences under selection, are adjacent to an exonic-splicing silencer, recently demonstrated to function through the binding of the splicing repressor polypyrimidine-tract-binding protein (PTB) (15). The original functional SELEX experiments were limited by the inability to accommodate all possible 20mers within an *in vitro* splicing reaction. In this study, we have performed the SF2/ASF functional SELEX experiments in a different exonic context, namely a *BRCA1* three-exon minigene pre-mRNA (24), to examine the effects of sequence context on motif derivation and functionality. The position of the enhancer within the middle exon makes this test system typical of the majority of naturally occurring ESEs. Furthermore, the shorter lengths of random sequence inserted made it possible to achieve total representation of all possible sequences available for selection. We demonstrate that the new SF2/ASF-specific motifs we identified are functional ESEs, both *in vitro* and *in vivo*, with motif scores correlating with levels of exon inclusion. Moreover, a reduction in SF2/ASF motif score correlates with the exon-skipping phenotypes associated with a number of disease-causing mutations.

RESULTS

Identification of ESE motifs recognized by SF2/ASF under splicing conditions

Our previous functional SELEX experiments utilized libraries of random sequence 20 nt in length, in the context of a two-exon IgM minigene construct (22,23). The derived SR-protein-specific matrices have proven to be very useful tools for ESE prediction. We wanted to refine the original SF2/ASF score matrix for a number of reasons: (1) to investigate the effects of sequence context on motif selection; (2) to ascertain whether there are differences in the motifs selected from internal rather than terminal exons; (3) to derive a more quantitative measurement of the threshold value, i.e. the score above which we consider a given motif to be significant and therefore a probable ESE; (4) to improve the robustness of the ESE predictions. We chose to introduce random libraries of 7 and 14 nt into the three-exon *BRCA1* minigene, because the motifs recognized by SF2/ASF are 7mers (22). When we used the original 20mer library we did not know *a priori* the length of the recognition motif; the new libraries make it possible to achieve complete representation in the context of an *in vitro* splicing reaction, and they also avoid or reduce the complication of having multiple ESEs within each insert. The n7 library provides a fixed context, whereas the n14 library allows the influence of potential position effects and flanking-sequence upon motif functionality to be directly assessed.

In vitro splicing conditions were optimized using previously described minigenes containing either the wild-type (WT) *BRCA1* SF2/ASF-specific exon 18 ESE or the naturally occurring E1694X nonsense mutation (MT) (24), in both nuclear extract and S100 extract complemented by recombinant SF2/ASF. In addition, *in vitro* splicing experiments were performed in which the WT SF2/ASF-specific ESE was replaced with individual 14-nt sequences, such as two tandem WT or MT heptamers, or a single WT or MT heptamer at a central position within the 14 nt sequence. The expected pattern of exon inclusion was observed in experiments carried out both in nuclear extract and in S100 extract complemented by recombinant SF2/ASF, namely predominant exon inclusion with constructs containing the WT ESE, and exon-skipping with MT-containing constructs (data not shown).

Two SELEX libraries, one with n7, and one with n14, randomized regions in exon 18 of *BRCA1* in place of the SF2/ASF ESE heptamer, were constructed by overlap-extension PCR. Randomized regions of 7 and 14 nt allow complete sequence representation in standard *in vitro* splicing reactions. The position of the ESE (beginning at nucleotide +4 in exon 18) required the engineering of a *Bae*I restriction site within exon 17 for rebuilding the full-length minigene construct after each round of SELEX. The engineered construct was

tested by *in vitro* splicing and behaved in an identical manner to the parental construct (data not shown). An additional construct, utilized solely for cloning purposes (*BRCA1* C), containing the reciprocal *Bae*I site in exon 18 was generated to allow rebuilding of the full-length construct. The SELEX libraries were cloned into pCR-Blunt (Invitrogen), and random clones sequenced. Sequence analysis revealed no strong bias towards any of the four nucleotides, and therefore the starting pools were optimal for the SELEX experiments. The nucleotide composition for the n14 library was 25.8% A, 19.9% C, 27.2% G, 27.1% T (73 clones sequenced), and 28.1% A, 18.9% C, 28.1% G, 24.9% T (55 clones sequenced) for the n7 library. About 6×10^9 molecules of pre-mRNA were used as the input for the SELEX experiments. A conservative estimate of coverage of all possible 7mers and 14mers in the n7 and n14 libraries, respectively, can be made based upon the limiting nucleotide in the random pools. For both libraries, the limiting nucleotide is C (18.9% in the n7 library, 19.9% in the n14 library). For the n14 library all possible 14mers should be represented at least once, with the exception of C14, which is expected to occur $6 \times 10^9 \times (0.199)^{14} = 0.92$ times. Coverage of all possible heptamers is complete. For example, C7, the most infrequent heptamer based upon the limiting nucleotide, is expected to occur $6 \times 10^9 \times (0.189)^7 = 5 \times 10^4$ times.

Figure 1 illustrates the functional SELEX procedure. The SELEX libraries were spliced in S100 extract complemented by recombinant SF2/ASF. As controls, equivalent samples from the SELEX libraries were spliced in nuclear extract, and *BRCA1* WT and MT minigenes were spliced under the same conditions. Spliced exon-18-containing mRNAs were recovered and rebuilt into full-length splicing constructs. Three rounds of selection were performed with both libraries. The winner pools from each round of selection were subcloned into pCR-Blunt (Invitrogen), and random clones sequenced. The results of splicing of the round-three winner pools are shown in Figure 2A. As previously reported, *BRCA1* WT predominantly includes exon 18 in both nuclear extract and in S100 extract complemented by SF2/ASF (Fig. 2A, lanes 1 and 3), whereas exon 18 is predominantly skipped when *BRCA1* MT is spliced, especially in the S100 + SF2/ASF sample (Fig. 2A, lane 6). Both the n7 and n14 round-three winner pools undergo splicing with almost complete inclusion of exon 18 (Fig. 2A, lanes 7–12). The experiment was performed in triplicate and the reaction products quantified by phosphorimage analysis. The data were normalized to the levels of splicing obtained with the *BRCA1* WT construct, and expressed as both normalized exon inclusion [included mRNA/(included mRNA + skipped mRNA)] (Fig. 2B), and normalized inclusive splicing [included mRNA/(included mRNA + skipped mRNA + pre-mRNA)] (Fig. 2C). Expression of the data in this way allows both the level of exon inclusion and the overall splicing efficiency afforded by the selected RNA pools to be compared. The SELEX winner pools included exon 18 to a higher degree than *BRCA1* WT, and showed a greatly enhanced level of splicing activity in response to SF2/ASF.

Sequencing of random clones from both winner pools demonstrated a striking degree of both positive and negative selection, reflected in the overall nucleotide composition (Table 1). The round-three winner pools had very similar GC contents of 70.96 and 69.64% for the n7 and n14 libraries, respectively, increasing from 47% GC in the unselected library pools. The increase in GC content is accounted for by a decrease in the A content and a dramatic decrease in T content. The IgM SF2/ASF SELEX-winner pool had a GC content of 62%; however, the GC content of the IgM starting pool was 58% (22), and therefore the increase observed for the *BRCA1* SELEX winners is more significant.

Consensus motifs and score matrices were derived from the n7 and n14 round-three winner sequences. An alignment step was not required for the n7 winners, due to the ESE position being fixed by the length of the random sequence insert (see below). The n14 round-three winners were aligned using three different motif-finding algorithms: Gibbs sampler (37), MEME (38) and DME (39). A fourth alignment was performed by scoring the n14 winners with the n7-derived matrix. The highest scoring 7mer from each winner was then used to generate the consensus motif. The matrix derived from alignment of the n14 winners scored with the n7-derived matrix was subsequently found to be the most accurate predictor of both *in vitro* and *in vivo* splicing when compared with the n14 matrices derived using the alignments generated by the motif-finding algorithms. Therefore, we have limited our discussion to the results obtained with this matrix, designated n14(n7). It should be noted that the motifs derived using MEME and Gibbs were somewhat similar, whereas the DME-derived motif resembled the motif derived from the n14(n7) matrix. When used to predict *in vitro* splicing activity, only the DME and n14(n7)-derived matrices resulted in statistically significant correlations.

The aligned winner sequences and consensus motifs are shown in Figure 3. The consensus motifs are relatively degenerate, as reported for the previously derived SR protein-specific motifs (22,23). The winner sequences were then used to derive score matrices according to the frequency of each nucleotide at each position of the consensus motif, with an adjustment to take into account the compositional bias of the initial random pools (22,23). A second n7 matrix was derived, in which the initial matrix was used to score the round-three winner sequences plus three exonic nucleotides upstream of the ESE and six nucleotides downstream (constant flanking regions). If this resulted in a higher score for a given winner, the flanking nucleotide(s) were included in a new alignment of that winner, instead of the original 7mer motif. This second matrix proved to be slightly less accurate at predicting experimental splicing, and therefore the original n7 matrix was retained. Clones with the *BRCA1* WT ESE sequence were found in both the n7 unselected pool and rounds two and three winner pools, and were not included in the sequence analysis as they probably represent PCR contamination.

The scores of the n7 round-three winners ranged from 0.344 to 4.337, with a mean score of 2.649. The 66 clones sequenced from the unselected (round-zero) pool had scores that range from −4.637 to 3.529, with a mean score of −0.952. Only three sequences in the round-zero pool had scores higher than the mean of the winner pool, whereas 14 sequences in the winner pool had scores higher than this mean, and all 27 winner clones had scores higher than the mean of the round-zero pool. The scores of the round-zero and winner pools

