8928–8937 *Nucleic Acids Research, 2011, Vol. 39, No. 20* *Published online 27 July 2011*
doi:10.1093/nar/gkr481

# Efficient internal exon recognition depends on near equal contributions from the 3′ and 5′ splice sites

Peter J. Shepard[1], Eun-A. Choi[1], Anke Busch[1] and Klemens J. Hertel[1,*]

[1]Department of Microbiology & Molecular Genetics, University of California, Irvine, Irvine, CA 92697-4025, USA

Received September 20, 2010; Revised May 13, 2011; Accepted May 24, 2011

ABSTRACT

Pre-mRNA splicing is carried out by the spliceosome, which identifies exons and removes intervening introns. In vertebrates, most splice sites are initially recognized by the spliceosome across the exon, because most exons are small and surrounded by large introns. This gene architecture predicts that efficient exon recognition depends largely on the strength of the flanking 3′ and 5′ splice sites. However, it is unknown if the 3′ or the 5′ splice site dominates the exon recognition process. Here, we test the 3′ and 5′ splice site contributions towards efficient exon recognition by systematically replacing the splice sites of an internal exon with sequences of different splice site strengths. We show that the presence of an optimal splice site does not guarantee exon inclusion and that the best predictor for exon recognition is the sum of both splice site scores. Using a genome-wide approach, we demonstrate that the combined 3′ and 5′ splice site strengths of internal exons provide a much more significant separator between constitutive and alternative exons than either the 3′ or the 5′ splice site strength alone.

INTRODUCTION

The removal of intronic regions is carried out by the spliceosome, which recognizes key splicing signal sequences on the pre-mRNA molecule and catalyzes the removal of introns and the joining of exons to form mature mRNA molecules. The spliceosome displays a high degree of fidelity by efficiently pairing constitutively spliced exons separated by introns up to 10^5 nucleotides in length (1). At the same time, >90% of genes contain at least one exon that is alternatively spliced (2). How the spliceosome achieves this balance between splicing flexibility and fidelity is a long-standing question in the RNA processing field. The alternative splicing decision is largely determined by the primary sequence of the pre-mRNA molecule itself. The best-known factors influencing the splicing decision are the 3′ and 5′ splice site sequences at either end of an internal exon. Indeed, many known disease causing point mutations disrupt splice site signals, thereby altering the normal splicing pattern (3,4). Thus, identifying the contribution that these splice site signals make towards efficient exon recognition is a fundamental aspect in understanding exon recognition.

In vertebrates, most exons are small (the vast majority between 50 and 200 nt) surrounded by much larger introns (5). Initial recognition of splice sites is favored when the distance between the splice sites is short (6). Therefore, recognition of splice sites occurs across the intron of lower eukaryotes that have small introns and large exons (7). However, as the length of the intron increases beyond 250 nt, splice site recognition across the intron becomes less efficient (6). The average size of vertebrate exons is 137 nt separated by much longer introns, an architecture which favors splice site recognition across the exon (exon recognition) (8). Indeed, small exons surrounded by large introns are efficiently recognized by the spliceosome, but exons expanded beyond 300 nt are skipped or an internal cryptic splice site is utilized yielding a shorter exon (7). Exon recognition is also inefficient for very short exons. This was demonstrated by deletion studies that showed that reducing the size of an internal exon below 50 nt induced a skipping phenotype (9). Therefore, the optimal unit of recognition for the vertebrate spliceosome is an exon between 50 and 250 nt in length.

Two crucial components in the initial recognition of exons are the 5′ and the 3′ splice sites. The 5′ splice site is defined by a single 9 nt sequence. The 3′ splice site is defined by a branch point sequence usually within 40 nt of the intron/exon junction, a polypyrimidine tract and the 3′ intron/exon junction (10). The 5′ splice site is initially recognized by the U1 snRNP, which binds to the 5′ exon/intron junction. Initial recognition of the intron/exon 3′ splice site requires U2AF association with the

*To whom correspondence should be addressed. Tel: +949 824 2127; Fax: +949 824 8598; Email: khertel@uci.edu

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

polypyrimidine tract and U2 snRNP with the branch point sequence. Because the sequence specificity of these interactions is driven by pre-mRNA/snRNA interactions and the U2AF binding preference for polypyrimidines, splice sites are classified by their complementarity to U1 snRNA (5′ splice site) and the extent of the polypyrimidine tract (3′ splice site). Greater complementarity with U1 snRNA and longer polypyrimidine tracts translate into higher affinity binding sites for these spliceosomal components and, thus, more efficient exon recognition (11,12).

To test the contribution that the 3′ and 5′ splice sites make toward efficient exon recognition, we constructed minigenes harboring an internal exon under the control of exon definition. We tested a range of splice site strengths at the 3′ and 5′ splice sites to show that the level of internal exon inclusion is almost equally influenced by the 3′ and 5′ splice site strength. The experimental evaluation was extended by performing a genome-wide analysis of constitutively and alternatively spliced exons. Our results show that the combined splice site strength of the exons distinguishes the two classes of exons much better than either splice site individually. We conclude that efficient recognition of internal exons is dictated nearly equally by the spliceosomal binding potential of the 3′ and 5′ splice sites.

MATERIALS AND METHODS

RNAs

To generate 3-exon minigenes with the internal exon under the control of exon definition, we first made two 2-exon minigenes with long introns (>250 nt). The pAdML minigene (13) intron was expanded by inserting a 236 nt intronic fragment obtained from *SMN* intron 6. The β-globin minigene intron 1 was expanded as described elsewhere (14). A fragment containing the first exon, intron and 3′ splice site of the expanded AdML (forward primer CGTTCTAGATCCATCGATGAATTCGAGC TC, reverse primer GGTCTAGAGCGTCGACCTGCA GCTGTG) was inserted upstream into the XbaI site of the expanded β-globin to generate a 3-exon minigene that contains exon 1 and intron 1 from AdML, an internal β-globin exon, followed by β-globin intron 1 and exon 2. The 3′ splice site of the internal exon was replaced by exchanging a fragment between the HindIII and PstI restriction sites to generate five different 3′ spice sites of variable strengths (Table 1). The 5′ splice site of the internal exon was replaced using an ExoIII cloning approach (15). The internal exon was amplified using one forward (Fwd) and five different reverse (Rev) primers each containing a 5′ splice site of different strength

(Fwd GGGTTTCCTTGAAGCTTTCGTGCTGACC;

10.9-Rev TCAAGCTAGCTTAAGTCTGTCTTGTAGG CTTGATACTTACCTGCTCG;
8.1-Rev TCAAGCTAGCTTAAGTCTGTCTTGTAGG CTTGATACCAACCTGCTCG;
−0.5-Rev TCAAGCTAGCTTAAGTCTGTCTTGTAG GCTTGATTGACACCTCCTCG;
−5.2-Rev TCAAGCTAGCTTAAGTCTGTCTTGTAG GCTTGATATCGACCGACTCG;
−9.1-Rev TCAAGCTAGCTTAAGTCTGTCTTGTAG GCTTGATATCCACAGTCTCG).

