Intrinsic differences between authentic and cryptic 5' splice sites

Krainer, Adrian R.; Sachidanandam, Ravi; Roca, Xavier

2003

Roca, X., Sachidanandam R., & Krainer A. R. (2003). Intrinsic differences between authentic and cryptic 5' splice sites. Nucleic Acids Research, 31(21), 6321-6333.

https://hdl.handle.net/10356/95535

https://doi.org/10.1093/nar/gkg830

Intrinsic differences between authentic and cryptic 5′ splice sites

Xavier Roca, Ravi Sachidanandam and Adrian R. Krainer*

Cold Spring Harbor Laboratory, PO Box 100, Cold Spring Harbor, NY 11724, USA

**To whom correspondence should be addressed. Tel: +1 516 367 8417; Fax: +1 516 67 8815; Email: krainer@cshl.edu*

ABSTRACT

Cryptic splice sites are used only when use of a natural splice site is disrupted by mutation. To determine the features that distinguish authentic from cryptic 5' splice sites (5'ss), we systematically analyzed a set of 76 cryptic 5'ss derived from 46 human genes. These cryptic 5'ss have a similar frequency distribution in exons and introns, and are usually located close to the authentic 5'ss. Statistical analysis of the strengths of the 5'ss using the Shapiro and Senapathy matrix revealed that authentic 5'ss have significantly higher score values than cryptic 5'ss, which in turn have higher values than the mutant ones. β-Globin provides an interesting exception to this rule, so we chose it for detailed experimental analysis *in vitro*. We found that the sequences of the β-globin authentic and cryptic 5'ss, but not their surrounding context, determine the correct 5'ss choice, although their respective scores do not reflect this functional difference. Our analysis provides a statistical basis to explain the competitive advantage of authentic over cryptic 5'ss in most cases, and should facilitate the development of tools to reliably predict the effect of disease-associated 5'ss-disrupting mutations at the mRNA level.

INTRODUCTION

Accurate splicing of pre-mRNA is a critical step in the gene expression pathway in eukaryotes. The exon–intron boundaries, known as the 5' and 3' splice sites, are

defined by conserved sequences that are critical for the reaction. Because the splice-site consensus motifs are degenerate, many matches to each consensus are present along pre-mRNAs, but the vast majority of these sequences, known as pseudo splice sites, are never selected for splicing (1). Cryptic splice sites also match the consensus motifs, and by definition they are splice sites that are not detectably used in wild-type pre-mRNA, but are only selected as a result of a mutation elsewhere in the gene, most often at the authentic splice site.

The 5' splice site (5'ss) consensus sequence in higher eukaryotes comprises nine partially conserved nucleotides at the exon−intron boundary, and corresponds to nearly perfect Watson−Crick base pairing to the U1 snRNA 5' terminus (2). There is strong genetic and biochemical evidence for the critical contribution of this base pairing to selection of the 5'ss (3−5). However, U1 snRNA appears to be dispensable for splicing of some pre-mRNAs; in such cases, U6 snRNA and SR proteins can make up for the absence of functional U1 snRNA in depleted extracts (6−8). *In vitro* selection of functional 5'ss sequences surprisingly yielded the same consensus sequence in splicing reactions containing either wild-type or 5'-end-deleted human U1 snRNAs (9). Consistent with this finding, yeast U1 snRNP can still bind to a 5'ss in the absence of the 5' end of U1 snRNA (10), and it has been proposed that the U1C polypeptide is responsible for this interaction (11). In addition, U1 snRNP binding to a 5'ss is not always followed by splicing at that site, as several U1 particles can simultaneously base pair to competing 5'ss, even when only one of the sites is selected for splicing (12). After the initial recognition of the 5'ss by U1, U5 and U6 snRNPs also bind at or around the 5'ss. U6 snRNA base pairs to the 5'ss in a mutually exclusive manner with U1 snRNA (13), and likely participates in splicing catalysis. However, U1 and U6 snRNAs can bind to adjacent, non-overlapping sequences of a pre-mRNA, and in this case the actual site of transesterification is defined by U6 snRNA (14,15).

Insights into the mechanisms of cryptic splice-site activation have been obtained experimentally in several systems. Thalassemia-associated mutations of the 5'ss of

intron 1 of the human β-globin gene, *Hbb*, activate the use of three cryptic 5'ss, upstream and downstream of the natural site (16). Coimmunoprecipitation of wild-type and mutant β-globin substrates with U1 snRNP antibodies, combined with RNase T1 digestion, revealed differences between these sequences in their apparent affinities for U1 (17). Furthermore, mutations in a β-globin exon 1 cryptic 5'ss at position −38 that enhance complementarity to U1 snRNA activate this splice site in the presence of a wild-type authentic 5'ss (18). Increasing the levels of SR proteins or hnRNP A/B proteins affects the relative use of each of the three cryptic splice sites in mutant substrates, both in vitro and in transfected cells (19−21). However, in no case was a shift in splice-site usage seen with the wild-type β-globin pre-mRNA, indicating that an excess of these splicing factors does not abrogate the distinction between authentic and cryptic 5'ss. Experiments with an adenovirus pre-mRNA showed that the use of a cryptic 5'ss depends on secondary structure in the upstream exon, and that the kinetics of splicing via the cryptic site is slower than that via the authentic site (22). A genetic screen in *Caenorhabditis elegans* unveiled a dominant, allele-specific suppressor mutation in *sup-39* that affects the choice among two cryptic 5'ss and a mutant 5'ss in *unc-73(e936)*, such that use of the mutant 5'ss is favored (23,24); the sup-39 gene remains to be identified, so it is not yet known whether it encodes a protein or an RNA. U snRNA complementation experiments in *Saccharomyces cerevisiae* revealed that both U5 (25) and U6 snRNAs (26) are involved in cryptic 5'ss activation in yeast, and recent experiments in *Schizosaccharomyces pombe* also showed that mutant versions of U1 can activate cryptic 5'ss (27). Finally, Eperon and colleagues carried out a competition analysis between different 5'ss sequences, and found that the three cryptic 5'ss they analyzed competed poorly in relation to authentic and alternative 5'ss (28).

Point mutations resulting in splicing defects account for at least 15% (29) and in some cases as many as 50% of known alleles in human-disease genes (30,31). In higher eukaryotes, mutations that disrupt a 5'ss usually cause skipping of the exon that precedes it (32). The second most frequent consequence of such mutations is

cryptic splice-site activation, whereas intron retention is very rare. Most reported splicing mutations in mutation databases have been described only at the genomic sequence level, and their effect at the protein level cannot be accurately predicted because it depends on which of these three splicing pathways are followed as a consequence of the mutation.

The latest compilation that includes cryptic 5'ss is the Aberrant Splicing Database, published in 1994, with 28 cryptic 5'ss and 15 cryptic 3'ss from different mammalian species (32). The aim of our study was to determine the general nature of the differences between authentic 5'ss and the corresponding cryptic 5'ss that are activated upon mutation. We addressed this question by compiling available examples of cryptic 5'ss in human genes, building a database, and analyzing the various splice sites by statistical methods. One particular case, derived from β-globin, was chosen for experimental analysis of the mechanism of cryptic 5'ss repression and activation. We show that, as a general rule, the extreme difference in splicing efficiency between competing authentic and cryptic 5'ss is determined by the sequences of the 9 nt 5'ss motifs.

MATERIALS AND METHODS

Construction of the database

Published examples of cryptic 5'ss activation were identified by searching PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi). We restricted the search to human genes, to avoid potential species-specific variability. We only included experimentally verified cryptic 5'ss. A few examples of cryptic 5'ss activation were omitted from the compilation because the genomic DNA sequence could not be located in the most recent release of the HGP database. The accessions were mapped to the genome using software we developed. Automatic corrections were made to ensure that the exon–intron boundaries were in the appropriate locations. Manual curation was used to improve the software and check the mapping results.

The cryptic 5'ss were then mapped to the sequences using information from the references.

We also collected human alternative 5'ss extracted from the ASAP database (33). We chose only alternative 5'ss that were labeled with high confidence as tissue specific. To ensure that the pool included only 5'ss that result from true alternative splicing events, we used the ASAP tools to select the alternative 5'ss supported by mRNA sequences, or by a minimum of seven ESTs, or two or more ESTs with different splicing patterns. 230 alternative 5'ss in ASAP met all these criteria (see the online supplement).