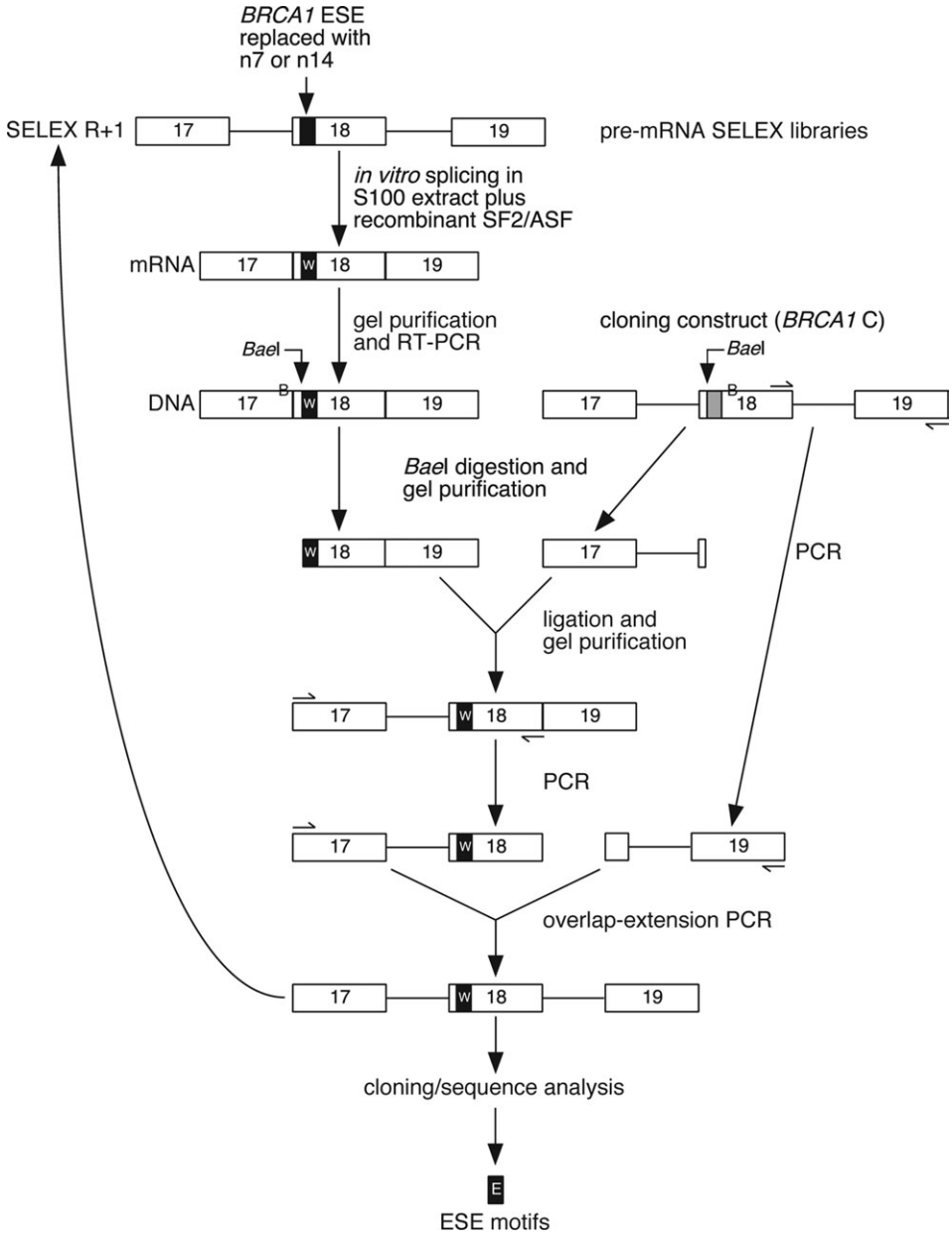

Figure 1. Experimental procedure for functional SELEX. The SF2/ASF-specific *BRCA1* exon 18 ESE was replaced by 7 or 14 nt of randomized sequence by overlap-extension PCR. *In vitro*-transcribed RNA was incubated under splicing conditions in HeLa S100 extract complemented by recombinant SF2/ASF. Spliced mRNA molecules containing SF2/ASF-responsive sequences (designated by the white W in a black box) were purified from denaturing polyacrylamide gels, and rebuilt into full-length intron-containing constructs. Following three rounds of selection, individual ESE-containing clones (E) were sequenced and consensus motifs and score matrices derived.

were compared by plotting the receiver operating characteristic (ROC) curve (Supplementary Material, Fig. S1). This analysis is a measure of the sensitivity (number of high-scores in the winner pool) and specificity (number of high-scores in the random pool) of the score matrix at all possible threshold values, from the minimum to the maximum score from both pools. Discrimination of the pools was observed for all threshold values (Supplementary Material, Fig. S1).

The n14 round-three winner scores ranged from 1.259 to 6.372, with a mean score of 3.919. The 73 clones from the round-zero pool had a score range of −3.467 to 4.410, with a mean score of 0.876. Only two sequences in the round-zero pool had scores higher than the mean of the winner pool, whereas 18 sequences in the winner pool had scores higher than this mean, and all 33 winner clones had scores higher than the mean of the unselected pool. Plotting the ROC curve for the n14 library demonstrated discrimination of the winner pool from the round-zero pool for all threshold values (Supplementary Material, Fig. S1). There is a significant bias for the highest-score motif being present at the beginning of the selected 14mer sequences. 13 out of the 33 winners have the highest score motif beginning at position 1

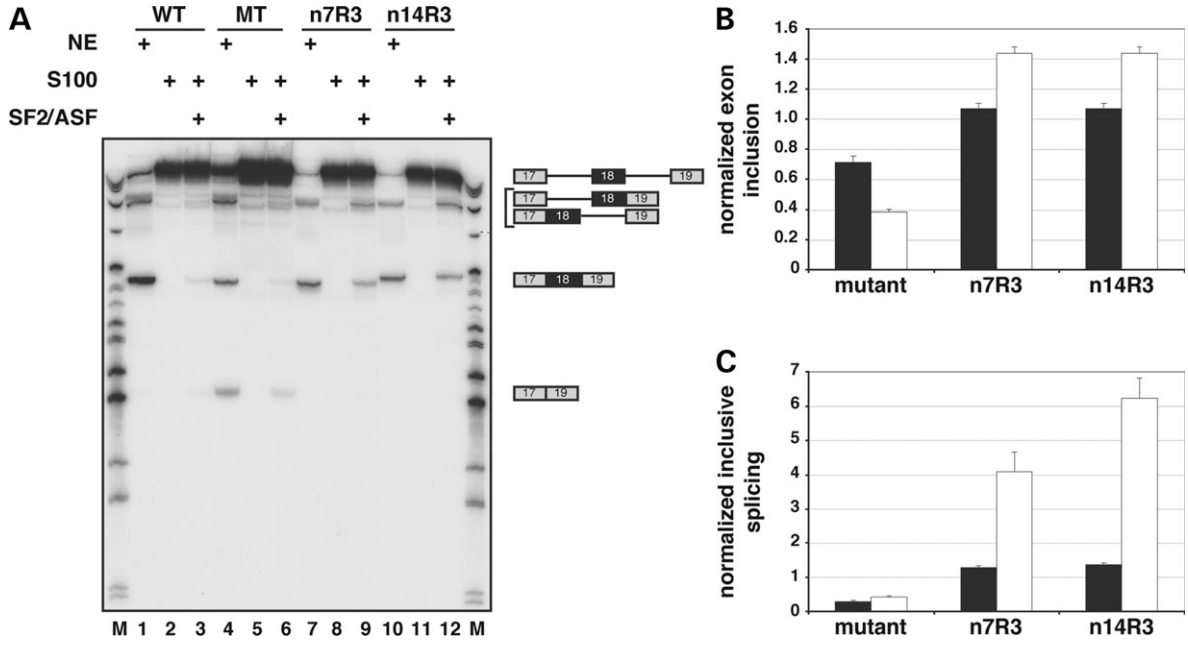

Figure 2. Splicing of the pre-mRNA pools following selection. (**A**) *In vitro* splicing of the n7 (lanes 7–9) and n14 (lanes 10–12) winner pools following three rounds of selection was performed in both HeLa nuclear extract (NE), and S100 extract complemented by recombinant SF2/ASF (A). As controls, *BRCA1* WT (lanes 1–3) and the nonsense E1694X MT (lanes 4–6) were also spliced. The structures of the precursor, intermediates and products are indicated next to the autoradiogram. The experiment was repeated three times and the data normalized to *BRCA1* WT levels of splicing for exon inclusion (**B**) and inclusive splicing (**C**). Black boxes indicate splicing in nuclear extract, white boxes splicing in S100 complemented by recombinant SF2/ASF and error bars equal the standard error.

Table 1. Nucleotide composition of the SELEX pools from the n7 and n14 libraries

Round	A	C	G	T	Number of clones
n7					
0	28.10	18.90	28.10	24.90	66
1	26.33	23.81	29.13	20.73	51
2	21.85	32.77	33.19	12.19	34
3	18.57	36.19	34.76	10.48	27
n14					
0	25.80	19.90	27.20	27.10	73
1	27.65	26.73	29.49	16.13	31
2	21.85	25.21	35.50	17.44	34
3	20.76	31.92	37.72	9.60	33

(26/33 at positions 1–4), indicating that there is selective pressure for functional ESEs being at this position, which corresponds to the position of the WT *BRCA1* ESE (Supplementary Material, Fig. S2).

The SELEX winners comprise functional ESEs

To investigate the functionality of the individual SELEX winner sequences, we tested the corresponding pre-mRNAs carrying individual sequences from the n7 and n14 round-three pools for their ability to include exon 18. The selected putative ESE motifs are in the same context as in the libraries used for functional SELEX, i.e. in place of the SF2/ASF-specific ESE of *BRCA1* exon 18. A representative experiment showing splicing of 10 n7 and 10 n14 winner sequences is shown in Figure 4A and B, respectively. The pre-mRNAs were incubated in both nuclear extract and S100 extract complemented by SF2/ASF. Under these splicing-reaction conditions, the WT *BRCA1* ESE is a weakly responsive SF2/ASF-specific ESE. As a control, we chose the level of enhancer function afforded by the *BRCA1* WT ESE as the lower limit for ESE function, and normalized all of the data to this level. Significantly, the winner sequences demonstrated a clear enhancement of splicing in response to SF2/ASF in comparison with the *BRCA1* WT construct (Fig. 4A and B), with many of the clones tested resulting in almost complete, and in a number of examples complete, exon inclusion.

Sixty-seven n7 and n14 winner sequences were spliced, in triplicate experiments, in S100 extract complemented by recombinant SF2/ASF, and the products of the splicing reactions quantified by phosphorimage analysis. The results were normalized to the *BRCA1* WT control run in each experiment, and expressed as normalized inclusive splicing [included mRNA/(included mRNA + skipped mRNA + pre-mRNA)], Fig. 4C to represent the degree of enhancer activity afforded by each of the tested motifs. For comparison, we also replotted the data as normalized exon inclusion [included mRNA/(included mRNA + skipped mRNA)] (Supplementary Material, Fig. S3). Sixty-six of the sequences supported levels of exon inclusion and inclusive splicing levels greater than that of the *BRCA1* WT control. The mean inclusive splicing level for the pooled data was 5.19 ± 3.17 for the n7 and n14 winners. Significantly, the single winner sequence that spliced poorly was found to contain a below-threshold (see below) motif when scored with the n7-derived matrix and the n14-derived matrix: 0.466 and 0.345, respectively.

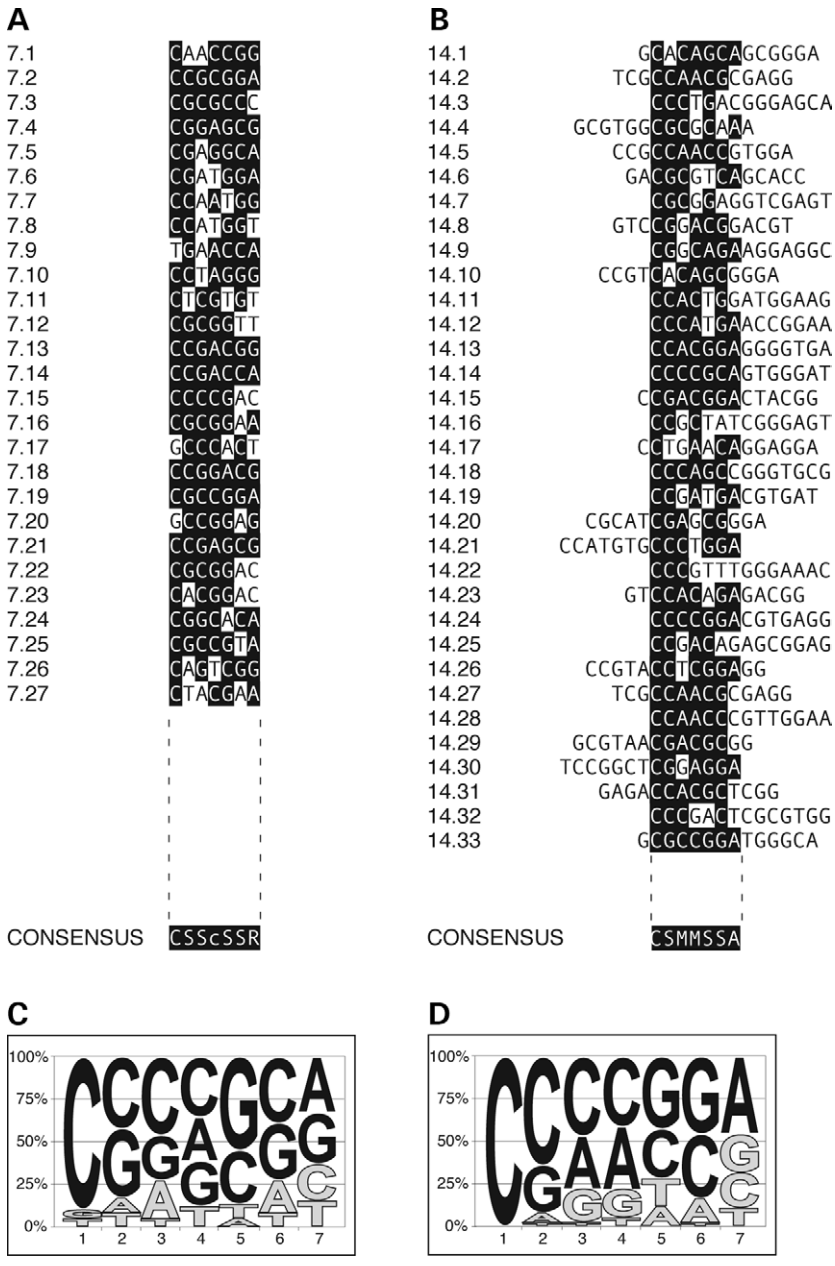

Figure 3. Analysis of the SF2/ASF-selected sequences. Sequence alignment and identification of consensus motifs from the n7 (**A**) and n14 round-three winner sequences (**B**). The n14 sequences were aligned on the basis of the highest score motif for each sequence. Nucleotides matching the consensus are shown as white on a black background, mismatched nucleotides are not shaded. The consensus shown is only an approximation that indicates the most frequent nucleotide(s) at each position. The lower case c at position 4 of the n7 consensus denotes a slight preference for this nucleotide over A and G, which occur at similar frequencies. R, purine; S, C or G; M, A or C. Pictogram representations (62) of the n7-derived consensus motif (**C**) and n14-derived consensus motif (**D**). The diagrams show the frequency of each nucleotide at each position of the heptamer consensus, adjusted for the compositional bias of the initial pool. The height of each letter is proportional to its frequency; black and grey letters indicate higher and lower than background frequencies, respectively. T is used instead of U for convenience.