The variable 3′ splice site plasmids were digested with HindIII and AflII and ExoIII cloning was performed (15) to obtain 25 minigenes. All plasmids were subsequently subcloned into a mammalian expression vector (intronless pCi) under the control of a CMV promoter.

Cell transfection

HeLa cells were plated to 1.5×10^5 cells/well in six well plates the day prior to transfection and grown in MEM (Cellgro) supplemented with 10% FBS (vol/vol), 2 mM glutamine and 10 mM sodium pyruvate. Lipofectamine 2000 (Invitrogen) was used to transfect 1 μg plasmid following manufacturer's protocol and cells were harvested 24 h later. Total RNA was extracted using TRIzol (Invitrogen) followed by phenol chloroform extraction and isopropyl alcohol precipitation. DNA contamination was removed using DNaseI (Invitrogen). RNA was reverse transcribed using iScript (BioRad). PCR reactions were carried out to detect spliced products and unspliced pre-mRNA using plasmid specific primers (Fwd GCTAA CGCAGTCAGTGCTTC) (Rev GTATCTTATCATGTC TGCTCG). PCR products were resolved on a 1% agarose gel. The fraction of exon inclusion was calculated as [included/(included + excluded)]. Each transfection was repeated at least three times to determine the variance between biological replicates.

Computational analysis of alternative splicing

The Alternative Splice Database (ASD) was used as the source for sequence and alternative splicing information (5). The complete list of exons from the ASD was downloaded and Perl scripts were used to extract 11 000 skipping events with only one exon in the skipping event (5). To create a list of constitutively spliced exons, internal exons of all known human isoforms were downloaded from the UCSC Genome Browser (16,17) using the UCSC Genes track of genome assembly hg19. From this list of internal exons all exons overlapping with an intron or with another exon were removed as well as exons that have alternative 3′ or 5′ splice sites or exons whose neighbors are involved in alternative splicing (as cassette exon or with an alternative 3′ or 5′ splice site). The remaining list was further reduced by filtering it through a list of spliced ESTs (UCSC track: Spliced ESTs, table: intronEst) and mRNAs (UCSC track: Human mRNAs, table: all_mrna) downloaded from UCSC. Again, exons were neglected if they overlap with an intron or exon, or if they or their neighbors are involved in alternative splicing. Exons that were not found in ESTs or only occur as first or last exons in ESTs were also removed. The remaining list includes 37 473 exons. Only exons with a high EST coverage (20 or more EST entries) were used for analysis, resulting in a final data set of 3280 constitutively

Table 1. Sequences used for 3′ and 5′ splice sites

MaxEnt	S&S	Sequence	Quality
3′ splice sites			
12.6	92	UGUCCCUUUUUUUUCCACAG/CUG	Very strong
9.2	88	UUUUUUUUUUUUUUGUCUAG/CUG	Strong
5.7	77	CUUUACUUCUAUGACUGUAG/CUG	Medium
5.3	73	GUGACUGUGUGUAUGCACAG/CUG	Weak
−4.8	59	AUUGUGAUCGCAGCCAAUAG/CUG	Very weak
5′ splice sites			
10.9	100	CAG/GUAAGU	Very strong
8.1	80	CAG/GUUGGU	Strong
−0.5	64	GAG/GUGUCA	Medium
−5.2	55	UCG/GUCGAU	Weak
−9.1	51	ACU/GUGGAU	Very weak

/ denotes the intron/Exon junction. S&S refers to splice site strength scores based on the originally proposed nucleotide weight tables (30). S&S values were calculated using the Analyzer-Splice-Tool calculator (http://ibis.tau.ac.il/ssat/SpliceSiteFrame.htm).

splice exons. The alternative and constitutive exon datasets were purged of any overlapping exons and the splice site sequence +20 to −3 for the 3′ splice site and −3 to +6 for the 5′ splice site were downloaded using the faToTwoBit download tool (http://hgdownload.cse.ucsc.edu/downloads.html). The splice site scores were obtained for each 5′ and 3′ splice site sequence using the Maximum Entropy scores (18) (http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html).

Estimation of exon inclusion from EST entries

From the list of 11 000 skipped exons containing only one exon in the skipping event (above), we estimated the inclusion level of the skipped exons for each skipping event. This was done by dividing the number of ESTs that have isoforms containing the exon in the skipping event by the total number of ESTs representing both the included and excluded isoforms for the skipping event. The EST representation for each isoform was obtained from the ASD (5).

The Linear Discriminant Analysis (LDA) was performed using the MASS package from statistical software program R (http://www.r-project.org/).

High-throughput sequencing

HeLa cells were plated in T21 flasks and grown to confluency in MEM (Cellgro) supplemented with 10% FBS (vol/vol), 2 mM glutamine and 10 mM sodium pyruvate. Total RNA was isolated using the RNAeasy protocol (Quiagen). Sequencing libraries were generated using the Illumina mRNAseq kit according to manufactory protocols. The cDNA library was sequenced on an Illumina GAII instrument generating approximately 32.5 million 76 nt paired end reads (65 million reads in total). Alignment of the sequencing reads was carried out using an in-house computational platform that includes matching reads to an extensive list of splice junctions. Internal exon inclusion levels were determined by calculating the fractional representation from splice junction reads that define exon skipping and splice junction reads that define exon inclusion.

ROC analysis

In order to determine the ability of the 5′ or 3′ splice site strength, or their additive combination to distinguish between high- and low inclusion exons, we performed a ROC curve analysis on EST verified alternative splicing events. All alternatively spliced exons with an inclusion level of 90% and higher were combined with constitutive exons to generate a new set of highly included exons. They were compared to a set of rarely included exons, consisting of all alternatively spliced exons with an inclusion level of 10% and lower. All exons with an EST coverage of at least 10 EST entries were retained for analysis.

Determining the free energy of sequences flanking splice sites

RNA structure analysis was performed as previously described (24). The RNA free energy minimization program RNAfold was downloaded locally from http://www.tbi.univie.ac.at/RNA/, and used to find the minimum free energy of 60-mers.