Statistical tests

To measure the strengths of the various 5'ss, we used the Shapiro and Senapathy (S&S) consensus matrix, which reflects the degree of conservation in different positions resulting from the alignment of 1446 5'ss (34,35). The consensus 5'ss sequence is MAG|GURAGU (M = A or C; R = purine), and spans from position –3 (the third nucleotide from the 3' end of the upstream exon) to +6 (the sixth nucleotide in the intron). Although position –3 is often ignored in these matrices, we took it into account because of the significant preference for C or A at this position.

The S&S scores for the different kinds of 5'ss (authentic, cryptic, mutant, pseudo and alternative) are not distributed normally around their means. Therefore, we used non-parametric tests to assess the significance of the score values for the five classes of 5'ss. To test the significance of deviation from zero of the difference between pairs of scores, we used the Wilcoxon signed rank test; to test the significance of the deviation of the overall means from each other for the different classes of 5'ss, we used the Mann–Whitney rank test. The pairwise analysis involves a case-by-case evaluation of the intrinsic differences between corresponding 5'ss associated with the same exon, and reflects the natural context in which splice-site selection occurs.

To calculate the free-energy parameters for the stability of the RNA duplexes between the various 5'ss sequences and the U1 snRNA 5' terminus, we used the Turner energy rules as described in http://www.bioinfo.rpi.edu/–zukerm/rna/energy/ (11/3/2000 update). These rules are based on measurements with synthetic oligoribonucleotides and reflect the contribution of hydrogen bonding, base stacking, mismatches and Watson–Crick or G–U base pairs (36).

Cloning procedures

All β-globin substrates were inserted into a pcDNA3.1+ plasmid (Invitrogen, Carlsbad, CA), and all bear a mutation of the cryptic 5'ss at position +13—a T to C transition at +14—that inactivates this 5'ss. For the mutants in Figure 4B, we used site-directed mutagenesis with oligonucleotides that carry the different mutations. For each pair of primers, 16 cycles of PCR with Pfu I Turbo (Stratagene) were performed. PCR products were digested with DpnI (New England Biolabs, Beverly, MA), followed by transformation of competent *Escherichia coli* DH5a. The following mutations were introduced: (i) a mutation of the authentic 5'ss at +1: CAG/AACCCG; (ii) a duplication of the authentic 5'ss CAG/GTTGGT at positions –16 and +1; (iii) a duplication of the cryptic 5'ss at –16 GTG/ GTGAGG at positions –16 and +1; and (iv) a swap of sequences between the 5'ss at –16 and +1. All mutant constructs were verified by sequencing on an ABI3700 Automated Sequencer.

In vitro splicing experiments

Human β-globin templates were made by PCR using the following primers: β-5-T7, 5'-AATTTAATACGACTCACTATAGGCTTACATTTGCTTCTG-3'; and R-Ex2-Bam-txn, 5'-GATCCACGTGCAGCTTGTCACAGTG-3'. The products were purified with a PCR-purification kit (Qiagen, Valencia, CA). Capped, ^{32}P-labeled pre-mRNA substrates were transcribed from purified PCR

products with T7 RNA polymerase (Promega, Madison, WI) and gel purified (37).

HeLa cell nuclear extract preparation and splicing reactions were carried out as described (37,38). For *in vitro* splicing reactions, 20 fmol of ^{32}P-labeled, $^{7}CH_3$-GpppG-capped T7 transcripts were incubated in 12.5 µl splicing reactions with 30% nuclear extract and 3.2 mM $MgCl_2$, for 3 h at 30 ºC. Samples were analyzed by electrophoresis on 5.5% polyacrylamide/7 M urea gels. Gels were exposed overnight onto X-OMAT films (Kodak).

RESULTS

A compilation of human cryptic 5'ss

Table 1 shows all the cases of cryptic splicing in human genes included in the present analysis. Cryptic 5'ss are those that are only used when a mutation disrupts use of the authentic splice site. Note that cryptic splice sites differ from splice sites that are created *de novo* by a mutation in an exon or intron that increases the match to a splice-site consensus. However, use of such a created splice site can frequently occur in conjunction with activation of a cryptic splice site, such that, e.g. a created 5'ss and an activated cryptic 3' splice site upstream together define a new exon. Figure 1 shows an example of cryptic 5'ss activation in the porphobilinogen deaminase (*PBGD*) gene (39), associated with acute intermittent porphyria. In this case, a mutation in the last nucleotide of exon 10 activates a cryptic 5'ss (C) 9 nt upstream of the mutant 5'ss (M), whereas other good matches to the 5'ss consensus around this area are never used, so they are considered pseudo 5'ss (P). We found 76 reported examples of cryptic 5'ss in 46 human genes, and this sample size enables meaningful statistical tests to be conducted. Mutations that activate cryptic 5'ss mapped to the authentic 5'ss sequence, with one exception in the *COL7A1* gene (40), in which a 28 bp genomic deletion in the middle of exon 73 activates use of a cryptic 5'ss 10 nt downstream of the deletion. In all but two cases (41,42), the mutation is associated with a genetic disease, i.e. hereditary syndromes and/or cancer.

General features of cryptic 5'ss

Table 2 summarizes the general features of the splice sites we analyzed. The cryptic 5'ss were found with equal frequencies in exons and introns; they were found both in first exons (seven out of 61 exons) and in internal exons. The cryptic 5'ss did not show a bias towards a particular reading-frame phase, relative to that established by the position of the authentic 5'ss (Table 2). Forty two percent of the cryptic 5'ss were in the same reading frame as the authentic 5'ss, i.e. the distance between the authentic and cryptic splice sites was a multiple of three nucleotides; 32% were shifted by +1 nucleotide and 26% were shifted by +2 nucleotides. These frequencies are not appreciably different from a random distribution.

The observed distribution of distances between authentic and cryptic 5'ss resembles a normal distribution with its peak centered at position +11 (Fig. 2a). The observed absolute average distance (63 nt) and its large standard deviation (SD) (64 nt) correspond to a broad-range distribution, spanning from –141 in the exon to +398 in the intron (Table 2). This distribution indicates that cryptic 5'ss can in a few cases be located very far from the original site, even when the size of the resulting exon becomes sub-optimally small or large. The distribution is not completely symmetrical because the density of cryptic 5'ss located in the upstream exon is slightly higher around the authentic site than that of the cryptic 5'ss in the downstream intron (Fig. 2b).

In general, the length of the abnormal exons that result from cryptic 5'ss usage is within the range for most internal exons [between 50 and 250 nt (43)]: out of 54 internal exons (average length 159 nt) whose 5'ss disruption leads to cryptic 5'ss activation, eight (15%) are beyond these limits; out of 64 internal exons (average length 174 nt) that result from cryptic 5'ss activation, 14 (22%) are beyond these length constraints (Table 2).

Recently, it was reported that in-frame stop codons located between an authentic 5'ss and a downstream latent or alternative 5'ss favor the use of the upstream 5'ss (44). In contrast, a subsequent study found no enrichment of in-frame stop codons in pseudo-exons (intronic sequences flanked by strong splice sites), or in intronic regions immediately downstream of exons (45). We analyzed the presence of in-frame stop codons between authentic and cryptic 5'ss, when the latter are located within the downstream intron. Of 39 intronic cryptic 5'ss in 27 genes, only half (19 cryptic 5'ss) were preceded by in-frame stop codons. The numbers do not enable statistical tests, but there does not appear to be an enrichment of stop codons between authentic and cryptic 5'ss in this data set that would account for the correct specification of the upstream authentic sites in the wild-type pre-mRNAs.

Analysis of the consensus value of the 5'ss by the Shapiro and Senapathy matrix

Table 3 shows a comparison of the S&S consensus values (34,35) for the five types of 5'ss: authentic, cryptic, mutant, pseudo and alternative 5'ss. The mutant category refers to mutated versions of an authentic 5'ss, which affect the efficiency of splicing at that site. To construct a set of reference pseudo 5'ss relevant for this analysis, we chose all the sequences in the Table 1 gene set with S&S scores above a threshold arbitrarily set at 60 (which detects >90% of the cryptic 5'ss), and located closer to the authentic 5'ss—on either side of it—than the cryptic 5'ss. Figure 3 displays the average and SD of the S&S scores for each category of 5'ss (Fig. 3A), and the differences in the scores for each pair of 5'ss that occurs in the same pre-mRNA (Fig. 3B). The average S&S score of cryptic 5'ss located in exons is very similar to the average score of the cryptic 5'ss in introns (71.01 versus 73.69).