In contrast, when a total of 45 random pre-mRNAs from the n7 and n14 unselected round-zero pools were tested, the levels of splicing observed were much lower. A representative experiment showing splicing of 10 n14 round-zero sequences spliced in S100 extract complemented by recombinant SF2/ASF is shown in Figure 5A. A number of sequences resulted in inclusion of exon 18 at levels above the *BRCA1* WT control (e.g. Fig. 5A, clones 9 and 10). Significantly, many of the sequences that spliced well were subsequently found to contain high-score motifs. The experiments were performed in triplicate, and the mean inclusive splicing levels for each sequence are shown in Figure 5B. The data in Figures 4C and 5B are plotted with the same ordinate scale to allow easier comparison. The mean inclusive splicing level for the n7 and n14 round-zero sequences was 1.54 ± 1.21, significantly lower than the splicing observed with the round-three winner sequences ($P < 10^{-10}$, two-sample *t*-test).

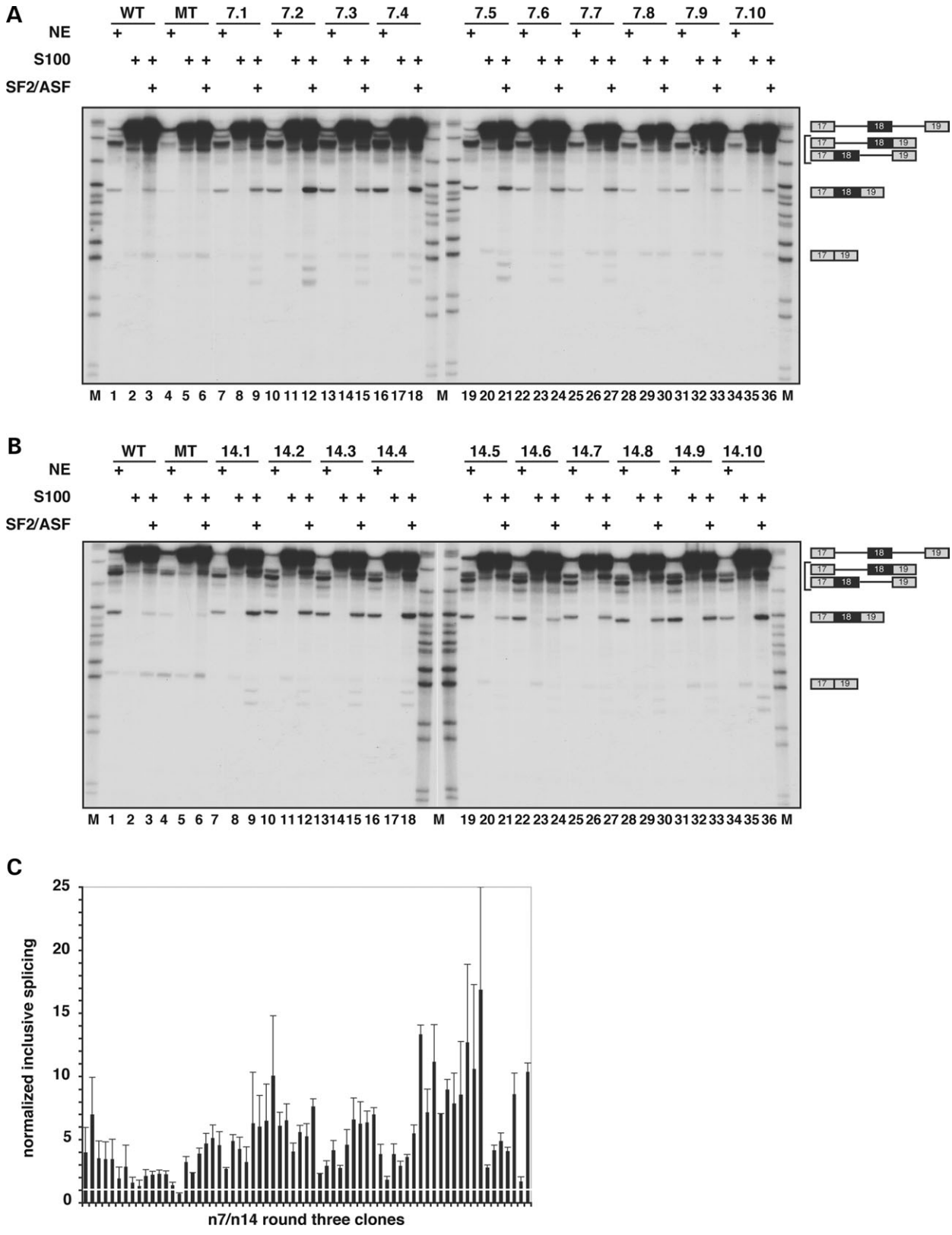

Figure 4. The winner sequences are functional SF2/ASF-dependent ESEs. *In vitro* splicing of 10 n7 (**A**) and 10 n14 (**B**) round-three winner sequences in HeLa nuclear extract, S100 extract alone and S100 complemented by recombinant SF2/ASF. The structures of the precursor, intermediates and products are indicated next to the autoradiograms. (**C**) Phosphorimage analysis of *in vitro* splicing, in S100 extract complemented by recombinant SF2/ASF, of 67 n7 and n14 round-three winner sequences. The experiment was performed three times and the data expressed as normalized inclusive splicing [included mRNA/(included mRNA + skipped mRNA + pre-mRNA)]. The horizontal white line represents the *BRCA1* WT level of inclusive splicing (1), error bars the standard error.

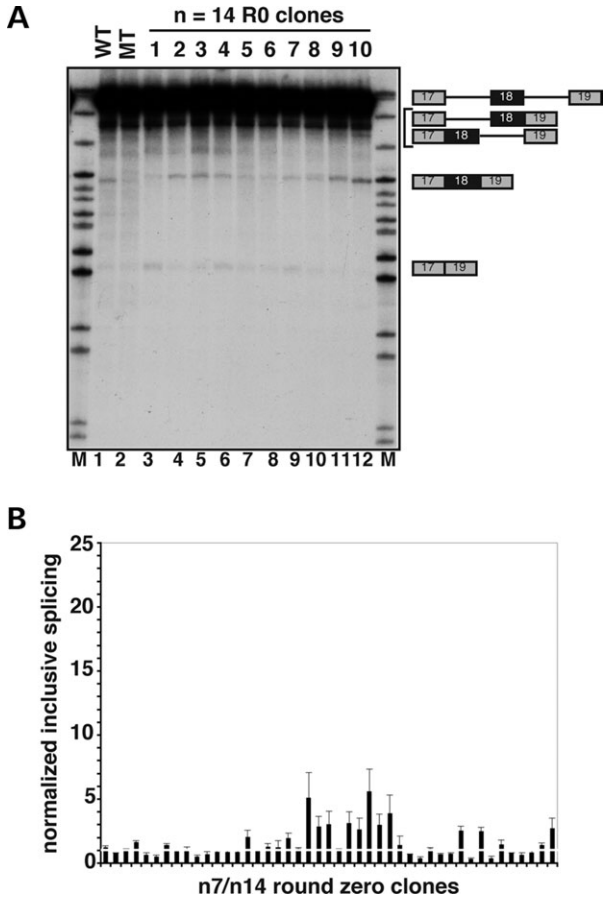

Figure 5. Unselected (round-zero) sequences contain a limited number of SF2/ASF-dependent ESEs. (**A**) *In vitro* splicing of 10 n14 round-zero clones in S100 complemented by recombinant SF2/ASF. The structures of the precursor, intermediates and products are indicated next to the autoradiogram. (**B**) Phosphorimage analysis of *in vitro* splicing, in S100 complemented by recombinant SF2/ASF, of 45 n7 and n14 round-zero clones. The experiment was performed three times and the data expressed as normalized inclusive splicing [included mRNA/(included mRNA + skipped mRNA + pre-mRNA)]. The horizontal white line represents the *BRCA1* WT level of inclusive splicing (1), error bars the standard error.

The SELEX winners function in a heterologous exonic context

To investigate whether the ESEs identified from our SELEX experiments are *BRCA1*-specific, or if they are able to function in a different exonic context, we transferred a number of winner sequences into the *SMN1* minigene. The *SMN1* construct is a three-exon minigene containing an SF2/ASF-specific ESE at position +6 in the middle exon, exon 7 (25,40). As a negative control for enhancer function, we used the *SMN2* minigene. *SMN2*, a paralog of *SMN1*, is identical to *SMN1* except for a few nucleotide differences in non-coding regions and a C to T transition at position +6 of exon 7 (41). The mutation inactivates the ESE and this correlates with a reduction in the ESEfinder SF2/ASF score to below-threshold (25,40). Homozygous loss or mutation of the *SMN1* gene causes spinal muscular atrophy (42). *SMN2* is unable to fully compensate for the lack of functional SMN1 protein as the predominant skipping of exon 7 results in an unstable protein (SMNΔ7) (43). Three round-three winner sequences from the n7 library and two round-zero sequences from the starting pool were inserted in place of the *SMN1* exon 7 ESE by overlap-extension PCR. The *SMN* clones were incubated in nuclear extract (Fig. 6), with *SMN1* and *SMN2* serving as positive and negative controls, respectively. All of the winner sequences promoted levels of exon inclusion that were greater than those observed for *SMN1* (Fig. 6A, lanes 5–7 compared with lane 1). In contrast, the two round-zero clones were similar to *SMN2*, resulting in predominant skipping of exon 7 (Fig. 6A, lanes 3 and 4). The experiment was performed in triplicate and the levels of exon inclusion for each construct normalized to those of *SMN1* (Fig. 6B).

Derivation of motif-score threshold values and matrix analysis

Motif scores are considered potentially significant if they are above a defined threshold value, in which case they are classified as high-score motifs. The default thresholds in ESEfinder were defined as the median of the highest scores for each sequence in a set of 30 randomly chosen 20-nt sequences from the starting pool used for the previous functional SELEX experiments (26). This statistical value gives a good differentiation between the starting pool and the SELEX winners in terms of scores, but does not take into consideration the splicing activity associated with different motif scores. We chose to derive a threshold value based on the experimentally determined extent of splicing of individual *BRCA1* SELEX clones. The n7- and n14-derived matrices were used to score the 112 clones that had been spliced *in vitro* in S100 extracts complemented with SF2/ASF. The amount of exon inclusion afforded by the WT *BRCA1* ESE was selected as the lower limit for positive enhancer activity, and all data were normalized to this level. The experimental threshold was defined as the minimum motif score that results in no false positives. In other words, this threshold means that all of the tested high-score clones give levels of exon inclusion at least equal to the WT ESE.