RESULTS

The level of internal exon inclusion is affected by both the 3′ and 5′ splice site strengths

To measure the influence of the 3′ and 5′ splice site strength on internal exon recognition, we constructed a series of 3-exon minigenes with the internal exon under the control of exon definition (Figure 1A). Included in these minigenes were restriction sites flanking the 3′ and 5′ splice sites of the internal exon that enabled replacing the splice sites of this exon. Five 3′ splice site sequences of variable strength ranging from strong to weak and five 5′ splice site sequences of variable strength ranging from strong to weak (Table 1) were designed and all possible combinations of these sequences were inserted into the 3-exon minigene resulting in 25 splicing substrates (The complete sequence of one of these is provided in Supplementary Figure S1). The selection of the 5′ and 3′ splice sites was based on predicted splice site strength using maximum entropy (MaxEnt) scores (18), a computationally derived probability function that a given splice site will be used. Internal exon inclusion ratios of the resulting 25 minigenes were measured from transfection experiments in HeLa cells and the results are summarized in Table 2, arranged in five groups based on the 3′ splice site strength.

When comparing internal exclusion levels as a function of 5′ splice site strength, a clear correlation is seen between its binding potential with U1 snRNP and its ability to rescue weak 3′ splice sites. For example, a strong 5′ splice site is able to direct efficient spliceosomal recruitment even when the internal exon is flanked by a 3′ splice site that is not expected to support splicing (Figure 1B). However if the 5′ splice site itself is relatively weak, only strong 3′ splice sites allow exon recognition (Figure 1C). A reciprocal relationship is observed when evaluating internal exon selection in the context of fixed 3′ splice sites (Figure 1D and E).

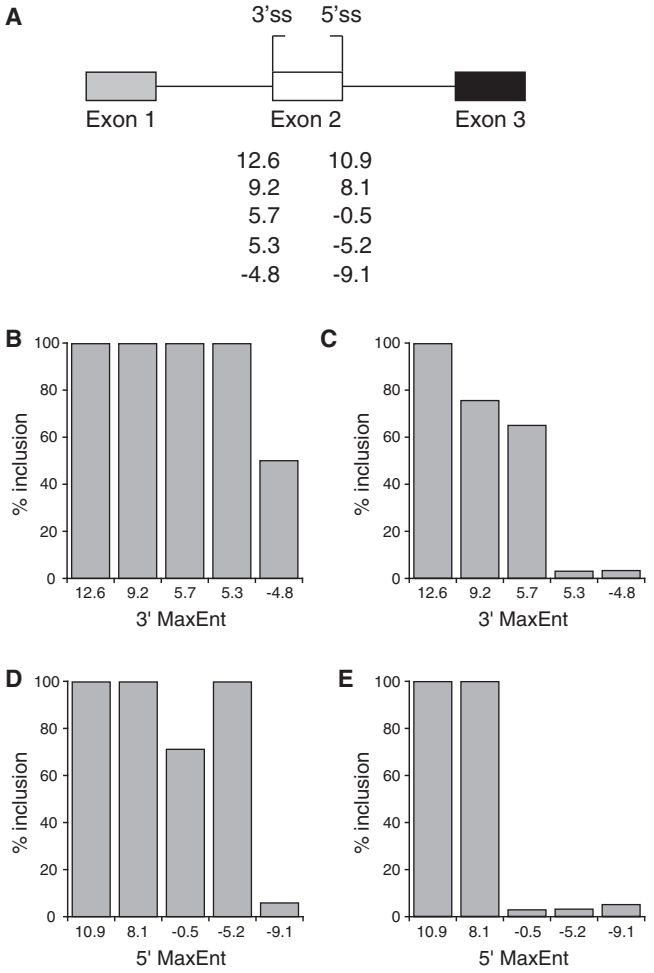

Figure 1. Interdependence of splice site strength. (**A**) Isogenic pre-mRNAs were tested for internal exon inclusion, which differ only in their 5′ or 3′ splice site sequence. Five different sequences at the 5′ splice site were matched with five different sequences at the 3′ splice site to generate 25 test substrates. The numbers below the 5′ and the 3′ splice site denote the MaxEnt score for the splice sites used. The sequences of the splice sites are listed in Table 1. Each graph shows the internal inclusion level of variable 3′ splice sites (**B** and **C**) or 5′ splice sites (**D** and **E**), while the complementing splice site was held constant. Panel B shows exon inclusion levels of variable 3′ splice sites in the context of a very strong 5′ splice site of MaxEnt score 10.9. (C) Same as in (B), however in the context of a weak 5′ splice site (−5.2). (D) and (E) show exon inclusion levels of variable 5′ splice sites in the context of a very strong 3′ splice site (12.6) or a weak 3′ splice site (5.3), respectively.

Table 2. Internal exon inclusion levels

Substrate no.	3′/5′ splice site	Inclusion/ Exclusion ratio	SD ratio	Inclusion (%)
1	12.6/10.9	332	0.4	99.7
2	12.6/8.1	332	0.4	99.7
3	12.6/−0.5	2.5	0.7	71.2
4	12.6/−5.2	500	0.2	99.8
5	12.6/−9.1	0.062	0.009	5.9
6	9.2/10.9	500	0.2	99.8
7	9.2/8.1	332	0.09	99.7
8	9.2/−0.5	21	0.08	95.4
9	9.2/−5.2	3.1	0.06	75.5
10	9.2/−9.1	[a]	[a]	[a]
11	5.7/10.9	250	0.1	99.6
12	5.7/8.1	250	0.2	99.6
13	5.7/−0.5	250	0.1	99.6
14	5.7/−5.2	1.8	0.08	64.9
15	5.7/−9.1	0.01	[b]	1
16	5.3/10.9	250	0.5	99.6
17	5.3/8.1	332	0.02	99.7
18	5.3/−0.5	0.03	0.02	2.7
19	5.3/−5.2	0.03	0.004	3.2
20	5.3/−9.1	0.05	[b]	5.1
21	−4.8/10.9	1.0	0.3	50
22	−4.8/8.1	[a]	[a]	[a]
23	−4.8/−0.5	0.23	0.09	18.7
24	−4.8/−5.2	0.03	0.02	3.3
25	−4.8/−9.1	0.03	0.02	2.7

Semi-quantitative PCR analysis was carried out to determine the exon inclusion ratio. The experiments were carried out at least three different times to determine the variation between biological replicates.
[a]Not detected, indicating that neither exon inclusion nor exon exclusion was detected.
[b]No standard deviation was calculated because only two repetitions of the experiments were available.