By comparing the authentic and mutant groups, it is clear that the differences between the wild-type 5'ss and the mutant versions are very significant, both in comparisons of the sets of sequences and of pairs of sequences in the same gene, as seen in previous studies (29,30). Indeed, in nearly all cases, the mutation of the 5'ss decreases the S&S score. There is only one reported exception, in the *COL7A1* gene,

in which an A to G transition at position -2 of the 5'ss of intron 3 improved the S&S score, but nevertheless reduced splicing at that site (46).

Likewise, the group and pairwise comparisons between the cryptic and mutant 5'ss gave consistent results. The S&S scores of the cryptic 5'ss are significantly higher than those of the mutant 5'ss. This higher score value is consistent with the activation of the cryptic 5'ss in the context of mutant pre-mRNAs. It should be noted that not all these mutations completely abolish splicing at the authentic 5'ss, but all do reduce the efficiency of the reaction at that site.

Most interestingly, both the group and pairwise comparisons of the S&S scores between authentic and cryptic 5'ss showed that the latter values are significantly lower. Thus, as a general rule, intrinsic differences in the 9 nt consensus 5'ss sequence can account for the cryptic 5'ss remaining completely silent in the presence of a wild-type authentic site.

For comparison, we also included a set of 230 alternative 5'ss extracted from the ASAP database (33). The average S&S score for this alternative 5'ss data set is 78.26 (Fig. 3 and Table 3), i.e. between the averages of authentic and cryptic 5'ss, but closer to the former. Statistical comparisons between the S&S scores of alternative and those of cryptic, mutant and pseudo 5'ss show that the alternative 5'ss have significantly higher scores than those of the latter three categories (Table 3 and data not shown). The S&S scores of authentic 5'ss, in turn, are also significantly higher than those of alternative 5'ss, although the difference of averages of the two kinds of scores is only 4.7.

Next, we calculated the potential matches to the 5'ss consensus that have an S&S score above a threshold of 60, and that are located closest to a given authentic 5'ss. This threshold was arbitrarily chosen instead of the lowest score of the cryptic 5'ss (49.73), to exclude many potential 5'ss with very low scores. These potential 5'ss were compared with the total pool of cryptic 5'ss. Although five of the cryptic 5'ss have an

S&S score lower than 60, eliminating them from the analysis did not significantly alter the results (data not shown). Of 61 different exons in which one or more cryptic 5'ss are activated when the authentic splice site is disrupted (Table 1), in 35 cases the closest potential site coincides with the cryptic 5'ss, or with one of several cryptic 5'ss. This observation suggests that the cryptic 5'ss(s) activated when a mutation inactivates or weakens an authentic 5'ss is usually the nearest sequence with a good match to the S&S consensus. However, in seven of the 35 cases, other cryptic 5'ss located further in the exon or in the intron are also activated.

The remaining 26 cases, in which the closest potential 5'ss does not coincide with the cryptic 5'ss, were considered pseudo 5'ss. To increase the number of samples, we also included the second spatially closest match to the S&S consensus that is not also a cryptic 5'ss. The S&S scores of these pseudo sites are significantly lower than the scores of the authentic sites. According to the above hypothetical rule, a pseudo site located between an authentic site and a cryptic site should have a lower S&S score than the cryptic site. However, although the mean value for the cryptic sites is higher than that for the pseudo sites, the difference is only marginally significant. The pairwise comparison failed to show statistically significant differences. This result argues against the simple model that, given a genomic sequence, one can reliably predict which potential 5'ss will act as a cryptic 5'ss when the authentic site is mutated.

We also calculated the strength of base pairing between the various categories of 5'ss and the U1 snRNA 5' terminus (see Supplementary Material). Although the differences observed using the S&S matrix were also seen using the calculated base-pairing measure, there were more exceptions in the comparison between authentic and cryptic 5'ss: 21 cryptic 5'ss (28%) had a higher predicted U1 base-pairing stability than their corresponding authentic 5'ss. In only nine of these cases (and four others), the S&S score of the cryptic 5'ss was higher than that of the corresponding authentic 5'ss. This comparison suggests that discrimination between authentic and cryptic 5'ss might be explained by differential duplex stability with U1 snRNA in some of the samples, but not in all.

The strengths of the authentic and cryptic 5'ss were also calculated by four other methods (Table 4): the neural network (NN) method (47) (http://www.fruitfly.org/seq_tools/splice.html), and the maximum entropy (MAXENT), maximum dependence decomposition (MDD) and first-order Markov (MM) models (48) (http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html). The number of exceptions to the competitive advantage of authentic versus cryptic 5'ss varied among them, with NN and MDD being the best discriminators between these 5'ss (six exceptions for each method), and the free-energy parameter model being the least accurate (21 exceptions). Of the 76 cryptic 5'ss, only three gave higher scores than their corresponding authentic 5'ss using all six methods, whereas 47 cryptic 5'ss had lower scores than their authentic 5'ss according to all six methods.

β-Globin as an atypical model of cryptic splice-site activation

Whereas most of the cryptic 5'ss have significantly lower S&S consensus values than the authentic sites (63 out of 76 cases), the β-globin cryptic 5'ss at position −16 of exon 1, which is activated by several β-thalassemia mutations (16), has a slightly higher score than the authentic 5'ss (81.84 versus 80.10, Fig. 4A). Moreover, the distance between the two sites is only 16 nt, which is much smaller than the absolute average distance for the whole set of cryptic 5'ss (63 nt, Table 2). We chose this example for experimental analysis to try to understand how the authentic and −16 cryptic 5'ss are distinguished. We used *in vitro* splicing reactions with a panel of mutant β-globin substrates and measured the relative use of the authentic (+1) and cryptic (−16) 5'ss.

Previous studies suggested that the cryptic 5'ss at −16 is potentially defective in U1 snRNA binding (17). Other studies showed that the use of certain 5'ss depends on U1 binding to a downstream sequence, which in turn facilitates binding of U6 to the actual 5'ss (15). To test whether splicing via the β-globin −16 cryptic 5'ss

depends on U1 binding to the mutant 5'ss downstream, we mutated the +1 authentic 5'ss to a sequence whose intronic portion completely lacks complementarity to the U1 snRNA 5' terminus (authentic 5'ss changed from CAG/GTTGGT to CAG/AACCCG). Splicing of this pre-mRNA in HeLa cell nuclear extract proceeded through use of the -16 cryptic 5'ss (Fig. 4B, lane 1). However, this site was used much less efficiently than in the β-thalassemia IVS1-G1A mutant (Fig. 4A and B, lane 6). We conclude that the −16 5'ss can be recognized independently of the natural 5'ss, although a weak—or an inactive but still recognizable—5'ss at +1 can enhance the use of the −16 site.

To address possible effects of the different sequence contexts surrounding the −16 and +1 5'ss, we designed the following three mutants: a duplication of the 9 nt sequence of the +1 5'ss, at positions −16 and +1; a duplication of the −16 5'ss at the same positions; and a swap of the sequences between the −16 and +1 5'ss. The first mutant substrate spliced via both +1 sequences at the two different positions, showing no preference for the sequence context at position +1 (Fig. 4B, lane 2). The second mutant substrate spliced only via the distal 5'ss, suggesting that under these conditions, the context surrounding the upstream site and/or the position itself provides a competitive advantage (Fig. 4B, lane 3). The swap mutant spliced only via the natural 5'ss sequence placed at the −16 position, confirming that the 9 nt sequence at +1 is necessary and sufficient to explain the splice-site selection specificity of this substrate (Fig. 4B, lane 4). Strikingly, the overall splicing efficiencies of these three β-globin mutants are strongly reduced (>10-fold), compared with the efficiencies of the wild-type and IVS1-G1A mutant β-globin substrates (Fig. 4B, compare lanes 2−4 with lanes 5 and 6).

We conclude that the intrinsic sequence differences between the splice site sequences at −16 and +1 are sufficient to explain why the +1 site is normally selected, even though these differences are not reflected in their respective S&S

consensus values.