The n14 clones contain eight 7mers within each winner sequence. The highest scoring 7mer from each clone was counted as the putative ESE motif. Plotting *in vitro* splicing data against motif score (Fig. 7) allowed experimental threshold values of 1.121 and 1.748 to be derived for the n7 and n14 matrices, respectively. It should be noted that scores generated from different matrices are not numerically comparable, because of the different degree of matrix degeneracy. Using this stringent criterion for threshold derivation results in highly accurate segregation of low motif-score clones that do not enhance splicing (lower left quadrants, Fig. 7A and B), from high motif-score clones that result in above WT levels of exon inclusion (upper right quadrants). Using these threshold values, only nine (of the 112 tested) clones behaved as enhancers that were not predicted by the n7 matrix, and 19 in the case of the n14 matrix. For both the n7 and n14 matrices there is a highly significant correlation between motif score and observed enhancer activity (Table 2). Significantly, the n7 matrix scored the WT *BRCA1* ESE as above threshold (1.225) and the E1694X nonsense mutation as below threshold (−0.128). The n14 matrix correctly

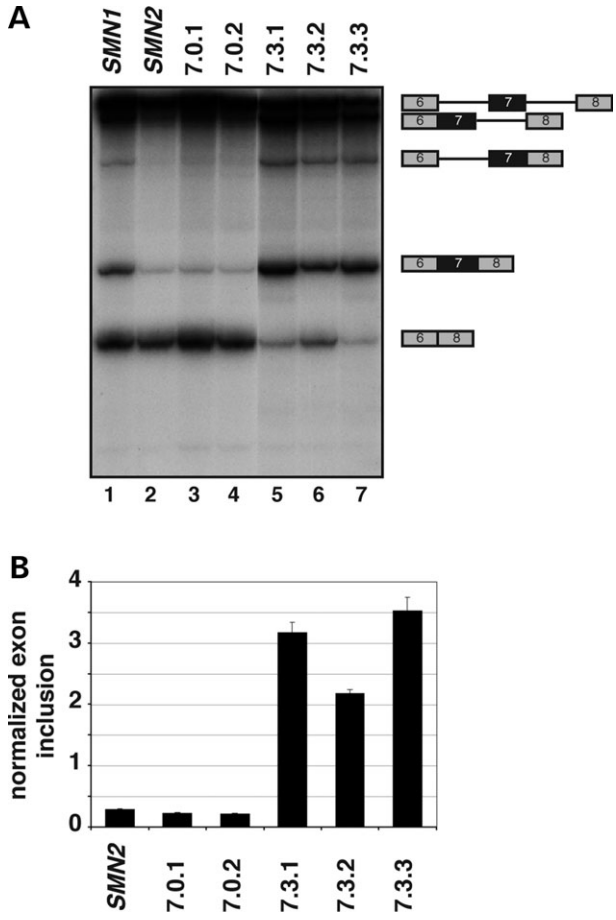

Figure 6. The SF2/ASF winner sequences function as ESEs in a heterologous context. (**A**) Three n7 round-three winners and two round-zero sequences were inserted in place of the naturally occurring SF2/ASF-responsive ESE in exon 7 of *SMN1*. The *SMN* constructs were spliced in nuclear extract with *SMN1* and *SMN2* serving as positive and negative controls, respectively. (**B**) Phosphorimage analysis of *in vitro* splicing. The splicing experiment was performed in triplicate and the data expressed as mean normalized exon inclusion relative to the *SMN1* control; error bars equal the standard error.

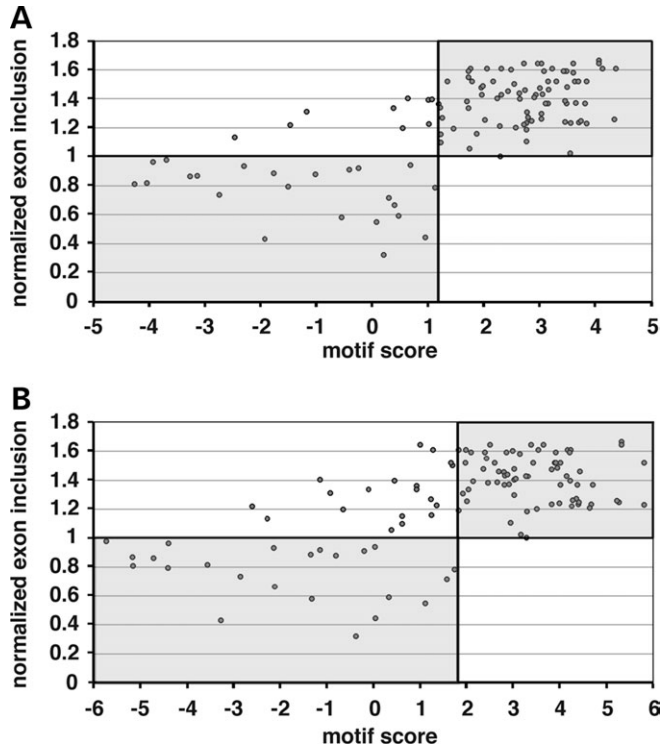

Figure 7. Correlation of SF2/ASF motif scores with *in vitro* splicing activity and experimental threshold derivation. The motif scores of 112 *BRCA1* clones were plotted against *in vitro* splicing activity for the n7 matrix (**A**) and the n14 matrix (**B**). Individual clones were spliced, in triplicate experiments, in S100 extract complemented with recombinant SF2/ASF. The horizontal line indicates the level of exon inclusion of the WT *BRCA1* ESE construct used as a control for each experiment. The vertical line represents the deduced experimental threshold corresponding to 1.121 and 1.748 for the n7 and n14 matrices, respectively. Quadrants containing clones whose splicing activity was correctly predicted are shaded grey.

ascertained a below-threshold score to E1694X, but did not score the WT ESE as above-threshold. These results suggest that the n7-derived matrix is the more robust predictor of ESE potential.

The matrices and experimentally derived thresholds were then tested against an independent panel of *SMN* constructs that were spliced *in vivo* (40). In that study, 30 *SMN* minigene constructs containing either the WT *SMN1* exon 7 SF2/ASF-dependent ESE, or mutations of the ESE motif, were transiently transfected into HEK293 cells and exon inclusion/skipping measured by semi-quantitative RT–PCR. We compiled the results into a dataset presented as scatter plots to demonstrate the relationship between exon 7-inclusion and matrix score for the n7 (Fig. 8A) and n14 (Fig. 8B) matrices, respectively. As was observed for the *in vitro BRCA1* splicing data, an accurate segregation of low-score clones that do not include exon 7 (lower left quadrants) from high-score clones that demonstrate high levels of exon inclusion (upper right quadrants) was achieved. The n7 matrix was again slightly more accurate in terms of predicting enhancer activity than the n14 matrix (23/30 correct predictions versus 21/30), and correctly predicted the splicing activity of both *SMN1* and *SMN2* (Fig. 8A). In addition, as mentioned earlier, the WT *BRCA1* ESE and the version with the E1694X nonsense mutation have high and low scores, respectively, with the n7 matrix. In this heterologous assay, in which these elements are tested in the context of an *SMN* minigene, the *BRCA1* WT ESE construct includes exon 7 at levels greater than *SMN2* (although below our 80% exon-inclusion threshold), whereas the *BRCA1* MT ESE results in complete exon-skipping. There is a significant correlation between motif scores and exon 7 inclusion (Table 2). The results are in accordance with those obtained with the SF2/ASF matrix in ESEfinder (40), and provide further evidence for exon 7 skipping in *SMN2* being a consequence of loss of a functional SF2/ASF ESE.

We repeated the matrix analysis utilizing other threshold values based on statistical discrimination of the winner pools from the unselected round-zero sequences. The n14 round-zero sequences were used because they represent a large number of random sequences and, as described earlier, exhibit no nucleotide-composition bias. We calculated a threshold value corresponding to the 60th percentile value of the highest scores from the 73 round-zero clones. This

Table 2. Correlation of ESE motif scores with *in vitro* and *in vivo* splicing

Matrix	n7[a]			n14[a]		
	Value	Sensitivity[d]	Specificity[e]	Value	Sensitivity[d]	Specificity[e]
Threshold E[b]	1.121	89.9	92	1.748	78.7	83.0
Threshold 60[c]	1.313	84.3	87.5	1.276	83.2	84.8
Statistical analysis[f]						
Number of XY pairs	112			112		
Pearson *r*	0.66			0.62		
95%CI	0.55–0.76			0.49–0.73		
P-value (two-tailed)	$P < 10^{-10}$			$P < 10^{-10}$		
R^2	0.44			0.39		
	n7[g]			n14[g]		
	Value	Correct predictions[h]		Value	Correct predictions[h]	
Threshold E	1.121	23		1.748	21	
Threshold 60	1.313	21		1.276	22	
Statistical analysis						
Number of XY pairs	30			30		
Pearson *r*	0.68			0.62		
95%CI	0.42–0.83			0.49–0.86		
P-value (two-tailed)	$P < 0.0001$			$P < 0.00001$		
R^2	0.46			0.53		

[a]The n7- and n14-derived matrices were used to score 112 *BRCA1* clones that were spliced *in vitro* in S100 extract plus recombinant SF2/ASF in triplicate. Motif scores were correlated against exon inclusion (normalized to *BRCA1* WT).
[b]Experimentally derived threshold.
[c]Threshold equal to the 60th percentile value of the highest score from each of the 73 sequenced n14 round-zero clones.
[d]Sensitivity is equal to the number of correctly predicted clones that splice/number of clones that splice.
[e]Specificity is equal to the number of correct predictions (clones that splice + clones that do not splice)/number of spliced clones.
[f]Statistical analysis correlating motif score to exon inclusion was performed using R software (GNU Project).
[g]The n7- and n14-derived matrices were used to score 30 *SMN* exon 7 ESE mutants analyzed for their ability to support exon inclusion *in vivo* by transfection into HEK293 cells (2–4 replicates) (40). Motif scores were correlated against % exon inclusion, with threshold values described above applied.
[h]The number of correct predictions (out of 30). Number of predictions by motif score that correlate with the observed splicing (exon inclusion or skipping).

threshold, when applied to the *BRCA1 in vitro* splicing data (Table 2) and *in vivo SMN* splicing experiments (Table 2) gave results that were in good accordance with the results obtained with the experimentally derived thresholds. It would appear that the ability of the matrices to predict the enhancer activity of a given motif is due less to the precise threshold value applied than to the discriminatory power of the matrices.

To investigate the degree of overlap between the n7 and n14 matrices, we scored all possible heptamers (16 384 sequences) with both matrices using the thresholds described earlier. The number of heptamers recognized as high scores by both matrices had a highly significant degree of overlap. For example, when the statistically derived thresholds were used, 81% of the heptamers recognized as high score by the n14 matrix were also high score with the n7 matrix, and 79% of the heptamers recognized as high score by the n7 matrix were high scores with the n14 matrix ($P < 10^{-10}$, Fisher's exact test, for the comparison of the expected number of overlapping high scores with the observed number of overlapping high scores).

An increased specificity score matrix for SF2/ASF

An important objective of this study was the derivation of an SF2/ASF score matrix with increased specificity. The current SF2/ASF matrix in ESEfinder generates a number of false positive high-scores, which may be at least in part a consequence of the context from which the motifs were derived. Increasing the specificity of the score matrix will enable more robust ESE prediction. A score matrix that incorporates information from both the original IgM SELEX experiments and the *BRCA1* SELEX experiments should allow recognition of ESE motifs capable of functioning in multiple exonic contexts.