When correlating internal exon inclusion levels with either the 3′ splice site strength alone (Figure 2A) or with the 5′ splice site strength alone (Figure 2B), it is apparent that considering only one splice site at a time is insufficient to accurately predict internal exon inclusion levels. While the measured inclusion levels follow the expected general trend that stronger splice sites mediate greater exon recognition (11,14), this is not always the case. We found a weak correlation between levels of internal exon inclusion and the 3′ splice site score (Pearson correlation = 0.48) or the 5′ splice site score (Pearson correlation = 0.66) alone. However, strong correlations (Pearson correlation = 0.82) can be obtained between exon inclusion levels and splice site scores if the sum of the splice site scores is evaluated (Figure 2C). Interestingly, the transition between exon inclusion and exclusion appears to be quite narrow, resulting in practically full inclusion or exclusion (Figure 2C and D). It is possible that such a narrow transition between exon inclusion and exclusion could be the consequence of the splice site scoring method used. However, analyzing the splicing data in the context of other splice site scoring methods demonstrates that the MaxEnt scoring method applied is not a cause for the steep exon inclusion transition (Supplementary Figure S2). We conclude that internal exon inclusion levels are tightly tuned to 5′ and 3′ splice site strength.

While the overall trend of the data presented in Figure 2C supports the notion that the sum of the splice site scores is an accurate predictor of exon inclusion levels, there are some outliers. In some cases, the observed deviations may be explained by differential propensities of the cloned splice sites to form local RNA secondary structures. For example, splicing substrate 23 (the exon with flanking splice site strengths of 5′ss = −4.1/3′ss = −0.5) displays a higher than expected inclusion level (18.7%) when compared to the next closest in the series, substrate 18 (5.3/−0.5, 2.7% inclusion, Table 2). The RNA secondary structure potential around the 3′ splice site differs significantly between these constructs (ΔG −15.4 kcal/mol

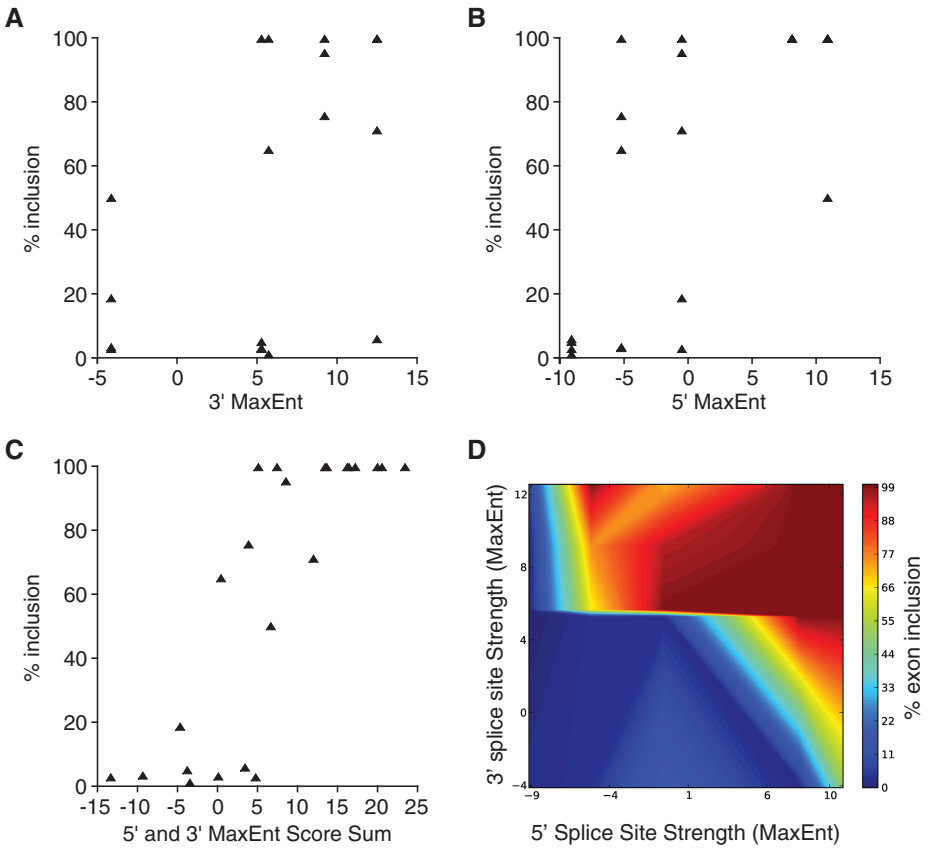

Figure 2. Correlation between splice site strength and internal exon inclusion. Internal exon inclusion levels measured for the 25 splicing substrates are correlated to (**A**) 3′ splice site MaxEnt scores (Pearson correlation = 0.48), (**B**) 5′ splice site MaxEnt scores (Pearson correlation = 0.66), (**C**) or the sum of 5′ and 3′ splice site MaxEnt scores (Pearson correlation = 0.82). (**D**) Interdependence of splice site strength and exon inclusion. The three-dimensional plot correlates 3′ (*y*-axis) and 5′ (*x*-axis) splice site strength with exon inclusion levels (color scale) to highlight the narrow transition between exon inclusion and exclusion. Exon inclusion levels are represented from dark red (fully included) to dark blue (fully excluded) on the *z*-axis.

for #23 versus −21.2 kcal/mol for #18), suggesting that the more single stranded character of the 3′ splice site of construct 23 promoted higher than anticipated exon inclusion levels. In other cases, deviations from the overall trend can be explained by the observation that when an exon is flanked by one very poor splice site, exon inclusion cannot be rescued, even if the other splice site is extremely strong. This is the case for substrate #5, where the strongest 3′ splice site is matched with the weakest 5′ splice site tested. Based on similar grounds the lower than anticipated exon inclusion levels of substrate #21 can be explained. We conclude that a strong splice site cannot rescue a very weak splice site at the other end of the exon and that local RNA secondary structure formation may interfere with splice site recognition.

One striking feature of the exon inclusion analysis was the observation that the transition between exon exclusion and exon inclusion is very tight, within a window of approximately 2–3 MaxEnt units. In addition, the splicing pattern erred on either the completely included or completely excluded side, with few substrates displaying intermediate inclusion levels (Figure 2C). To evaluate the narrow splice pattern transition in an alternative splice site selection scheme, we analyzed alternative 5′ splice site selection in the context of a fixed distal splice site and variable proximal splice sites (Figure 3A). In agreement with our exon inclusion analysis, the transition between alternative splice site usage occurs within a very narrow window, essentially transitioning completely from exclusive distal to exclusive proximal splice site selection (Figure 3B). These results underscore the notion that splice site selection and exon recognition is a highly fine tuned process resulting in either minimal or maximal recognition by the splicing machinery.