DISCUSSION

Certain aspects of the molecular mechanisms of cryptic splice-site activation have been previously outlined (17,18,22), and it was previously shown in some individual cases, that the cryptic 5'ss had lower scores than the authentic 5'ss (28,30,49−54). However, the general basis for the different splicing efficiencies between the authentic and cryptic 5'ss has not been determined because the small number of samples analyzed previously did not allow the use of statistical tests. Analysis of the present compilation of 76 cryptic splice sites in 46 human genes provides statistical evidence consistent with the fact that this category of splice sites is not used in the wild-type context. All the cryptic 5'ss were given equal weight, even though when activated they are used with widely different efficiencies, which might be a source of bias in the statistical comparisons. However, we are interested in defining sequences that can function as 5'ss when not out-competed by the authentic 5'ss, irrespective of the levels of use of such splice sites in the mutant genes.

We found no biases in the distribution of cryptic 5'ss in exons versus introns, or in each of the three reading frames (defined relative to that of the authentic 5'ss), suggesting that these parameters do not have a determining influence on the use of the cryptic splice sites. The absence of preference for cryptic 5'ss being in exons suggests that proximity to the upstream 3' splice site does not provide a significant contextual advantage for cryptic splice-site activation, in contrast to a previous proposal (55). Surprisingly, we detected no bias against the use of cryptic 5'ss involving a frameshift with respect to the authentic 5'ss, even though in many cases frameshifted mRNAs are expected to trigger the nonsense-mediated mRNA decay pathway (NMD). This pathway specifically degrades mRNAs with premature-termination codons (PTC)—generated by point mutations and/or as a result of splicing alterations (e.g. exon skipping)—whose

translation may be deleterious for the cell (56). We provide two possible explanations for this finding. First, NMD usually downregulates but does not completely eliminate mRNAs with PTCs (56), and therefore, the activated cryptic 5'ss that result in frameshifting might appear to be used less efficiently, but their use is still detectable. Secondly, use of some, but not all, cryptic 5'ss gives rise to mRNAs that appear to be resistant to NMD. For example, it has recently been shown that aberrant β-globin mRNAs that result from use of two cryptic 5'ss have different sensitivities to NMD, one of them being completely unresponsive to this pathway (57). Likewise, use of two cryptic 5'ss in the *C.elegans unc-73(e936)* mutant gene gives rise to NMD-resistant mRNAs (23). Understanding the molecular basis for this resistance to NMD would help explain the detection of some of the cryptic 5'ss included in this compilation, and perhaps also the failure to detect other potential cryptic 5'ss.

The mean distance of 11 nt between cryptic and authentic 5'ss places the typical cryptic 5'ss close to the natural exonic−intronic boundary, suggesting that the distance from the authentic 5'ss is an important constraint for cryptic splice-site activation. However, the absolute mean distance is 65 nt, indicating that the region where a given cryptic 5'ss is located can be relatively large. Furthermore, in many cases, several alternative cryptic 5'ss are used for a given mutant 5'ss, even though their S&S scores are sometimes very different. Taken together, these findings suggest that proximity to the authentic 5'ss is an important, but not strict determinant for cryptic 5'ss activation. The fact that the length of the internal exons that result from cryptic 5'ss activation is usually maintained within the optimal range suggests that cryptic 5'ss activation is subject to the usual constraints for optimal internal exon inclusion, consistent with the exon definition model (43). The distribution of 5'ss further suggests that the competitive advantage of the authentic over the cryptic 5'ss is not related to the length of the resulting exons. Furthermore, the hypothesis that the cryptic 5'ss are those good matches to the S&S matrix that are located closest to the authentic site is not supported by our statistical tests. Although distance to the authentic 5'ss is an important constraint, in some cases a cryptic and a pseudo 5'ss are

indistinguishable with respect to their S&S scores. Context sequences may explain the differential 5'ss selection in these cases. Indeed, the role of exonic splicing enhancers (ESEs) in 5'ss selection can be essential, and ESEs appear to be widespread (58). Specific sequences within introns can also be critical for 5'ss selection, as demonstrated, e.g. for the intronic G-triplets that are frequently present near 5'ss (59). The effect of splicing silencers, a less well defined category of splicing-repressor elements (60), may explain the competitive advantage of the cryptic over the pseudo 5'ss in these exceptions. Further experimental studies with specific substrates, to address the context requirements for cryptic splice-site activation, might prove informative.

Most importantly, the present analysis shows that in most cases, the differences in the 9 nt long 5'ss, measured with the S&S matrix, are sufficient to explain the competitive advantage of the authentic 5'ss versus the cryptic ones. This difference can explain why only the former are used in wild-type pre-mRNAs. The quantitative analysis of the 5'ss scores enables us to propose a gradation of 5'ss: splice-site selection at authentic 5'ss is more efficient than at cryptic 5'ss, which in turn is more efficient than at mutated versions of the authentic sites. As a point of reference, alternative 5'ss have scores that lie between those of authentic and cryptic 5'ss, and are slightly closer to the former. At present, we cannot predict which of the potential splice sites located close to a mutated authentic site will be activated, because the differences between cryptic 5'ss and other potential nearby (pseudo) 5'ss are subtle.

The question of how splicing can proceed with apparently 100% specificity, given that authentic splice sites are in competition with cryptic splice sites, remains unanswered. The contribution of putative specificity factors—proteins or RNAs—that bind the pre-mRNA and are involved in splice-site recognition, may be critical in some cases to distinguish between authentic and cryptic or pseudo-splice sites. Studies in our laboratory showed that the relative efficiencies of utilization of the three cryptic 5'ss in the human β-globin intron 1 can be changed, both *in vitro* and *in vivo*, depending on the levels of certain protein splicing factors (19−21).

SF2/ ASF, the founding member of the SR protein family, favors the use of the proximal (downstream) cryptic 5'ss, whereas hnRNP A1 favors the use of the distal (upstream) cryptic 5'ss. This effect was also observed with naturally occurring alternative 5'ss, suggesting interesting mechanistic and evolutionary links between cryptic splice-site activation and alternative splicing. In our study, alternative 5'ss are an intermediate category, with scores lower than those of authentic 5'ss, but higher than those of sub-optimal (cryptic, pseudo or mutated) 5'ss. However, this finding has to be interpreted with caution, given that the pool of alternative 5'ss we used is heterogeneous, as reflected by the high SD of their scores (Table 3). In contrast to the regulation of alternative 5'ss, for which several specific proteins have been described to influence their relative use, no protein factor is so far known to be involved in the discrimination between authentic and cryptic splice sites in a wild-type context. Alternatively, the NMD pathway might eliminate low levels of aberrant splicing arising from leaky use of certain cryptic 5'ss in a wild-type pre-mRNA, although this scenario remains hypothetical.

Consistent with the models of ‘exon definition’ (43) and ‘multiple weak interactions’ (61), exons whose 5'ss is disrupted by a mutation are either skipped during pre-mRNA splicing or included (at least partially) by using a cryptic 5'ss. Indeed, intermediate cases in which both pathways work alternatively are frequent. We suggest that inactivation of the 5'ss of an exon should result in exon skipping when the overall signals that define this sequence as an exon become sub-optimal as a result of the mutation, as proposed earlier (55). In contrast, exons associated with a 5'ss mutation that are included (at least in part) by utilizing a cryptic 5'ss still have a ‘splicing-favorable’ context; thus, other splicing signals, such as the 3' splice site and the ESEs, are sufficient to specify this pre-mRNA segment as an exon. Such a context allows a sub-optimal, neighboring 5'ss—the cryptic splice site—to be used. The threshold of signals that defines a sequence as an exon is still largely unknown because of our incomplete knowledge of all the relevant elements and their relative contributions to exon definition.