A combined IgM-*BRCA1* SF2/ASF-specific score matrix was created. The *BRCA1* n7 matrix was chosen, as it gave the most accurate predictions of both *in vitro* and *in vivo* splicing activity when compared with the n14 matrix. We scored all possible heptamers (16 384) with both the n7 matrix and the ESEfinder SF2/ASF matrix. For increased specificity we used the original statistically derived IgM SF2/ASF matrix threshold (26), and the threshold corresponding to the 60th percentile of the *BRCA1* n14 R0 highest scores. Heptamer motifs that were recognized as high-scores by both matrices were used to generate the combined matrix. This analysis resulted in 395 overlapping high-score sequences, compared with the 678 IgM ESEfinder high-scores. Given the percentage of all possible heptamers that are high-score motifs with the n7 matrix, by chance one would expect only 91 overlapping sequences, so the observed degree of overlap is highly significant ($P < 10^{-10}$, Fisher's exact test). A consensus motif (Fig. 9A) and score matrix were derived from the overlapping high-scores, adjusted for the genomic exonic nucleotide

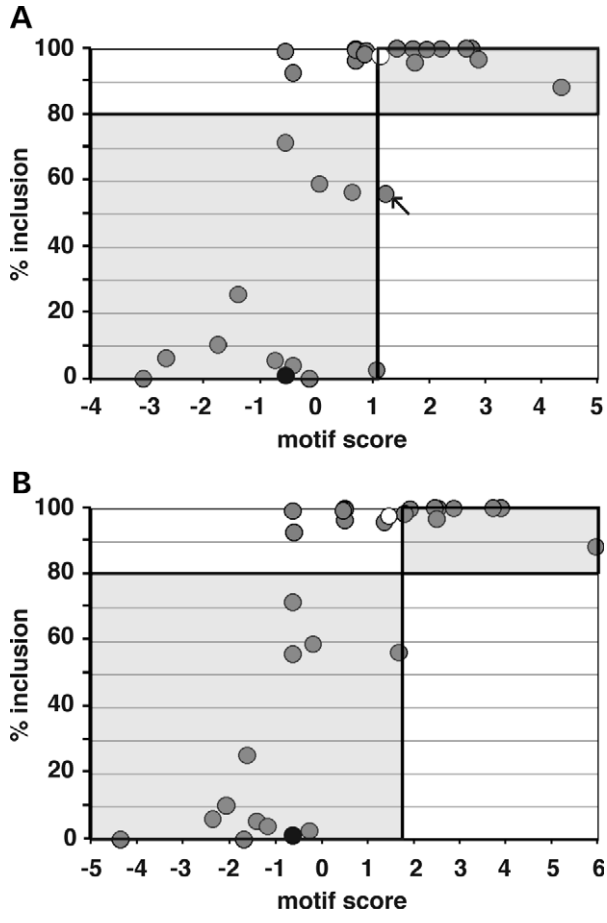

Figure 8. Correlation of SF2/ASF motif scores with *in vivo* splicing activity. The motif scores of 30 *SMN* exon 7 ESE mutants were plotted against exon 7 inclusion for the n7 matrix (**A**) and the n14 matrix (**B**). Splicing activity was analyzed by transiently transfecting plasmids harboring *SMN* minigenes into HEK293 cells and measuring the subsequent levels of exon 7 inclusion by semi-quantitative RT-PCR (2–4 repeats) (40). *SMN1* and *SMN2* are colored white and black, respectively. *BRCA1* exon 18 WT ESE is indicated by an arrow. The vertical line represents the deduced experimental thresholds. The horizontal line represents 80% exon inclusion (our lower limit for enhancer activity). Quadrants containing clones whose splicing activity was correctly predicted are shaded grey.

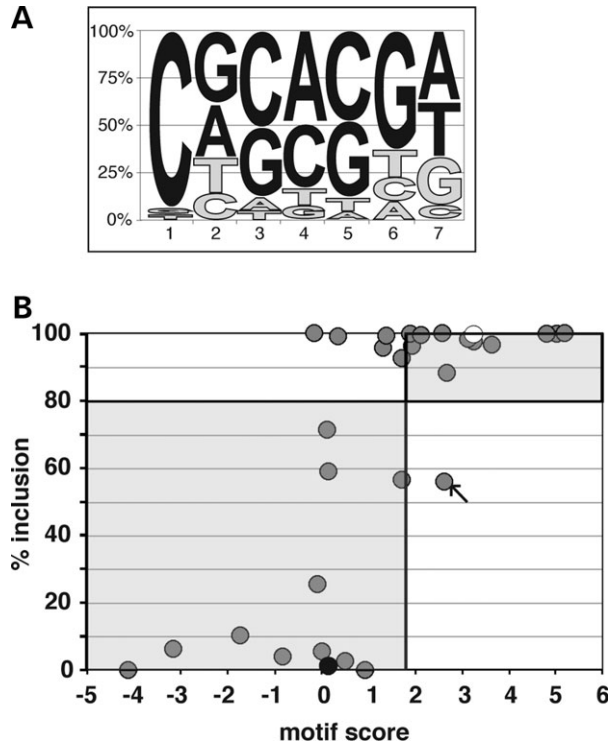

Figure 9. A combined IgM-*BRCA1* SF2/ASF score matrix. (**A**) Pictogram representation of the consensus motif for functional SF2/ASF-specific ESEs derived from the IgM and *BRCA1* SELEX experiments. The diagram shows the frequency of each nucleotide at each position of the heptamer consensus, adjusted for exonic nucleotide frequency. The height of each letter is proportional to its frequency; colored and grey letters indicate higher and lower than background frequencies, respectively. T is used instead of U for convenience. (**B**) Correlation of SF2/ASF motif scores with the *in vivo* splicing activity of 30 *SMN* exon 7 ESE mutants. Splicing activity was analyzed by transiently transfecting plasmids harboring *SMN* minigenes into HEK293 cells and measuring the subsequent levels of exon 7 inclusion by semi-quantitative RT-PCR (2–4 repeats) (40). *SMN1* and *SMN2* are colored white and black, respectively. *BRCA1* exon 18 WT ESE is indicated by an arrow. The vertical line represents the motif score threshold value of 1.867. The horizontal line represents 80% exon inclusion (our lower limit for enhancer activity). Quadrants containing clones whose splicing activity was correctly predicted are shaded grey.

frequency as the background (25.9691% A, 25.0262% C, 25.4594% G, 23.5453% T, derived from a large set of human internal protein-coding exons) (44). It is highly probable that both the IgM and the *BRCA1*-derived matrices predict a number of motifs that function only in the context from which they were derived. Therefore the threshold for the combined matrix was derived statistically, rather than using the experimental threshold based on *BRCA1 in vitro* splicing (1.867, equal to the 70th percentile of the highest score from the *BRCA1* R0 clones). About 57.5% of the combined matrix high-score motifs are recognized by the ESEfinder SF2/ASF matrix and 81.4% are recognized by the n7 matrix.

As an independent assay of the robustness of the new matrix, we scored the panel of *SMN* mutants analyzed in Figure 8, and correlated motif scores with the observed *in vivo* splicing activity. The results are shown as a scatter plot to demonstrate the relationship between *SMN* exon 7 inclusion and motif score (Fig. 9B). The combined matrix correctly predicts the ESE functionality of 25/30 of the motifs, including both *SMN1* and *SMN2*. In addition, *BRCA1* WT and the E1694X nonsense mutation have high and low scores, respectively. The correlation between motif score and exon 7 inclusion is highly significant (Pearson correlation coefficient of 0.72; 95%CI 0.48–0.86).

ESE inactivation by point mutations is a documented mechanism of disease (reviewed in reference 3). For example, the *BRCA1* nonsense mutation E1694X and the translationally silent C → T difference between *SMN1* and *SMN2* disrupt SF2/ASF-specific ESEs, causing skipping of the exons harboring these mutations and progression to breast cancer and SMA, respectively (24,25,40). We also analyzed a set of 73 exonic single-base substitutions in human genes that cause exon-skipping *in vivo* (3,45). This set was created from a list of 50 exonic mutations documented to cause

exon-skipping *in vivo* (45), and 42 exonic mutations associated with altered splicing *in vivo* (3). The list was filtered to remove mutations that cause increased exon inclusion, and mutations that fall in the last two nucleotides of an exon, overlapping the 5′ splice-site. We scored the WT and mutant sequences for the presence of above-threshold (>1.867) motifs using the IgM-*BRCA1* combined matrix. Table 3 summarizes the results obtained. For completeness we have included the results obtained with the score matrices previously derived for SRp40, SRp55 and SC35 (22,23,26). The motifs recognized by SF2/ASF and SRp40 are heptamers, SRp55 recognizes hexamers and SC35 octamers. Therefore, we scored 13mers for SF2/ASF and SRp40, 11mers for SRp55 and 15mers for SC35, with the mutant nucleotide at the central position. Twenty of the 73 WT sequences contained above-threshold SF2/ASF motifs that were reduced in the mutant sequences, and of these, 15 were reduced below the threshold. In contrast, only seven sequences had higher scores in the mutant than the WT sequence. To eliminate the background from sequences that lack putative SF2/ASF motifs in both WT and mutant versions, we compared only the sequences with a high-score motif in either the WT version, the mutant version, or both. The highest score motif in each of these sequences was compared with the highest score motif in the corresponding WT or mutant sequence. This analysis revealed a statistically significant difference in SF2/ASF motif scores between the WT and mutant sequences ($P < 0.01$, Fisher's exact test). The combined output of the four SR protein-specific matrices resulted in the prediction of 23 mutant sequences having loss of ESE function, compared with the creation of 12 putative ESEs in the mutants. The effects of mutations in putative SF2/ASF ESEs appear to be more detrimental than mutations occurring in putative ESEs for the other three SR proteins.

Two *ab initio* computational ESE prediction methods have recently been developed by the Burge and Chasin labs (33,34). We reported that the motifs recognized by ESEfinder do not overlap above the chance expectation with those predicted by RESCUE-ESE or the PESEs of Zhang and Chasin, with the exception of a significant overlap in the number of SF2/ASF high-score motifs in the set of PESEs (44). We hypothesized that the exon-skipping mutants predicted to be deleterious by our functionally derived matrices may be different from those recognized by the computational prediction methodologies, which presumably include ESEs recognized by a wide variety of proteins. We scored the 73 WT and mutant sequences for the presence of RESCUE-ESE motifs and PESEs or PESSs (PESX) (33,34). We scored 11mer sequences, with the mutant nucleotide at the central position, for the presence of RESCUE-ESE hexamers (http://genes.mit.edu/burgelab/rescue-ese/index.html), and 15mer sequences for the presence of PESX octamers (http://cubweb.biology.columbia.edu/pesx/). The Burge lab recently reported the results of an *in vivo* screen for motifs that function as ESSs (46). We therefore also scored the WT and mutant sequences (11mers) for the presence of motifs that match the putative ESS motifs (FAS-hex2 hexamers, http://genes.mit.edu/fas-ess/). Supplementary Material, Tables S1 and S2 summarize the results of this analysis. RESCUE-ESE, FAS-hex2 and PESX were able to predict the exon-skipping behavior of 25, 21 and 32 out of the 73 mutations, respectively. As for the ESEfinder analysis, predictions were taken to correlate with exon-skipping if there was an overall loss of putative ESEs, or gain of ESSs. Without experimental validation, it is not possible to conclude which of multiple motifs within any given sequence is functional. Overall, RESCUE-ESE predicted the loss of 25 ESEs and the creation of 15 new ESEs. FAS-hex2 predicted the creation of 21 new ESSs and the loss of only four ESS motifs in the mutant sequences. The predictions of exon-skipping with PESX were robust; there was a net loss of 20 PESE-containing sequences in the mutants, with the creation of six PESEs. The mutant sequences contain 15 new PESSs and lose only one. Significantly, most of the mutants whose ESE functionality was correctly predicted by RESCUE-ESE and PESE were different from those predicted by ESEfinder.

The combined output of scoring the set of exon-skipping mutants with all four methods is shown in Table 4. In total, the exon-skipping phenotype of the 73 point mutations is predicted correctly 61 times, a significant increase over the ability of any one of the methods alone. None of the mutants were predicted to skip by all three ESE predictors. It should be noted that a direct comparison of the three methods is not possible, as ESEfinder searches for specific motifs for only four of the SR proteins, whereas RESCUE-ESE and PESE are not protein-specific. Including more SR protein score matrices should increase the number of mutations amenable to this type of analysis.