Relative contributions of the 3′ and 5′ splice site in exon recognition

To complement the biochemical analysis using isogenic pre-mRNAs that deviate from each other only in the splice sites of the internal exon, we carried out bioinformatic analyses of constitutive and alternative exons. A database was created containing human constitutively spliced exons and alternatively included exons from ESTs. To correlate the combined splice site score with the levels of exon inclusion, we determined the fraction of exon inclusion for each alternative splicing event by comparing the number of EST entries that support exon inclusion or exon exclusion (see 'Materials and Methods' section). In agreement with the experimentally determined

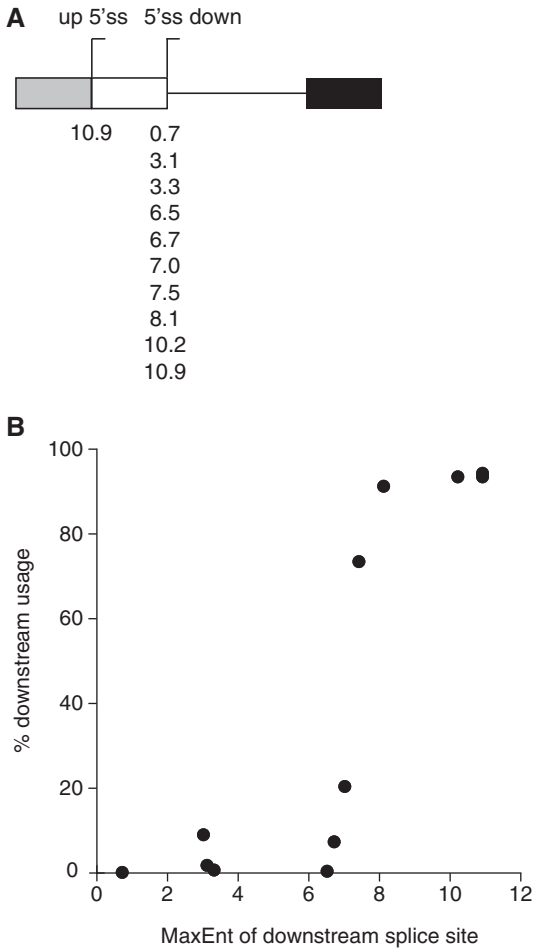

Figure 3. Correlation between splice site strength and alternative 5′ splice site selection. (**A**) Isogenic pre-mRNAs were tested for downstream splice site usage. Ten different sequences at the downstream 5′ splice site were matched with a constant upstream 5′ splice site. The numbers below the splice sites denote the MaxEnt score for the splice sites used. (**B**) Downstream to upstream splice site usage measured for the 10 splicing substrates are correlated to the downstream splice site strength.

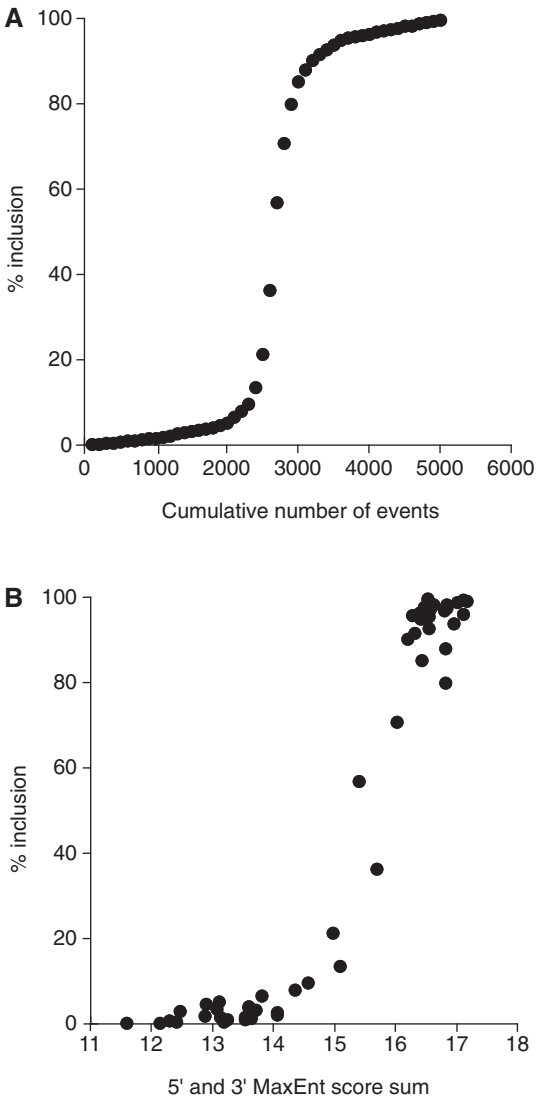

Figure 4. Exon inclusion levels determined from EST sequence entries. (**A**) EST databases were used to extract the levels of internal exon inclusion. The plot shows the relationship between exon inclusion levels and the cumulative number of events. (**B**) Exon inclusion levels are correlated with the average strength of the combined 5′ and 3′ splice site scores for groups of 100 EST entries sorted by inclusion levels.

steep transition between preferential exon inclusion and exclusion (Figure 2C), we observe that the majority of alternative exon inclusion events are either less than 10% included or >90% included, i.e. preferentially excluded or included. Only a small fraction of the alternative splicing events analyzed displayed inclusion levels between 10% and 90% (Figure 4A). These observations support the idea that splicing decisions tend to be skewed toward the extremes of the inclusion spectrum. A prediction of our experimental analysis is that the transition between preferential exon inclusion and preferential exon exclusion tracks with the combined splice site score. Indeed, the correlation between the combined splice site score and exon inclusion levels is sigmoidal and displays a sharp transition between exon exclusion and inclusion (Figure 4B). As was observed experimentally, the transition into preferential exon inclusion occurs within a narrow combined splice site score window ranging from MaxEnt 15 and 16. Importantly, this sharp transition is independent of the splice site scoring method used (Supplementary Figure S3). Given the relatively strong splice sites of preferentially excluded exons, these observations suggest that additional splicing features that reduce the overall capacity of exon definition, such as splicing silencers, are likely to dominate splicing decisions.