Comparison of most of the pairs of authentic and cryptic 5'ss shows significant differences in their S&S scores, which can explain why the correct splice sites are selected. There are a few exceptions to this trend, such as the example of human β-globin exon 1. Both *in vivo* and *in vitro* splicing experiments with β-globin pre-mRNA have shown that the cryptic 5'ss at−16, although it is close to, and has a higher score than, the correct splice site (at +1), is never used in the wild-type context (16,19,62). However, the authentic 5'ss at +1 has seven potential Watson–Crick base pairs plus one wobble base pair to the U1 snRNA 5' terminus, whereas the cryptic 5'ss at16 has only five Watson–Crick plus one wobble base pair to U1. The calculated free-energy parameters (–14 kcal/mol for +1 versus –9.9 kcal/mol for –16) are also consistent with a greater base-pairing potential of the +1 site to U1 snRNA. Thus, this exception could be attributable to a limitation of the S&S algorithm, which in this case does not reflect the intrinsic differences in efficiency between 5'ss at +1 (authentic 5'ss) and –16 (major cryptic 5'ss). For instance, one of the limitations of the S&S matrix is that it assumes independence between each of the nine positions. However, dependencies between each of the consensus positions are likely to exist, e.g. due to nearest-neighbor effects on RNA base-pairing stability. We have also shown that the lack of splicing via the 5'ss at –16 in the wild-type pre-mRNA is not due to the relative position of this splice site, or to its local sequence context, but rather to intrinsic differences between the cryptic and the competing authentic 5'ss. The results with both duplication mutants (Fig. 4B, lanes 2 and 3) suggest that splicing via the –16 5'ss sequence, but not the +1 sequence, can only take place at its natural position at –16, which could be due to its proximity to the 5' cap structure. On the other hand, the –16 cryptic 5'ss is not intrinsically defective, because it can function in the context of an inactive or debilitated site at +1, and moreover, identical 9 nt sequences function as normal 5'ss in other genes (such as the authentic 5'ss of exon 48 in the human *COL1A1* gene). Our findings thus far do not account for the remarkable fidelity of the splicing reaction, i.e. they do not explain why use of the –16 5'ss is undetectable in the wild-type β-globin context.

All the new β-globin mutant substrates tested in this study show reduced

splicing efficiency (Fig. 4B). Thus, these mutations could reduce splice-site recognition at both 5'ss or affect a subsequent step of the splicing reaction. One possible explanation for this phenomenon is that these mutations could affect the overall secondary structure of the β-globin pre-mRNA, thereby compromising the efficiency of the reaction. Alternatively, simultaneous assembly of splicing factors at nearby 5'ss could result in steric hindrance effects that may account for the reduced overall splicing efficiency of these mutants (12,63). The latter hypothesis would explain the reduced efficiency of both duplication mutants (Fig. 4B, lanes 2 and 3). However, the high splicing efficiencies of β-globin substrates bearing IVS1-G5C or T6C mutations (16), which result in splicing via both +1 and –16 5'ss, argue against a steric hindrance model. The reduced efficiency of the mutant with swapped sequences between the –16 and +1 5'ss (Fig. 4B, lane 4) also suggests that the inefficient splicing at position –16 is not specific to the sequence of the cryptic 5'ss at –16. Furthermore, the +1 5'ss mutant G1A (16), shown in lane 6 of Figure 4B, splices very efficiently via the –16 5'ss. Taken together, these results suggest that efficient splicing requires an optimal context, present only in the wild-type and single mutant substrates, and this context is disrupted by all tested extensive mutations of any of these 5'ss sequences.

Interestingly, the analysis of the strength of base pairing between the different categories of 5'ss and U1 snRNA revealed that 21 cryptic 5'ss have a stronger base-pairing potential than their corresponding authentic 5'ss, whereas only 13 cryptic 5'ss have higher S&S scores than their corresponding authentic 5'ss. In addition, nine of these exceptions were found by both methods. Thus, the S&S matrix appears to be a better heuristic discriminator between the different categories of 5'ss than the calculated stability of the 5'ss-U1 snRNA duplex. This observation suggests that stronger base pairing of authentic 5'ss to U1 snRNA may not be the general feature that distinguishes authentic and cryptic 5'ss, but rather other intrinsic sequence patterns within the consensus 5'ss sequence—somehow implicit in the S&S matrix—may contribute to the correct discrimination between these two categories of 5'ss. Another possibility, which remains to be tested, is that the

sequences surrounding the 5'ss can play a pivotal role in discriminating between the authentic and the cryptic 5'ss, similar to what we proposed above for the discrimination between cryptic and pseudo 5'ss.

We also analyzed the scores of authentic and cryptic 5'ss using four other methods (Table 4), and compared them with the S&S scores and the free-energy parameters. These four new matrices are, at least to some extent, better discriminators between these 5'ss, but the combination of exceptions suggests that none of these methods is sufficiently reliable: for instance, out of the six exceptions found using the NN and six using the MDD methods, only three were found by both. However, we believe that the sample size in this study is not large enough to accurately compare the reliability of these 5'ss scoring methods. All the methods gave the same relative ranking of authentic > alternative > cryptic > mutant 5'ss (data not shown). We did not calculate pseudo splice sites by the other methods because the threshold chosen for calculations by one method tends to give very low cut-offs for the other ones.

Accurate prediction of the splicing patterns that result from a mutation, as well as of the structural and functional consequences for the corresponding protein, will help to understand allele-specific differences in various diseases. Furthermore, therapeutic approaches to correct splicing defects relate to cryptic splice-site activation. For instance, Dominski and Kole have used an antisense approach to suppress the use of a cryptic 5'ss and thereby rescue a mutant 5'ss (64). The present study and further evaluation of cryptic 5'ss activation will be helpful in the application of such techniques for the treatment of diseases caused by splice-site mutations.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We are thankful to Jun Zhu for sharing extracts and reagents, and to other members of the laboratory for helpful advice and discussions. We are also grateful to Michelle Hastings and Luca Cartegni for useful comments on the manuscript. X.R. and A.R.K. acknowledge support from NCI grant CA13106.

REFERENCES

1. Sun,H. and Chasin,L.A. (2000) Multiple splicing defects in an intronic false exon. *Mol. Cell. Biol.*, **20**, 6414–6425.
2. Horowitz,D.S. and Krainer,A.R. (1994) Mechanisms for selecting 5' splice sites in mammalian pre-mRNA splicing. *Trends Genet.*, **10**, 100–106.
3. Zhuang,Y. and Weiner,A.M. (1986) A compensatory base change in U1 snRNA suppresses a 5' splice site mutation. *Cell*, **46**, 827–835.
4. Siliciano,P.G. and Guthrie,C. (1988) 5' splice site selection in yeast: genetic alterations in base-pairing with U1 reveal additional requirements. *Genes Dev.*, **2**, 1258–1267.
5. Seraphin,B., Kretzner,L. and Rosbash,M. (1988) A U1 snRNA:pre-mRNA base pairing interaction is required early in yeast spliceosome assembly but does not uniquely define the 5' cleavage site. *EMBO J.*, **7**, 2533–2538.
6. Crispino,J.D., Blencowe,B.J. and Sharp,P.A. (1994) Complementation by SR proteins of pre-mRNA splicing reactions depleted of U1 snRNP. *Science*, **265**, 1866–1869.
7. Crispino,J. and Sharp,P.A. (1995) U6 snRNA:pre-mRNA interaction can be rate-limiting for U1-independent splicing. *Genes Dev.*, **9**, 2314–2323.
8. Tarn,W. and Steitz,J. (1994) SR proteins can compensate for the loss of U1 snRNP functions *in vitro*. *Genes Dev.*, **8**, 2704–2717.
9. Lund,M. and Kjems,J. (2002) Defining a 5' splice site by functional selection in the presence and absence of U1 snRNA 5' end. *RNA*, **8**, 166–179.
10. Du,H. and Rosbash,M. (2001) Yeast U1 snRNP-pre-mRNA complex formation without U1 snRNA-pre-mRNA base pairing. *RNA*, **7**, 133–142.
11. Du,H. and Rosbash,M. (2002) The U1 snRNP protein U1C recognizes the 5' splice site in the absence of base pairing. *Nature*, **419**, 86–90.
12. Eperon,I.C., Ireland,D.C., Smith,R.A., Mayeda,A. and Krainer,A.R. (1993) Pathways for selection of 5' splice sites by U1 snRNPs and SF2/ASF. *EMBO J.*, **12**, 3607–3617.
13. Wassarman,D.A. and Steitz,J.A. (1992) Interactions of small nuclear RNA's

with precursor messenger RNA during *in vitro* splicing. *Science*, **257**, 1918–1925.