SF2/ASF motif frequencies in human protein-coding exons

Functional ESEs should be present at a higher frequency in exons compared with their flanking introns, as we observed for the original SR-protein-specific matrices in ESEfinder (44) and as was found for the RESCUE-ESE hexamers and PESE octamers (33,34). We created a dataset of 16 635 constitutively spliced internal protein-coding exons ≥ 106 nt in length (100 consecutive heptamers), plus 100 nt each of flanking upstream and downstream intronic sequence. We created a second dataset of 5041 exons that undergo alternative splicing. 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. We scored the datasets with the IgM-*BRCA1* matrix for the presence of high-score motifs (above the threshold of 1.867).

SF2/ASF scores were calculated at each position and high-score motif frequencies plotted at the first position of the motif (Fig. 10). As expected, given the SF2/ASF consensus motif, there is a sharp decrease in the frequency of SF2/ASF motifs found in the areas corresponding to the polypyrimidine tract and 5′ splice-site. ESE motif frequencies were approximately constant within the exons, as we previously observed with the ESEfinder high-score motif frequency distributions within exons (44). Comparison of the ESE-motif frequencies revealed that the higher density of ESE motifs in exons than in introns was statistically significant for both the constitutive and alternatively spliced exons (paired *t*-test, $P < 10^{-10}$). Analysis of the datasets with ESEfinder SF2/ASF (26) also revealed a significant difference in ESE-motif frequency between exons and introns (data not shown), in accordance with our previous data (44).

Table 3. Disease-associated mutations that cause exon-skipping correlate with a reduction in SF2/ASF motif scores

Gene[a]	Mutation[b]	sub[c].	SF2/ASF	SRp40 SRp55 SC35
Missense mutations				
ADA	A215T	G → A	↓↑	40 ↑
ATM	E2032K	G → A		40 ↑35 ↑
ATP7A	G1302R	G → A		
BRCA1	E1694K	G → A	↓	
CFTR	G85E	G → A		40 ↓ 55 ↓
CFTR	D565G	A → G	↑	40 ↑ 35 ↓
F8	R1997W	C → T		
HEXB	P404L	C → T	↓	55 ↓
HPRT	G40V	G → T	↑	55 ↑
HPRT	R48H	G → A		55 ↓
HPRT	A161E	C → A	↓	40↓↑35 ↓
HPRT	G180E	G → A		
HPRT	G180V	G → T		
HPRT	G182K	G → A		
HPRT	P184L	C → T	↓	40 ↑35 ↓
HPRT	D194Y	G → T		40 ↑35 ↓
HPRT	E197K	G → A		55 ↓
HPRT	E197V	A → T		55 ↑
HPRT	D201V	A → T		
IVD	R21C	C → T		55 ↓
IVD	R21P	G → T		55 ↓
IVD	D40N	G → A		40 ↑
MLH1	R659P	G → C		40 ↑ 35 ↑
MLH1	R659L	G → T		40 ↑↑35↑
PDHA1	A175T	G → A	↓	40 ↑35 ↑
PMM2	E139K	G → A		
RHAG	G380V	G → T	↓	40 ↓
Silent mutations				
APC	R623R	G → T	↓	40 ↓
AR	S888S	C → T		55 ↑
CYP27A1	G112G	G → T		35 ↑
FBN1	I21181I	C → T		35 ↑
HPRT	F199F	C → T		40 ↓55 ↓
ITGB3	T420T	G → A	↓↑	40 ↓
NF1	K354K	G → A		
PAH	V399V	A → T		
PDHA1	G185G	A → G		
RET	I647I	C → T		40 ↑ 55 ↓
SMN1	F280F	C → T	↓	
TNFRSF5	T136T	A → T	↓	40 ↓ 35 ↑

Gene[a]	Mutation[b]	sub[c].	SF2/ASF	SRp40 SRp55 SC35
Nonsense mutations				
ADA	R142X	C → T		
ATP7A	R645X	C → T		40 ↓
BRCA1	E1694X	G → T	↓	55 ↑
CFTR	E60X	G → T		35 ↓
CFTR	R75X	C → T	↓↑	55 ↓
CFTR	R553X	C → T		35 ↑
CFTR	W1282X	G → A	↓↓	40 ↓
DMD	E1211X	G → T		
F8	E1987X	G → T	↓	40 ↓
F8	R2116X	C → T		55 ↑
FANCC	R185X	C → T	↓	40 ↓
FBN1	Y2113X	T → G	↑	40 ↓ 55 ↓
HMGCL	E37X	G → T		55 ↑
HPRT	E30X	G → T	↓	55 ↓↑
HPRT	E47X	G → T		40 ↑
HPRT	R51X	C → T		40 ↓
HPRT	K55X	A → T		
HPRT	C66X	T → A	↑	40 ↑55 ↓
HPRT	K103X	A → T	↓	55 ↑35 ↓
HPRT	G119X	G → T		
HPRT	L125X	T → G		40 ↓
HPRT	G180X	G → T		
HPRT	E182X	G → T		
HPRT	E197X	G → T		40 ↑55 ↓
HPRT	Y198X	C → G		40 ↓↑55 ↓
IDUA	Y64X	C → A		55 ↓
MLH1	R659X	C → T		40 ↑
NF1	Y2264X	C → A		40 ↓
NF1	Y2264X	C → G		40 ↓
OAT	W178X	G → A	↓	40 ↓ 35 ↓
OAT	W275X	G → A		
PROS1	S62X	C → G		35 ↑
TG	R717X	C → T		
WAS	Q99X	C → T	↓	

SR	↓[d]	↑[e]	*P*-value[f]
SF2/ASF	21	7	**0.007**
SRp40	18	17	1
SRp55	15	8	0.062
SC35	7	9	0.704
Total	61	41	

Seventy-three point mutations that cause exon-skipping and disease were scored with the combined IgM-*BRCA1* SF2/ASF score matrix and the score matrices for SRp40, SRp55 and SC35 from ESEfinder. Sequence motifs for the same or for a different SR protein can overlap. Only the WT or MT sequence motifs with scores greater than or equal to the threshold for the corresponding SR protein were considered. Downward arrows denote a reduction, or elimination (bold) of the motif score as a result of the mutation. Upward arrows denote a higher score in the MT than the WT, with bold denoting an increase from below to above the threshold (creation of new putative ESE motif).

[a]Genes and their encoded proteins are as follows: *ADA*, adenosine deaminase; *APC*, adenomatous polyposis coli; *AR*, androgen receptor; *ATM*, ataxia telangiectasia mutated; *ATP7A*, ATPase, Cu^{2+} transporting, α-polypeptide; *BRCA1*, breast cancer 1, early onset; *CFTR*, cystic fibrosis transmembrane conductance regulator; *CYP27A1*, sterol-27-hydroxlase; *DMD*, dystrophin; *F8*, coagulation factor VIII; *FANCC*, Fanconi anemia, complementation group C; *FBN1*, fibrillin 1; *HEXB*, hexosaminidase B, β-polypeptide; *HMGCL*, 3-hydroxymethyl-3-methylglutaryl-Coenzyme A lyase; *HPRT*, hypoxanthine phosphoribosyltransferase 1; *IDUA*, α-L-iduronidase; *ITGB3*, integrin-β3; *IVD*, isovaleryl coenzyme A dehydrogenase; *MLH1*, mutL homologue; *NF1*, neurofibromin 1; *OAT*, ornithine amino-transferase; *PAH*, phenylalanine hydroxylase; *PDHA1*, pyruvate dehydrogenase (lipoamide) α1; *PMM2*, phosphomannomutase 2; *PROS1*, protein S-α; *RHAG*, Rhesus blood group-associated glycoprotein; *SMN1*, survival of motor neuron 1; *TG*, thyroglobulin; *TNFRSF5*, tumor-necrosis factor receptor superfamily, member 5 (*CD40*); *WAS*, Wiskott–Aldrich syndrome.

[b]The specific mutations are identified by the WT amino acid in the one-letter code, followed by the residue number in the protein sequence and the MT amino acid (X denotes one of the three nonsense codons) as it would be in the absence of exon-skipping.

[c]Nucleotide substitution.

[d]The number of high-score motifs that are reduced in the MT sequences compared with the WT.

[e]The number of high-score motifs that are increased or created in the MT sequences compared with the WT.

[f]Comparison of the highest score in the WT sequence with the highest score in the MT when a high score is present in either the WT or MT sequence (Fisher's exact test).

Table 4. Prediction of the effects of exon-skipping disease-associated point mutations by scoring for the presence of ESEfinder, RESCUE-ESE, FAS-hex2 and PESX motifs

Gene	Mutation	Exon-skipping predicted				Gene	Mutation	Exon-skipping predicted			
		ESEfinder[a]	RESE[b]	FAS hex2[c]	PESX[d]			ESEfinder[a]	RESE[b]	FAS hex2[c]	PESX[d]
Missense mutations						Nonsense mutations					
ADA	A215T					*ADA*	R142X				
ATM	E2032K				Yes	*ATP7A*	R645X	Yes			Yes
ATP7A	G1302R		Yes			*BRCA1*	E1694X			Yes	
BRCA1	E1694K	Yes		Yes		*CFTR*	E60X	Yes			Yes
CFTR	G85E	Yes	Yes			*CFTR*	R75X				
CFTR	D565G		Yes		Yes	*CFTR*	R553X		Yes		Yes
F8	R1997W			Yes		*CFTR*	W1282X	Yes			
HEXB	P404L	Yes				*DMD*	E1211X		Yes		Yes
HPRT	G40V				Yes	*F8*	E1987X	Yes	Yes	Yes	Yes
HPRT	R48H	Yes				*F8*	R2116X				
HPRT	A161E					*FANCC*	R185X	Yes			
HPRT	G180E		Yes			*FBN1*	Y2113X		Yes	Yes	
HPRT	G180V		Yes		Yes	*HMGCL*	E37X		Yes	Yes	
HPRT	G182K				Yes	*HPRT*	E30X		Yes	Yes	Yes
HPRT	P184L			Yes	Yes	*HPRT*	E47X		Yes		Yes
HPRT	D194Y				Yes	*HPRT*	R51X	Yes			
HPRT	E197K	Yes	Yes			*HPRT*	K55X		Yes	Yes	Yes
HPRT	E197V		Yes		Yes	*HPRT*	C66X				
HPRT	D201V			Yes		*HPRT*	K103X		Yes		Yes
IVD	R21C	Yes			Yes	*HPRT*	G119X		Yes		
IVD	R21P	Yes				*HPRT*	L125X	Yes			
IVD	D40N					*HPRT*	G180X		Yes		Yes
MLH1	R659P			Yes		*HPRT*	E182X				Yes
MLH1	R659L			Yes		*HPRT*	E197X		Yes		Yes
PDHA1	A175T					*HPRT*	Y198X		Yes	Yes	
PMM2	E139K		Yes		Yes	*IDUA*	Y64X	Yes		Yes	
RHAG	G380V	Yes			Yes	*MLH1*	R659X				
Silent mutations						*NF1*	Y2264X	Yes			
APC	R623R	Yes				*NF1*	Y2264X	Yes			
AR	S888S			Yes		*OAT*	W178X	Yes		Yes	Yes
CYP27A1	G112G			Yes		*OAT*	W275X				Yes
FBN1	I21181I					*PROS1*	S62X				Yes
HPRT	F199F	Yes	Yes	Yes	Yes	*TG*	R717X				Yes
ITGB3	T420T					*WAS*	Q99X	Yes			Yes
NF1	K354K				Yes	Total		23	25	21	32
PAH	V399V		Yes	Yes	Yes						
PDHA1	G185G		Yes	Yes		Combined prediction[e]			61/73		
RET	I647I				Yes						
SMN1	F280F	Yes	Yes	Yes							
TNFRSF5	T136T										

[a]Seventy-three point mutations that cause exon-skipping and disease were scored with the combined IgM-*BRCA1* SF2/ASF score matrix and the score matrices for SRp40, SRp55 and SC35 from ESEfinder.
[b]Scored for the presence of RESCUE-ESE (RESE) hexamers.
[c]Scored for the presence of FAS-hex2 hexamers.
[d]Scored for the presence of PESX octamers.
[e]Total number of correct predictions when all four of the motif analysis methodologies are used.