A similar analysis was also carried out by correlating exon inclusion levels, as determined from high throughput sequencing, with the combined splice site score of the corresponding exons. Genome-wide exon inclusion levels were calculated from HeLa cell mRNA sequencing reads that either define exon inclusion or exon exclusion. As was observed for the EST database analysis, only a small fraction of alternative exon inclusion resides within the intermediate range of 20–80% inclusion and the majority of reads represent preferential exon inclusion or exclusion events (Figure 5A). The correlation between

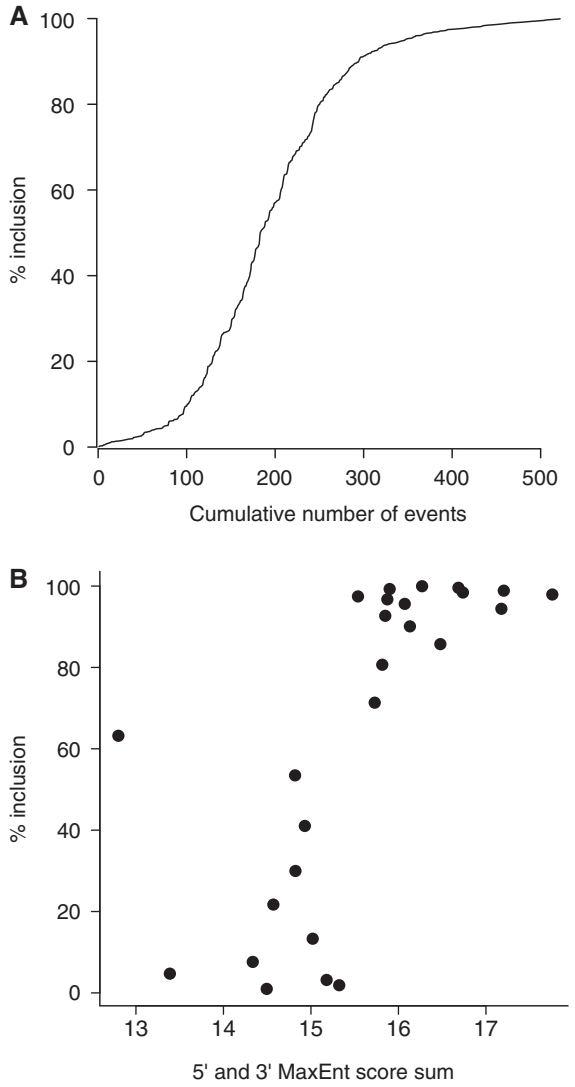

Figure 5. Exon inclusion levels determined from high-throughput sequencing of HeLa cells. (**A**) High-throughput sequence reads were used to calculate the levels of internal exon inclusion. The plot shows the relationship between exon inclusion levels and the cumulative number of events. (**B**) Exon inclusion levels are correlated with the average strength of the combined 5′ and 3′ splice site scores for groups of 20 alternative splicing events sorted by inclusion levels.

exon inclusion levels and combined splice site score also displays a sigmoidal shape with a narrow transition between combined scores of MaxEnt 15 and 16 (Figure 5B). We conclude that the strength of the combined splice site score of alternatively spliced exons tracks with exon inclusion levels.

To evaluate splice site score separation, we compared the distribution of 3′ splice site strengths of constitutively spliced exons, preferentially included alternative exons and preferentially excluded alternative exons. We found that on average, preferentially included alternative exons have stronger 5′ splice sites compared to preferentially excluded alternative exons, as expected from previous work (19) (Figure 6A; $P < 1.3\,e^{-69}$). We then compared the distribution of 3′ splice site strengths of constitutively spliced exons and preferentially excluded alternative exons and found that on average, preferentially included alternative exons have stronger 3′ splice sites (Figure 6B; $P < 1\,e^{-84}$). In agreement with our experimental analysis, a significantly better classifier for alternative exons can be obtained when evaluating the sum of splice site scores between these groups (Figure 6C; $P < 3.1\,e^{-163}$). Interestingly, no significant difference was observed between constitutively spliced exons and preferentially included alternative exons (Supplementary Figure S4; $P < 0.47$), indicating that preferentially included exons behave similar to constitutive exons. The class separation improvement shown in Figure 6 is also reflected by ROC curve analyses that demonstrate that the sum of the splice site scores is a better predictor of alternative exon inclusion than either the 5′ or the 3′ splice site alone (Figure 7). To evaluate the hypothesis that one splice site might be more important for exon inclusion, we weighted the individual contribution of the 3′ and 5′ splice site score. To accomplish this, we used the Linear Discriminant Algorithm (LDA), which weighs the relative contribution of splice sites to optimize separation between the two classes of exons. While a linear discriminant of 0.49 for the 5′ splice site and 0.51 for the 3′ splice site was obtained, a ROC curve analysis indicates that the weighted contributions do not significantly improve exon inclusion predictions (data not shown). We conclude that splice sites display near equal contributions towards exon recognition and that the combined splice site score of internal exons is a major discriminant between alternatively and constitutively spliced exons.

DISCUSSION

The majority of genes in the human genome contain exons that are exon defined (7). Using a 3-exon minigene whose architecture fits the exon definition mode of splice site recognition, we examined the contribution of the 3′ and 5′ splice site strengths on internal exon inclusion. We have shown that when the 3′ or 5′ splice site is strong, most internal exons are efficiently recognized. However, even if one splice site is strong internal exons can be skipped when the other splice site at the opposite end of the exon is very weak. When the 3′ or 5′ splice site of an internal exon is of intermediate strength, a strong compensating splice site at the opposite end of the exon is required to support spliceosomal recognition. Using a Pearson correlation to measure the interdependence of the splice site strength and inclusion levels of an internal exon, we found that there is a much greater correlation between exon inclusion levels and the combined splice site score than between exon inclusion levels and the 3′ or 5′ splice site score alone. The concept of combined splice site strength as a strong discriminant between alternatively and constitutively spliced exons was further evaluated by performing genome-wide analyses. In agreement with previous work, the splice site strength of either the 3′ splice site or the 5′ splice site was significantly stronger in the constitutively spliced population of exons (19). However, when considering the combined 3′ and 5′ splice site scores an increased