14. Brackenridge,S., Wilkie,A.O. and Screaton,G.R. (2003) Efficient use of a 'dead-end' GA 5' splice site in the human fibroblast growth factor receptor genes. *EMBO J.*, **22**, 1620–1631.
15. Hwang,D.Y. and Cohen,J.B. (1996) U1 snRNA promotes the selection of nearby 5' splice sites by U6 snRNA in mammalian cells. *Genes Dev.*, **10**, 338–350.
16. Treisman,R., Orkin,S.H. and Maniatis,T. (1983) Specific transcription and RNA splicing defects in five cloned beta-thalassaemia genes. *Nature*, **302**, 591–596.
17. Chabot,B. and Steitz,J.A. (1987) Recognition of mutant and cryptic 5' splice sites by the U1 small nuclear ribonucleoprotein *in vitro*. *Mol. Cell. Biol.*, **7**, 698–707.
18. Nelson,K.K. and Green,M.R. (1990) Mechanism for cryptic splice site activation during pre-mRNA splicing. *Proc. Natl Acad. Sci. USA*, **87**, 6253–6257.
19. Cáceres,J.F., Stamm,S., Helfman,D.M. and Krainer,A.R. (1994) Regulation of alternative splicing in vivo by overexpression of antagonistic splicing factors. *Science*, **265**, 1706–1709.
20. Krainer,A.R., Conway,G.C. and Kozak,D. (1990) The essential pre-mRNA splicing factor SF2 influences 5' splice site selection by activating proximal sites. *Cell*, **62**, 35–42.
21. Mayeda,A. and Krainer,A.R. (1992) Regulation of alternative pre-mRNA splicing by hnRNP A1 and splicing factor SF2. *Cell*, **68**, 365–375.
22. Domenjoud,L., Kister,L., Gallinaro,H. and Jacob,M. (1993) Selection between a natural and a cryptic 5' splice site: a kinetic study of the effect of upstream exon sequences. *Gene Expr.*, **3**, 83–94.
23. Roller,A., Hoffman,D. and Zahler,A. (2000) The allele-specific suppressor sup-39 alters use of cryptic splice sites in *Caenorhabditis elegans*. *Genetics*, **154**, 1169–1179.

24. Run,J.Q., Steven,R., Hung,M.S., van Weeghel,R., Culotti,J.G. and Way,J.C. (1996) Suppressors of the unc-73 gene of *Caenorhabditis elegans*. *Genetics*, **143**, 225–236.
25. Newman,A. and Norman,C. (1992) U5 snRNA interacts with exon sequences at 5' and 3' splice sites. *Cell*, **68**, 743–754.
26. Kandels-Lewis,S. and Seraphin,B. (1993) Involvement of U6 snRNA in 5' splice site selection. *Science*, **262**, 2035–2039.
27. Alvarez,C.J. and Wise,J.A. (2001) Activation of a cryptic 5' splice site by U1 snRNA. *RNA*, **7**, 342–350.
28. 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.
29. 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.
30. Ars,E., Serra,E., Garcia,J., Kruyer,H., Gaona,A., Lazaro,C. and Estivill,X. (2000) Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. *Hum. Mol. Genet*. **9**, 237–247.
31. Teraoka,S.N., Telatar,M., Becker-Catania,S., Liang,T., Onengut,S., Tolun,A., Chessa,L., Sanal,O., Bernatowska,E., Gatti,R.A. and Concannon,P. (1999) Splicing defects in the ataxia-telangiectasia gene, ATM: underlying mutations and consequences. *Am. J. Hum. Genet*., **64**, 1617–1631.
32. Nakai,K. and Sakamoto,H. (1994) Construction of a novel database containing aberrant splicing mutations of mammalian genes. *Gene*, **141**, 171–177.
33. Lee,C., Atanelov,L., Modrek,B. and Xing,Y. (2003) ASAP: the Alternative Splicing Annotation Project. *Nucleic Acids Res*., **31**, 101–105.
34. Senapathy,P., Shapiro,M.B. and Harris,N.L. (1990) Splice junctions, branch point sites and exons: sequence statistics, identification and applications to genome project. *Methods Enzymol*., **183**, 252–278.
35. 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.

36. Serra,M.J. and Turner,D.H. (1995) Predicting thermodynamic properties of RNA. *Methods Enzymol*., **259**, 242−261.
37. Mayeda,A. and Krainer,A.R. (1999) Mammalian in vitro splicing assays. *Methods Mol. Biol*., **118**, 315−321.
38. Mayeda,A. and Krainer,A.R. (1999) Preparation of HeLa cell nuclear and cytosolic S100 extracts for *in vitro* splicing. *Methods Mol. Biol*., **118**, 309−314.
39. Delfau,M.H., Picat,C., De Rooij,F., Voortman,G., Deybach,J.C., Nordmann,Y. and Grandchamp,B. (1991) Molecular heterogeneity of acute intermittent porphyria: identification of four additional mutations resulting in the CRIM-negative subtype of the disease. *Am. J. Hum. Genet*., **49**, 421−428.
40. Sakuntabhai,A., Hammami-Hauasli,N., Bodemer,C., Rochat,A., Prost,C., Barrandon,Y., de Prost,Y., Lathrop,M., Wojnarowska,F., Bruckner-Tuderman,L. and Hovnanian,A. (1998) Deletions within COL7A1 exons distant from consensus splice sites alter splicing and produce shortened polypeptides in dominant dystrophic epidermolysis bullosa. *Am. J. Hum. Genet*., **63**, 737−748.
41. Colgin,L.M., Hackmann,A.F. and Monnat,R.J.,Jr (1999) Different somatic and germline HPRT1 mutations promote use of a common, cryptic intron 1 splice site. *Hum. Mutat*., **14**, 92.
42. MacLeod,J.N., Liebhaber,S.A., MacGillivray,M.H. and Cooke,N.E. (1991) Identification of a splice-site mutation in the human growth hormone-variant gene. *Am. J. Hum. Genet*., **48**, 1168−1174.
43. Berget,S. (1995) Exon recognition in vertebrate splicing. *J. Biol. Chem*., **270**, 2411−2414.
44. Li,B., Wachtel,C., Miriami,E., Yahalom,G., Friedlander,G., Sharon,G., Sperling,R. and Sperling,J. (2002) Stop codons affect 5' splice site selection by surveillance of splicing. *Proc. Natl Acad. Sci. USA*, **99**, 5277−5282.
45. Zhang,X., Lee,J. and Chasin,L.A. (2003) The effect of nonsense codons on splicing: a genomic analysis. *RNA*, **9**, 637−639.

46. Gardella,R., Belletti,L., Zoppi,N., Marini,D., Barlati,S. and Colombi,M. (1996) Identification of two splicing mutations in the collagen type VII gene (COL7A1) of a patient affected by the localisata variant of recessive dystrophic epidermolysis bullosa. *Am. J. Hum. Genet.*, **59**, 292–300.
47. Brunak,S., Engelbrecht,J. and Knudsen,S. (1991) Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.*, **220**, 49–65.
48. Yeo,G. and Burge,C.B. (2003) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. In Miller,W., Vingron,M., Istrail,S., Pevzner,P. and Waterman,M. (eds), *Proceedings of the 7th Annual International Conference on Computational Molecular Biology (RECOMB03)*. ACM Press, New York, NY, pp. 322–331.
49. Zielenski,J., Bozon,D., Markiewicz,D., Aubin,G., Simard,F., Rommens,J.M. and Tsui,L.C. (1993) Analysis of CFTR transcripts in nasal epithelial cells and lymphoblasts of a cystic fibrosis patient with 621 + 1G→T and 711 + 1G→T mutations. *Hum. Mol. Genet.*, **2**, 683–687.
50. Kuivaniemi,H., Kontusaari,S., Tromp,G., Zhao,M.J., Sabol,C. and Prockop,D.J. (1990) Identical G+1 to A mutations in three different introns of the type III procollagen gene (COL3A1) produce different patterns of RNA splicing in three variants of Ehlers-Danlos syndrome. IV. An explanation for exon skipping some mutations and not others. *J. Biol. Chem.*, **265**, 12067–12074.
51. Pinotti,M., Toso,R., Redaelli,R., Berrettini,M., Marchetti,G. and Bernardi,F. (1998) Molecular mechanisms of FVII deficiency: expression of mutations clustered in the IVS7 donor splice site of factor VII gene. *Blood*, **92**, 1646–1651.
52. Furihata,K., Drousiotou,A., Hara,Y., Christopoulos,G., Stylianidou,G., Anastasiadou,V., Ueno,I. and Ioannou,P. (1999) Novel splice site mutation at IVS8 nt 5 of HEXB responsible for a Greek–Cypriot case of Sandhoff disease. *Hum. Mutat.*, **13**, 38–43.
53. Maruyama,T., Miyake,Y., Tajima,S., Funahashi,T., Matsuzawa,Y. and Yamamoto,A. (1995) A single point mutation in the splice donor site of the low-

density-lipoprotein-receptor gene produces intron read-through, exon-skipped and cryptic-site-utilized transcripts. *Eur. J. Biochem.*, **232**, 700−705.