DISCUSSION

The prototypical SR protein SF2/ASF (47,48) functions in both the regulation of alternative splicing and in constitutive splicing (6,49), and participates in other cellular processes, including nonsense-mediated mRNA decay, maintenance of genomic stability, RNA export and translation (50–53). We have performed a functional SELEX screen for sequence motifs that are recognized as ESEs by SF2/ASF. Our data support the validity of the SR protein-specific ESE matrix derived in the context of the IgM two-exon minigene (22,23), and provide further information regarding the motifs functionally recognized by SF2/ASF.

We carried out parallel SELEX experiments utilizing one random library of seven nucleotides, and one of 14 nucleotides, in place of the naturally occurring SF2/ASF-specific ESE in *BRCA1* exon 18 (24). *In vitro* splicing conditions in S100 extracts complemented by recombinant SF2/ASF were optimized to minimize the possibility of non-functional motifs passing through the ESE screen. Three rounds of functional SELEX were performed, resulting in the derivation of consensus motifs and score matrices from both libraries. There was a highly significant overlap in the motifs derived from the two libraries, with both winner pools being greatly enriched for C and G nucleotides. The consensus motifs are

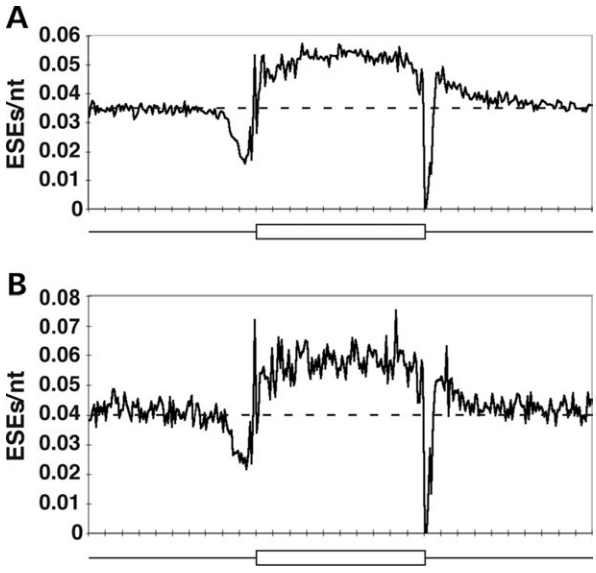

Figure 10. SF2/ASF motif frequency distribution in coding exons and flanking introns. The IgM-*BRCA1*-derived score matrix was used to analyze (**A**) 16 635 constitutive and (**B**) 5041 alternatively spliced coding exons ≥106 nt in length for the presence of high-score SF2/ASF motifs (≥1.867). The boxes represent composite exons 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, and high-scores plotted at the first position of the motif. The horizontal dashed lines represent the mean intronic motif densities of 0.0354 ESEs/nt and 0.0407 ESEs/nt for the constitutive and alternative exons, respectively.

degenerate, as was observed for the IgM-derived motifs (22,23,26), and this probably reflects the need for SF2/ASF-specific ESEs to function in multiple exonic contexts, with differential protein-coding specificities. The selected winner pools demonstrated enhanced splicing activity in response to SF2/ASF, compared with the WT *BRCA1* ESE control, indicating that functional selection had been successful. In addition, the enhancer activity of individual winner clones was higher than both the WT ESE and random clones from the unselected library pool for both the n7 and n14 sequences. In total 112 clones were tested, with all the winner clones splicing well, with the exception of one clone, later found to be a low-score motif. Our results support the conclusion that the position of functional ESEs within a given exon is constrained to a certain degree. The observation that many of the highest score motifs selected from the n14 library were in the same location, with respect to the 3′ splice-site, as the WT enhancer, implies that this position is optimal for enhancer function in *BRCA1* exon 18.

The context for ESE selection provided by the three-exon *BRCA1* minigene makes this splicing substrate more typical of the majority of exons, in that the exon under selection is internal, rather than the terminal M2 exon used in the IgM SELEX experiments. There is evidence that the mechanism of exon definition for 3′-terminal exons is different to the definition of internal exons, in that it involves coupling to 3′-end cleavage and polyadenylation of the pre-mRNA (54,55). Hypothetically, the function of ESEs in terminal exons may be subtly different to their function in internal exons, and this in turn may influence the ESEs selected from within a terminal exon. In addition, the local sequence environment may influence ESE selection. Our data support the conclusion that many of the motifs derived from the IgM-based SELEX are able to function in a different exonic context, in that we found significant overlap between the high-score motifs derived from this screen and our experiments in the context of *BRCA1*. In addition, we recently reported that ESEfinder high-score motifs are present at significantly higher frequencies in internal exons than in their flanking intronic regions (44). The overlap between the IgM- and *BRCA1*-derived motifs implies that, by analogy, many of the *BRCA1*-derived motifs will function as ESEs in other exons.

The score matrices in ESEfinder utilize threshold values, above which a given motif is considered significant, which were based on a statistical discrimination of the winner sequences from the unselected pool (26). We derived thresholds based on experimental data and compared them with statistical thresholds derived in an analogous manner to the ESEfinder threshold values. Our data suggest that the absolute threshold value is less important for the ability to predict splicing outcomes, both *in vitro* and *in vivo*, than the discriminatory power of the matrix. Our results, with the n7-derived matrix in particular, demonstrated that this score matrix was highly accurate in predicting the *in vitro* splicing behavior of individual *BRCA1* clones, and of an independently tested set of 30 *SMN* mutants spliced *in vivo* (40).

A number of winner sequences were demonstrated to enhance exon inclusion in a heterologous exonic context, indicating that the SELEX procedure does not result in motifs that are able to function only in the context of *BRCA1*. Furthermore, our new matrices were able to accurately predict the splicing outcome of the majority of 30 *SMN* mutants spliced *in vivo* (40). Our data give further support to the notion that the splicing defect in *SMN2* that contributes to spinal muscular atrophy is the consequence of the disruption of an SF2/ASF-specific ESE (25,40).

The large number of winner clones demonstrated to be functional ESEs validates our ESE screen. However, it remains possible that a number of the winner sequences are able to function only in the context of *BRCA1* exon 18. This limitation also applies to the functional IgM SELEX. For example, the IgM-derived matrix is not as accurate at predicting the splicing behavior of the *BRCA1* clones tested, although the results were still highly statistically significant ($P < 10^{-7}$, correlation of splicing with motif score). We have addressed this concern by creating a combined *IgM-BRCA1* score matrix. A combined matrix improves specificity, an important consideration for ESEfinder users, as the probability of any given motif being a real ESE is increased if the score is derived from information generated from independent experiments. The combined SF2/ASF matrix was derived from the significant number of overlapping high-score motifs from the IgM and *BRCA1* SELEX matrices. The combined matrix filters out many ESEs that function only in the context of *BRCA1* exon 18 or IgM exon M2, and includes ESEs able to function in both internal and terminal exons. A statistical, rather than the *BRCA1*-biased experimental threshold value was calculated for the combined matrix, based on scoring of the n14 round-zero clones. A value equal to the 70th percentile of the highest score from each

of the round-zero clones was chosen for increased stringency. A number of different threshold values were calculated and applied to the experimental data. The 70th percentile was selected as it gave the best correlation with observed ESE function and largest statistical discrimination between the putative exonic versus intronic ESE frequencies. Scoring 30 *SMN* mutants with the combined SF2/ASF matrix gave significant correlation with the observed *in vivo* levels of exon inclusion, including accurately predicting the behavior of both *SMN1* and *SMN2*.

An important outcome of our experiments was the derivation of a score matrix that could be applied to identify ESEs in other contexts. ESEfinder (26) has been widely utilized by the research community, and a number of groups have identified disease-associated mutations that correlate with reduced ESEfinder motif scores (24,25,27–32,40). In addition, a recent report described a significant correlation between antisense oligonucleotides able to induce specific exon-skipping with high-score ESEfinder motifs in the target exonic sequence (56). We analyzed the ability of the combined SF2/ASF plus the ESEfinder score matrices specific for SRp40, SRp55 and SC35 to predict the effects of 73 single-base substitutions in human disease-associated genes that cause exon-skipping *in vivo* (3,45). The majority of WT sequences containing SF2/ASF motifs had reduced scores in the mutated sequences, and many were reduced below the threshold. The results obtained with the combined matrix were better than using ESEfinder SF2/ASF (data not shown). The effects of point mutations in putative SRp40, SRp55 and SC35 were less conclusive, with a relatively high number of the mutant sequences containing higher score motifs than the WT sequences. One possibility is that a mutated SF2/ASF motif has more pronounced effects upon exon inclusion than mutations in the other SR protein-specific matrices, or that the motifs recognized are partially redundant. Alternatively, further refinement of these other score matrices may be necessary, as the data may be a consequence of the matrices incorrectly predicting some non-existent ESEs.

As a comparison to the predictive power of our new SF2/ASF matrix and ESEfinder, we scored the exon-skipping mutations with the *ab initio* computational methods of RESCUE-ESE and PESX (33,34). Both correctly predicted the exon-skipping phenotype associated with a number of the mutations. There was some overlap between the motifs recognized by the three methods, and this was more significant for motifs recognized by both RESCUE-ESE and PESE. Our data is in accordance with a recent report analyzing mutations in 22 predicted PESEs (57). Both ESEfinder and RESCUE-ESE motifs were affected by 11/18 of the PESE mutations demonstrated to have an effect on exon inclusion, and the majority of these were in different sequences (57). We analyzed the overlap between the motifs recognized by our new SF2/ASF matrix and RESCUE-ESE and PESEs, as we have previously calculated for the ESEfinder matrices (44). The motifs recognized as ESEs by the SF2/ASF matrix do not overlap with the RESCUE-ESE motifs. However, we found a significant overlap between our high-score SF2/ASF motifs and PESEs (data not shown), as we previously described for the IgM SELEX-derived SF2/ASF motifs (44). All three methods are useful tools for ESE prediction, and all have been experimentally validated. The constraints, experimental or computational, imposed by the three methods result in significant differences in the motifs predicted to be ESEs. Our data support the notion that all three methods should be used in parallel when searching for putative ESEs.

We recently reported that the frequency of putative SR-protein-ESE motifs is higher in constitutive exons than in their flanking intronic regions (44). We repeated this analysis with our new SF2/ASF matrix on a second set of constitutively spliced exons and a set of alternatively spliced exons, and observed a similar distribution of ESE motifs. Our data reinforce the hypothesis that ESEs are required for splicing of most, if not all, exons, and are not limited to a function in the splicing of alternative exons. The statistical difference in ESE frequency between exonic and intronic regions was greater with our second generation SF2/ASF matrix than the distribution of high-score SF2/ASF motifs generated by scoring the same datasets with the original SF2/ASF matrix. This large-scale analysis provides further evidence that the new matrix is more specific. ESEfinder has now been updated to include the new SF2/ASF matrix, as well as the original four SR protein-specific matrices, increasing the robustness of ESE prediction.