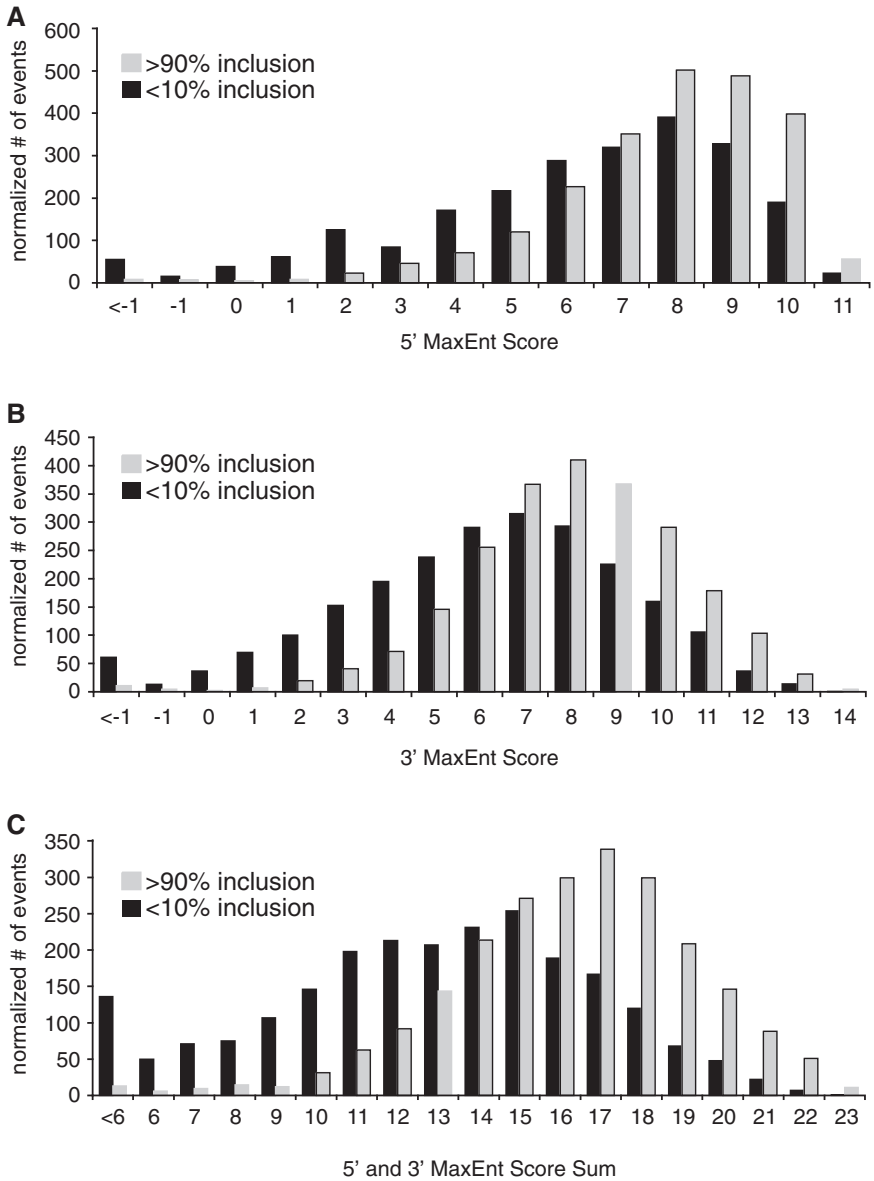

Figure 6. Computational analysis of splice site strength and exon inclusion. Splice site strength distributions (MaxEnt score) were derived for sets of alternatively spliced exons. Splice site scores were binned in MaxEnt scoring bins that increase by one. The *x*-axis displays increasing splice site strength from left to right. The *y*-axis represents the number of events. (**A**) 5′ splice site score distribution of alternative exons with >90% inclusion levels (light gray) and alternative exons with <10% inclusion levels (dark gray). A two-tailed *t*-test was performed to show the distribution profiles are significantly different, $P < 1.3\,e^{-69}$. (**B**) 3′ splice site score distribution of <10 and >90% inclusion exons ($P < 1\,e^{-84}$). (**C**) 5′ and 3′ splice site score sum distribution of <10 and >90% inclusion exons ($P < 3.1\,e^{-163}$).

separation of constitutive and alternative exons was accomplished. Recent work demonstrated that the application of multiple integrated splicing features, including 3′ and 5′ splice site scores, can significantly increase the discriminating between constitutive and alternative exons (20). Such a combinatorial considerations have also been demonstrated to be beneficial to differentiate between real and pseudo exons (21).

To determine the relative contributions of the 3′ and 5′ splice sites we used a Linear Discriminant Analysis model that finds the optimal weight for each variable to reliably predict class membership. The LDA approach showed that the contribution each splice site makes towards efficient exon recognition is close to equal, with the 3′ splice site contributing 51% and the 5′ splice site contributing 49%. Thus, neither the 5′ nor the 3′ splice site signal truly dominates in mediating spliceosomal recognition. The fact that weaker splice sites can be rescued by strong compensating splice sites at the opposite end of the exon further suggests that 5′ and 3′ splice site recognition in the exon definition mode occurs within the same rate-limiting step (22). In agreement with recent proposals, our results highlight the importance to evaluate the inclusion or exclusion fate of internal exons as a function of multiple

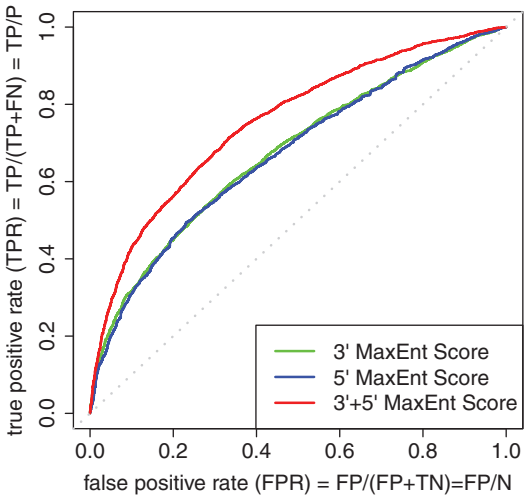

Figure 7. A ROC curve was used to determine the ability of the 5′ or 3′ splice site strength or their additive combination to distinguish between high and low inclusion exons. All exons with an EST coverage of at least 10 EST entries were used for analysis.

splicing elements (10). Indeed, improved splicing predictions can be achieved when splicing regulator signals are combined with exon/intron architectural specifications (20).

In our mini-gene assays, it was easy to differentiate the constitutively spliced exons from the excluded exons based on their combined splice site score. However, when the splice site strength distributions were compared between preferentially included and preferentially excluded exons, a sizable overlap between the two populations remained (Figure 6C). This extensive overlap of near constitutive and alternative exons with splice sites scores of similar strengths is likely due to the fact that exons have evolved to depend on multiple parameters in addition to the 3′ and 5′ splice site strength, such as the presence or absence of splicing regulators (23), RNA secondary structures (24–26), nucleosome enrichment and histone modification (27), the exon/intron architecture (6) and the process of pre-mRNA synthesis itself (28). Supporting this notion is the observation that many strong 3′ and 5′ splice sites separated by <250 nt exist within introns, yet they are not recognized as true exons (29). Evidently, real exons require the additional contribution of other splicing elements and these elements can compensate for weak splice sites in constitutive exons. The inclusion level of each individual exon is then determined by its own unique suite of recognition elements, also referred to as the splicing code (20). While the combined strength of the 3′ and 5′ splice sites is a fundamental aspect of efficient exon recognition, it must be considered in the context of the exon and its repertoire of recognition elements. This conclusion is underscored when considering that the majority of alternative exons displays a combined MaxEnt splice site score of >12 (Figure 6C), implying that many exons are alternatively spliced even though their splice sites are of sufficient strength to support splicing (18). This observation suggests that alternative splicing may be largely under the control of splicing repressor elements. Alternatively, it is possible that the MaxEnt scoring method, which is based on comparing used splice sites with pseudo splice sites, biases towards higher scores. Regardless of interpretation, the data demonstrate that the splice sites of alternatively spliced exons are much closer in sequence content to actual splice sites than background sequences.