54. Gotoda,T., Yamada,N., Murase,T., Inaba,T., Ishibashi,S., Shimano,H., Koga,S., Yazaki,Y., Furuichi,Y. and Takaku,F. (1991) Occurrence of multiple aberrantly spliced mRNAs upon a donor splice site mutation that causes familial lipoprotein lipase deficiency. *J. Biol. Chem.*, **266**, 24757−24762.
55. Robberson,B.L., Cote,G.J. and Berget,S.M. (1990) Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol. Cell. Biol.*, **10**, 84−94.
56. Maquat,L.E. (2000) Nonsense-mediated RNA decay in mammalian cells: a splicing-dependent means to down-regulate the levels of mRNAs that prematurely terminate translation. In Sonenberg,N., Hershey,J.W.B. and Mathews,M.B. (eds), *Translational Control of Gene Expression*. Cold Spring Harbor Press, Cold Spring Harbor, New York, NY, pp. 849−868.
57. Danckwardt,S., Neu-Yilik,G., Thermann,R., Frede,U., Hentze,M.W. and Kulozik,A.E. (2002) Abnormally spliced beta-globin mRNAs: a single point mutation generates transcripts sensitive and insensitive to nonsense-mediated mRNA decay. *Blood*, **99**, 1811−1816.
58. Cartegni,L., Chew,S.L. and Krainer,A.R. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Rev. Genet.*, **3**, 285−298.
59. McCullough,A. and Berget,S. (1997) G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Mol. Cell. Biol.*, **17**, 4562−4571.
60. Ladd,A.N. and Cooper,T.A. (2002) Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol.*, **3**, 0008.1−0008.16 (reviews).
61. Reed,R. (1996) Initial splice-site recognition and pairing during pre-mRNA splicing. *Curr. Opin. Genet. Dev.*, **6**, 215−220.
62. Krainer,A.R., Maniatis,T., Ruskin,B. and Green,M.R. (1984) Normal and mutant human beta-globin pre-mRNAs are faithfully and efficiently spliced *in vitro*. *Cell*, **36**, 993-1005.
63. Cunningham,S.A., Else,A.J., Potter,B.V. and Eperon,I.C. (1991) Influences

of separation and adjacent sequences on the use of alternative 5' splice sites. *J. Mol. Biol.*, **217**, 265–281.

64. Dominski,Z. and Kole,R. (1993) Restoration of correct splicing in thalassemic pre-mRNA by antisense oligonucleotides. *Proc. Natl Acad. Sci. USA*, **90**, 8673–8677.
65. Guimaraes,C.P., Lemos,M., Menezes,I., Coelho,T., Sa-Miranda,C. and Azevedo,J. (2001) Characterisation of two mutations in the ABCD1 gene leading to low levels of normal ALDP. *Hum. Genet.*, **109**, 616–622.
66. Nemeth-Slany,A., Talmud,P., Grundy,S.M. and Patel,S.B. (1997) Activation of a cryptic splice-site in intron 24 leads to the formation of apolipoprotein B-27.6. *Atherosclerosis*, **133**, 163–170.
67. Ris-Stalpers,C., Kuiper,G.G., Faber,P.W., Schweikert,H.U., van Rooij,H.C., Zegers,N.D., Hodgins,M.B., Degenhart,H.J., Trapman,J. and Brinkmann,A.O. (1990) Aberrant splicing of androgen receptor mRNA results in synthesis of a nonfunctional receptor protein in a patient with androgen insensitivity. *Proc. Natl Acad. Sci. USA*, **87**, 7866–7870.
68. Vega,A., Campos,B., Bressac-De-Paillerets,B., Bond,P.M., Janin,N., Douglas,F.S., Domenech,M., Baena,M., Pericay,C., Alonso,C., Carracedo,A., Baiget,M. and Diez,O. (2001) The R71G BRCA1 is a founder Spanish mutation and leads to aberrant splicing of the transcript. *Hum. Mutat.*, **17**, 520–521.
69. Scholl,T., Pyne,M.T., Russo,D. and Ward,B.E. (1999) BRCA1 IVS16+6T-*C is a deleterious mutation that creates an aberrant transcript by activating a cryptic splice donor site. *Am. J. Med. Genet.*, **85**, 113–116.
70. Jones,C.T., McIntosh,I., Keston,M., Ferguson,A. and Brock,D.J. (1992) Three novel mutations in the cystic fibrosis gene detected by chemical cleavage: analysis of variant splicing and a nonsense mutation. *Hum. Mol. Genet.*, **1**, 11–17.
71. Wang,Q., Forlino,A. and Marini,J.C. (1996) Alternative splicing in COL1A1 mRNA leads to a partial null allele and two in-frame forms with structural defects in non-lethal osteogenesis imperfecta. *J. Biol. Chem.*, **271**,

28617−28623.

72. Schwarze,U., Starman,B.J. and Byers,P.H. (1999) Redefinition of exon 7 in the COL1A1 gene of type I collagen by an intron 8 splice-donorsite mutation in a form of osteogenesis imperfecta: influence of intron splice order on outcome of splice-site mutation. *Am. J. Hum. Genet.*, **65**, 336−344.
73. Bateman,J.F., Chan,D., Moeller,I., Hannagan,M. and Cole,W.G. (1994) A 5' splice site mutation affecting the pre-mRNA splicing of two upstream exons in the collagen COL1A1 gene. Exon 8 skipping and altered definition of exon 7 generates truncated pro alpha 1(I) chains with a non-collagenous insertion destabilizing the triple helix. *Biochem.* J., **302**, 729−735.
74. Vanegas,O.C., Zhang,R.Z., Sabatelli,P., Lattanzi,G., Bencivenga,P., Giusti,B., Columbaro,M., Chu,M.L., Merlini,L. and Pepe,G. (2002) Novel COL6A1 splicing mutation in a family affected by mild Bethlem myopathy. *Muscle Nerve*, **25**, 513−519.
75. Botto,M., Fong,K.Y., So,A.K., Rudge,A. and Walport,M.J. (1990) Molecular basis of hereditary C3 deficiency. *J. Clin. Invest.*, **86**, 1158−1163.
76. Harada,N., Ogawa,H., Shozu,M., Yamada,K., Suhara,K., Nishida,E. and Takagi,Y. (1992) Biochemical and molecular genetic analyses on placental aromatase (P-450AROM) deficiency. *J. Biol. Chem.*, **267**, 4781−4785.
77. Chen,W., Kubota,S., Ujike,H., Ishihara,T. and Seyama,Y. (1998) A novel Arg362Ser mutation in the sterol 27-hydroxylase gene (CYP27): its effects on pre-mRNA splicing and enzyme activity. *Biochemistry*, **37**, 15050−15056.
78. Roest,P.A., Bout,M., van der Tuijn,A.C., Ginjaar,I.B., Bakker,E., Hogervorst,F.B., van Ommen,G.J. and den Dunnen,J.T. (1996) Splicing mutations in DMD/BMD detected by RT−PCR/PTT: detection of a 19AA insertion in the cysteine rich domain of dystrophin compatible with BMD. *J. Med. Genet.*, **33**, 935−939.
79. Wilton,S.D., Chandler,D.C., Kakulas,B.A. and Laing,N.G. (1994) Identification of a point mutation and germinal mosaicism in a Duchenne muscular dystrophy family. *Hum. Mutat.*, **3**, 133−140.
80. Hahn,S.H., Krasnewich,D., Brantly,M., Kvittingen,E.A. and Gahl,W.A. (1995)

Heterozygosity for an exon 12 splicing mutation and a W234G missense mutation in an American child with chronic tyrosinemia type 1. *Hum. Mutat.*, **6**, 66−73.