MATERIALS AND METHODS

Library construction and functional SELEX

Two SELEX libraries were constructed in which the naturally occurring SF2/ASF-specific 7-nt ESE in exon 18 of *BRCA1* (24) was replaced with random sequences of either 7 or 14 nt—n7 and n14, respectively—by sequential PCR using high-fidelity Deep Vent polymerase (New England Biolabs). The *BRCA1* template is a three-exon minigene with a T7 RNA polymerase promoter, allowing RNA transcripts to be derived. Several oligonucleotide pools containing the randomized regions were obtained in which the manufacturer (Sigma) had optimized the phosphoramidite concentrations to balance the nucleotide composition within the randomized region. Libraries were constructed, random clones sequenced and the n7 and n14 libraries that contained randomized regions closest to 25% representation of each nucleotide were chosen for the SELEX experiments. The location of the *BRCA1* exon 18 natural ESE at position +4 made it impossible to regenerate full-length constructs by PCR following the SELEX procedure. Therefore, for cloning purposes, a *Bae*I restriction site was created by site-directed mutagenesis within exon 17. A second construct (*BRCA1* C), in which the reciprocal *Bae*I site was engineered in *BRCA1* exon 18, was made solely for cloning purposes. The SELEX procedure was carried out as previously described, with some modifications (22,23). An overview of the procedure is illustrated in Figure 1. The *BRCA1* SELEX libraries were *in vitro* transcribed with T7 polymerase (Promega), generating radiolabeled pre-mRNA substrate pools. About 10 fmol of the pre-mRNA pools was incubated under *in vitro* splicing conditions in S100 extract plus 10 pmol of recombinant SF2/ASF in 25 μl reaction mixtures. The RNAs were separated by denaturing polyacrylamide gel electrophoresis, and

spliced mRNAs excised and eluted from the gel in 0.5 M sodium acetate, 1 mM ethylenediaminetetraacetic acid and 0.2% sodium dodecyl sulfate. RNAs were recovered by ethanol precipitation and reverse-transcribed using Superscript II, as described by the manufacturer (Invitrogen), using an exon-specific primer. Full-length cDNAs were amplified by PCR using Deep Vent polymerase, and then digested with *Bae*I (New England Biolabs) followed by agarose-gel purification (Qiagen gel-extraction kit). Exon 17 plus intron 17, produced from *Bae*I digestion of the cloning construct (C) was then ligated to the *Bae*I-digested SELEX products using T4 DNA ligase (Gibco-BRL). Ligation products were purified by agarose-gel electrophoresis, and full-length splicing constructs rebuilt by overlap-extension PCR. Oligonucleotide sequences are available upon request. Following each round of selection, the winner pools were sub-cloned into the vector pCR-Blunt (Invitrogen) and sequenced by use of a Dye Terminator Cycle Sequencing kit (Perkin-Elmer) and an automated ABI 377 sequencer.

Preparation of HeLa cell extracts and recombinant SF2/ASF

HeLa nuclear and S100 extracts, and recombinant SF2/ASF were prepared as described (58,59). The integrity and purity of SF2/ASF was checked by SDS–PAGE and the specific activity determined by *in vitro* splicing of β-globin pre-mRNA in S100 extract. SF2/ASF and extract concentrations for use in the SELEX experiments were optimized using the splicing minigenes *BRCA1*-WT and nonsense mutant *BRCA1* E1694X (*BRCA1*-MT), the latter of which is known to cause skipping of exon 18 (24). The same S100 and SF2/ASF preparations were used for all of the *in vitro* experiments.

In vitro splicing

Uniformly [α-^{32}P] UTP-labeled, 5′-capped, T7 runoff transcripts were produced from the SELEX libraries, individual SELEX clones, *BRCA1* control substrates and *SMN* constructs. Transcripts were purified by denaturing PAGE and spliced in HeLa cell nuclear extract or S100 post-nuclear extract under standard conditions with 1.6 mM $MgCl_2$ (60). SELEX experiments were performed in a volume of 25 μl, and all other splicing reactions were performed in a volume of 12.5 μl. After incubation at 30°C for 4 h, the reactions were phenol-extracted and the RNA precipitated with ethanol. The reaction products were resolved on 12% denaturing polyacrylamide gels, followed by autoradiography and phosphorimage analysis.

Sequence analysis and construction of score matrices

Consensus motifs were identified from the n14 winner sequences by alignment of the sequences using three motif-finding algorithms: (1) Gibbs sampler (37); (2) MEME (38); (3) DME (39). An additional alignment of the n14 winners was performed by scoring the sequences with the n7-derived matrix. Alignment of the n7 winners was not required because of the ESE position being fixed. The aligned sequences were used to derive consensus motifs from which nucleotide-frequency score matrices were constructed using established methods (22,23). The compositional bias of the initial RNA pools was taken into account. A detailed description of score-matrix construction is given in Liu *et al*. (22).

Analysis of genomic ESE motif frequencies

Datasets of constitutively spliced and alternatively spliced human internal protein-coding exons were retrieved from the Alternative Splicing Database (61). Genomic sequences and transcript annotations were downloaded from http://www.ebi.ac.uk.asd/. Constitutively spliced exons and their flanking intronic sequences were extracted by a Perl script from 6291 intron-containing genes without any EST or cDNA evidence of alternative splicing. The dataset was filtered to remove exons flanked by short introns (<250 nt), and short exons (<106 nt). Composite exons were created from this set of 16 635 exons, comprising 25 nt from each end and 50 nt from the center. The composite exons, and 100 nt of upstream and downstream flanking intronic sequence, were scored for the presence of high-score SF2/ASF motifs. A total of 5041 alternatively spliced cassette exons and their flanking intronic regions were extracted from 3644 genes, filtered and composite exons created as earlier.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.

ACKNOWLEDGEMENTS

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 U.S. Army Medical Research and Matériel Command to P.J.S.

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. *Nat. 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. 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.
6. 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.
7. Graveley, B.R. (2000) Sorting out the complexity of SR protein functions. *RNA*, **6**, 1197–1211.
8. 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.

9. 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.
10. 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.
11. 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.
12. Graveley, B.R., Hertel, K.J. and Maniatis, T. (2001) The role of U2AF35 and U2AF65 in enhancer-dependent splicing. *RNA*, **7**, 806–818.
13. Kan, 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.
14. 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.
15. Shen, H., Kan, J.L., Ghigna, C., Biamonti, G. and Green, M.R. (2004) A single polypyrimidine tract binding protein (PTB) binding site mediates splicing inhibition at mouse IgM exons M1 and M2. *RNA*, **10**, 787–794.
16. Huang, Y. and Steitz, J.A. (2005) SRprises along a messenger's journey. *Mol. Cell*, **17**, 613–615.
17. Ibrahim, E.C., Schaal, T.D., Hertel, K.J., Reed, R. and Maniatis, T. (2005) Serine/arginine-rich protein-dependent suppression of exon skipping by exonic splicing enhancers. *Proc. Natl Acad. Sci. USA*, **102**, 5002–5007.
18. 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.
19. Tian, H. and Kole, R. (1995) Selection of novel exon recognition elements from a pool of random sequences. *Mol. Cell. Biol.*, **15**, 6291–6298.
20. Boukis, L.A. and Bruzik, J.P. (2001) Functional selection of splicing enhancers that stimulate *trans*-splicing *in vitro. RNA*, **7**, 793–805.
21. 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.
22. 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.
23. 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.
24. 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.
25. 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.
26. 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.
27. 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.
28. Ferrari, S., Giliani, S., Insalaco, A., Al-Ghonaium, A., Soresina, A.R., Loubser, M., Avanzini, M., 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.
29. 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.
30. Mas, C., Taske, N., Deutsch, S., Guipponi, M., Thomas, P., Covanis, A., Friis, M., Kjeldsen, M.J., Pizzolato, G.P., Villemure, J.G. *et al*. (2004) Association of the *connexin36* gene with juvenile myoclonic epilepsy. *J. Med. Genet.*, **41**, e93.
31. 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.
32. 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.
33. 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.
34. Zhang, X.H. and Chasin, L.A. (2004) Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev.*, **18**, 1241–1250.
35. Pagani, F. and Baralle, F.E. (2004) Genomic variants in exons and introns: identifying the splicing spoilers. *Nat. Rev. Genet.*, **5**, 389–396.
36. Watakabe, A., Tanaka, K. and Shimura, Y. (1993) The role of exon sequences in splice site selection. *Genes Dev.*, **7**, 407–418.
37. Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F. and Wootton, J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
38. Bailey, T.L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
39. Smith, A.D., Sumazin, P. and Zhang, M.Q. (2005) Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc. Natl Acad. Sci. USA*, **102**, 1560–1565.
40. Cartegni, L., Hastings, M.L., Calarco, J.A., de Stanchina, E. and Krainer, A.R. (2006) Determinants of exon 7 splicing in the spinal muscular atrophy genes, *SMN1* and *SMN2*. *Am. J. Hum. Genet.*, **78**, 63–77.
41. Lorson, C.L., Hahnen, E., Androphy, E.J. and Wirth, B. (1999) A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. *Proc. Natl Acad. Sci. USA*, **96**, 6307–6311.
42. Lefebvre, S., Burglen, L., Reboullet, S., Clermont, O., Burlet, P., Viollet, L., Benichou, B., Cruaud, C., Millasseau, P., Zeviani, M. *et al*. (1995) Identification and characterization of a spinal muscular atrophy-determining gene. *Cell*, **80**, 155–165.
43. 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.
44. Wang, J., Smith, P.J., Krainer, A.R. and Zhang, M.Q. (2005) Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes. *Nucleic Acids Res.*, **33**, 5053–5062.
45. Valentine, C.R. (1998) The association of nonsense codons with exon skipping. *Mutat. Res.*, **411**, 87–117.
46. Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M. and Burge, B. (2004) Systematic identification and analysis of exonic splicing silencers. *Cell*, **119**, 831–845.
47. Krainer, A.R., Conway, G.C. and Kozak, D. (1990) Purification and characterization of pre-mRNA splicing factor SF2 from HeLa cells. *Genes Dev.*, **4**, 1158–1171.
48. Ge, H. and Manley, J.L. (1990) A protein factor, ASF, controls cell-specific alternative splicing of SV40 early pre-mRNA *in vitro*. *Cell*, **62**, 25–34.
49. Mayeda, A. and Krainer, A.R. (1992) Regulation of alternative pre-mRNA splicing by hnRNP A1 and splicing factor SF2. *Cell*, **68**, 365–375.
50. Zhang, Z. and Krainer, A.R. (2004) Involvement of SR proteins in mRNA surveillance. *Mol. Cell*, **16**, 597–607.
51. Li, X. and Manley, J.L. (2005) Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability. *Cell*, **122**, 365–378.
52. Huang, Y., Gattoni, R., Stévenin, J. and Steitz, J.A. (2003) SR splicing factors serve as adapter proteins for TAP-dependent mRNA export. *Mol. Cell*, **11**, 837–843.
53. Sanford, J.R., Gray, N.K., Beckmann, K. and Cáceres, J.F. (2004) A novel role for shuttling SR proteins in mRNA translation. *Genes Dev.*, **18**, 755–768.
54. Niwa, M., MacDonald, C.C. and Berget, S.M. (1992) Are vertebrate exons scanned during splice-site selection? *Nature*, **360**, 277–280.
55. Vagner, S., Vagner, C. and Mattaj, I.W. (2000) The carboxyl terminus of vertebrate poly(A) polymerase interacts with U2AF 65 to couple 3′-end processing and splicing. *Genes Dev.*, **14**, 403–413.
56. Aartsma-Rus, A., De Winter, C.L., Janson, A.A., Kaman, W.E., Van Ommen, G.J., Den Dunnen, J.T. and Van Deutekom, J.C. (2005) Functional analysis of 114 exon-internal AONs for targeted *DMD* exon skipping: indication for steric hindrance of SR protein binding sites. *Oligonucleotides*, **15**, 284–297.

57. Zhang, X.H., Kangsamaksin, T., Chao, M.S., Banerjee, J.K. and Chasin, L.A. (2005) Exon inclusion is dependent on predictable exonic splicing enhancers. *Mol. Cell. Biol.*, **25**, 7323–7332.
58. Mayeda, A. and Krainer, A.R. (1999) Preparation of HeLa cell nuclear and cytosolic S100 extracts for *in vitro* splicing. *Meth. Mol. Biol.*, **118**, 309–314.
59. Krainer, A.R., Mayeda, A., Kozak, D. and Binns, G. (1991) Functional expression of cloned human splicing factor SF2: homology to RNA-binding proteins, U1 70K, and *Drosophila* splicing regulators. *Cell*, **66**, 383–394.
60. Mayeda, A. and Krainer, A.R. (1999) Mammalian *in vitro* splicing assays. *Meth. Mol. Biol.*, **118**, 315–321.
61. Thanaraj, T.A., Stamm, S., Clark, F., Riethoven, J.J., Le Texier, V. and Muilu, J. (2004) ASD: the alternative splicing database. *Nucleic Acids Res.*, **32**, D64–D69.
62. Burge, C.B., Tuschl, T. and Sharp, P. (1999) Splicing of precursors to mRNAs by the spliceosomes. Gesteland, R.F., Cech, T.R. and Atkins, J.F. (eds), *The RNA world*, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 525–559.