When analyzing mini-genes or genome-wide data, we made the striking observation that the transition between preferential inclusion and preferential exclusion is quite narrow and steep (Figures 2–5). These results show that even minor differences in the combined splice site strength can trigger drastically different inclusion ratios and suggest that efficient exon recognition is realized by the spliceosome after reaching a particular threshold. As discussed above, reaching this threshold can be accomplished via several RNA splicing elements. Regardless of the mechanisms that increase exon recognition, it appears that the splicing machinery has evolved to preferentially maintain either high or low levels of exon inclusion. An advantage of such threshold kinetics is that it permits significant changes in alternative splicing upon minor alterations in the cellular environment. However, this increased sensitivity of promoting alternative splicing also comes at the cost of inducing splicing defects upon acquiring single base mutations (4).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are grateful to the Hertel laboratory for helpful comments on the manuscript. HeLa cells were obtained from the National Cell Culture Center (Minneapolis, MN).

FUNDING

National Institute of Health (grants GM 62287 and American Recovery and Reinvestment Act GM 62287S1 to K.J.H., and T15LM007443 to P.J.S.); Postdoc-Program of the German Academic Exchange Service (DAAD) (fellowship to A.B.). Funding for open access charge: National Institutes of Health (grants GM 62287).

Conflict of interest statement. None declared.

REFERENCES

1. Fox-Walsh,K.L. and Hertel,K.J. (2009) Splice-site pairing is an intrinsically high fidelity process. *Proc. Natl Acad. Sci. USA*, **106**, 1766–1771.
2. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
3. 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.
4. Cooper,T.A., Wan,L. and Dreyfuss,G. (2009) RNA and disease. *Cell*, **136**, 777–793.
5. Stamm,S., Riethoven,J.J., Le Texier,V., Gopalakrishnan,C., Kumanduri,V., Tang,Y., Barbosa-Morais,N.L. and Thanaraj,T.A. (2006) ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res.*, **34**, D46–D55.
6. Fox-Walsh,K.L., Dou,Y., Lam,B.J., Hung,S.P., Baldi,P.F. and Hertel,K.J. (2005) The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proc. Natl Acad. Sci. USA*, **102**, 16176–16181.
7. Sterner,D.A., Carlo,T. and Berget,S.M. (1996) Architectural limits on split genes. *Proc. Natl Acad. Sci. USA*, **93**, 15081–15085.
8. Berget,S.M. (1995) Exon recognition in vertebrate splicing. *J. Biol. Chem.*, **270**, 2411–2414.
9. Dominski,Z. and Kole,R. (1991) Selection of splice sites in pre-mRNAs with short internal exons. *Mol. Cell Biol.*, **11**, 6075–6083.
10. Hertel,K.J. (2008) Combinatorial control of exon recognition. *J. Biol. Chem.*, **283**, 1211–1215.
11. Lear,A.L., Eperon,L.P., Wheatley,I.M. and Eperon,I.C. (1990) Hierarchy for 5′ splice site preference determined in vivo. *J. Mol. Biol.*, **211**, 103–115.
12. Burge,C.B., Tuschl,T. and Sharp,P.A. (1999) In Gesteland,R.F.C.T.R. and Atkins,J.F. (eds), *The RNA World*, 2nd edn. CSHL Press, Cold Spring Harbor, NY, pp. 525–560.
13. Michaud,S. and Reed,R. (1991) An ATP-independent complex commits pre-mRNA to the mammalian spliceosome assembly pathway. *Genes Dev.*, **5**, 2534–2546.
14. Hicks,M.J., Mueller,W.F., Shepard,P.J. and Hertel,K.J. (2010) Competing upstream 5′ splice sites enhance the rate of proximal splicing. *Mol. Cell Biol.*, **30**, 1878–1886.
15. Li,C. and Evans,R.M. (1997) Ligation independent cloning irrespective of restriction site compatibility. *Nucleic Acids Res.*, **25**, 4165–4166.
16. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
17. Rhead,B., Karolchik,D., Kuhn,R.M., Hinrichs,A.S., Zweig,A.S., Fujita,P.A., Diekhans,M., Smith,K.E., Rosenbloom,K.R., Raney,B.J. *et al.* (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.*, **38**, D613–D619.
18. Yeo,G. and Burge,C.B. (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.*, **11**, 377–394.
19. Clark,F. and Thanaraj,T.A. (2002) Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum. Mol. Genet.*, **11**, 451–464.
20. Barash,Y., Calarco,J.A., Gao,W., Pan,Q., Wang,X., Shai,O., Blencowe,B.J. and Frey,B.J. (2010) Deciphering the splicing code. *Nature*, **465**, 53–59.
21. Zhang,X.H., Heller,K.A., Hefter,I., Leslie,C.S. and Chasin,L.A. (2003) Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. *Genome Res.*, **13**, 2637–2650.
22. Lam,B.J. and Hertel,K.J. (2002) A general role for splicing enhancers in exon definition. *RNA*, **8**, 1233–1241.
23. Black,D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.*, **72**, 291–336.
24. Shepard,P.J. and Hertel,K.J. (2008) Conserved RNA secondary structures promote alternative splicing. *RNA*, **14**, 1463–1469.
25. Hiller,M., Zhang,Z., Backofen,R. and Stamm,S. (2007) Pre-mRNA Secondary Structures Influence Exon Recognition. *PLoS Genet.*, **3**, e204.
26. Eperon,L.P., Graham,I.R., Griffiths,A.D. and Eperon,I.C. (1988) Effects of RNA secondary structure on alternative splicing of pre-mRNA: is folding limited to a region behind the transcribing RNA polymerase? *Cell*, **54**, 393–401.
27. Spies,N., Nielsen,C.B., Padgett,R.A. and Burge,C.B. (2009) Biased chromatin signatures around polyadenylation sites and exons. *Mol. Cell*, **36**, 245–254.
28. Kornblihtt,A.R. (2006) Chromatin, transcript elongation and alternative splicing. *Nat. Struct. Mol. Biol.*, **13**, 5–7.
29. Chasin,L.A. (2007) Searching for splicing motifs. *Adv. Exp. Med. Biol.*, **623**, 85–106.
30. Shapiro,M.B. and Senapathy,P. (1987) RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.*, **15**, 7155–7174.