81. Hutchinson,S., Wordsworth,B.P. and Handford,P.A. (2001) Marfan syndrome caused by a mutation in FBN1 that gives rise to cryptic splicing and a 33 nucleotide insertion in the coding sequence. *Hum. Genet.* **109**, 416−420.
82. Schrijver,I., Koerper,M.A., Jones,C.D. and Zehnder,J.L. (2002) Homozygous factor V splice site mutation associated with severe factor V deficiency. *Blood*, **99**, 3063−3065.
83. Attanasio,C., de Moerloose,P., Antonarakis,S.E., Morris,M.A. and Neerman-Arbez,M. (2001) Activation of multiple cryptic donor splice sites by the common congenital afibrinogenemia mutation, FGA IVS4 + 1 G→T. *Blood*, **97**, 1879−1881.
84. Sun,F., Knebelmann,B., Pueyo,M.E., Zouali,H., Lesage,S., Vaxillaire,M., Passa,P., Cohen,D., Velho,G., Antignac,C. *et al.* (1993) Deletion of the donor splice site of intron 4 in the glucokinase gene causes maturity-onset diabetes of the young. *J. Clin. Invest.*, **92**, 1174−1180.
85. Felber,B.K., Orkin,S.H. and Hamer,D.H. (1982) Abnormal RNA splicing causes one form of alpha thalassemia. *Cell*, **29**, 895−902.
86. Akli,S., Chelly,J., Kahn,A. and Poenaru,L. (1993) A null allele frequent in non-Jewish Tay-Sachs patients. *Hum. Genet.*, **90**, 614−620.
87. Buesa,C., Pie,J., Barcelo,A., Casals,N., Mascaro,C., Casale,C.H., Haro,D., Duran,M., Smeitink,J.A. and Hegardt,F.G. (1996) Aberrantly spliced mRNAs of the 3-hydroxy-3-methylglutaryl coenzyme A lyase (HL) gene with a donor splice-site point mutation produce hereditary HL deficiency. *J. Lipid Res.*, **37**, 2420−2432.
88. Hunter,T.C., Melancon,S.B., Dallaire,L., Taft,S., Skopek,T.R., Albertini,R.J. and O'Neill,J.P. (1996) Germinal HPRT splice donor site mutation results in multiple RNA splicing products in T-lymphocyte cultures. Somat. Cell Mol. *Genet.*, **22**, 145−150.
89. Bunge,S., Steglich,C., Zuther,C., Beck,M., Morris,C.P., Schwinger,E.,

Schinzel,A., Hopwood,J.J. and Gal,A. (1993) Iduronate-2-sulfatase gene mutations in 16 patients with mucopolysaccharidosis type II (Hunter syndrome). *Hum. Mol. Genet.*, **2**, 1871–1875.

90. Matsuura,S., Kishi,F., Tsukahara,M., Nunoi,H., Matsuda,I., Kobayashi,K. and Kajii,T. (1992) Leukocyte adhesion deficiency: identification of novel mutations in two Japanese patients with a severe form. *Biochem. Biophys. Res. Commun.*, **184**, 1460–1467.
91. Jin,Y., Dietz,H.C., Nurden,A. and Bray,P.F. (1993) Single-strand conformation polymorphism analysis is a rapid and effective method for the identification of mutations and polymorphisms in the gene for glycoprotein IIIa. *Blood*, **82**, 2281–2288.
92. Rugg,E.L., Rachet-Prehu,M.O., Rochat,A., Barrandon,Y., Goossens,M., Lane,E.B. and Hovnanian,A. (1999) Donor splice site mutation in keratin 5 causes in-frame removal of 22 amino acids of H1 and 1A rod domains in Dowling-Meara epidermolysis bullosa simplex. *Eur. J. Hum. Genet.*, **7**, 293–300.
93. Goyette,P., Frosst,P., Rosenblatt,D.S. and Rozen,R. (1995) Seven novel mutations in the methylenetetrahydrofolate reductase gene and genotype/phenotype correlations in severe methylenetetrahydrofolate reductase deficiency. *Am. J. Hum. Genet.*, **56**, 1052–1059.
94. Jacoby,L.B., MacCollin,M., Barone,R., Ramesh,V. and Gusella,J.F. (1996) Frequency and distribution of NF2 mutations in schwannomas. *Genes Chromosomes Cancer*, **17**, 45–55.
95. Mustajoki,S., Pihlaja,H., Ahola,H., Petersen,N.E., Mustajoki,P. and Kauppinen,R. (1998) Three splicing defects, an insertion and two missense mutations responsible for acute intermittent porphyria. *Hum. Genet.*, **102**, 541–548.
96. Tsujino,S., Tonin,P., Shanske,S., Nohria,V., Boustany,R.M., Lewis,D., Chen,Y.T. and DiMauro,S. (1994) A splice junction mutation in a new myopathic variant of phosphoglycerate kinase deficiency (PGK North Carolina). *Ann. Neurol.*, **35**, 349–353.

97. Peral,B., Gamble,V., San Millan,J.L., Strong,C., Sloane-Stanley,J., Moreno,F. and Harris,P.C. (1995) Splicing mutations of the polycystic kidney disease 1 (PKD1) gene induced by intronic deletion. *Hum. Mol. Genet.*, **4**, 569–574.

98. Celebi,J.T., Wanner,M., Ping,X.L., Zhang,H. and Peacocke,M. (2000) Association of splicing defects in PTEN leading to exon skipping or partial intron retention in Cowden syndrome. *Hum. Genet.*, **107**, 234–238.

99. Tsujino,S., Shanske,S., Nonaka,I., Eto,Y., Mendell,J.R., Fenichel,G.M. and DiMauro,S. (1994) Three new mutations in patients with myophosphorylase deficiency (McArdle disease). *Am. J. Hum. Genet.*, **54**, 44–52.

100. Dry,K.L., Manson,F.D., Lennon,A., Bergen,A.A., Van Dorp,D.B. and Wright,A.F. (1999) Identification of a 5' splice site mutation in the RPGR gene in a family with X-linked retinitis pigmentosa (RP3). *Hum. Mutat.*, **13**, 141–145.

101. Laubach,V.E., Ryan,W.J. and Brantly,M. (1993) Characterization of a human alpha 1-antitrypsin null allele involving aberrant mRNA splicing. *Hum. Mol. Genet.*, **2**, 1001–1005.

102. Ishioka,C., Sato,T., Gamoh,M., Suzuki,T., Shibata,H., Kanamaru,R., Wakui,A. and Yamazaki,T. (1991) Mutations of the P53 gene, including an intronic point mutation, in colorectal tumors. *Biochem. Biophys. Res. Commun.*, **177**, 901–906.

103. Sakai,E. and Tsuchida,N. (1992) Most human squamous cell carcinomas in the oral cavity contain mutated p53 tumor-suppressor genes. *Oncogene*, **7**, 927–933.

104. vanBakel,I., Sepp,T., Ward,S., Yates,J.R. and Green,A.J. (1997) Mutations in the TSC2 gene: analysis of the complete coding sequence using the protein truncation test (PTT). *Hum. Mol. Genet.*, **6**, 1409–1414.

105. Gantla,S., Bakker,C.T., Deocharan,B., Thummala,N.R., Zweiner,J., Sinaasappel,M., Roy Chowdhury,J., Bosma,P.J. and Roy Chowdhury,N. (1998) Splice-site mutations: a novel genetic mechanism of Crigler-Najjar syndrome type 1. *Am. J. Hum. Genet.*, **62**, 585–592.

List of Tables

List of Figures

Fig. 4 Analysis of cryptic 5'ss activation in human β-globin. (**A**) Cryptic 5'ss in the human β-globin gene. The diagram shows the first two exons (gray boxes) and first intron (horizontal thin line) of *HBB*. The sequence around the first intron 5'ss is shown below the diagram. Vertical lines represent the cryptic 5'ss, whose GT dinucleotides are underlined. The arrows indicate the cleavage/ligation sites. The phase, or position of the intron within a codon, is given in Roman numerals, and the number in parenthesis is the relative position of the cryptic 5'ss relative to the authentic 5'ss at +1. The numbers above each splice site are their S&S consensus values. The authentic (+1) and main cryptic (−16) 5'ss are shown in red. The cryptic 5'ss at +13 was mutated in all the constructs (see Materials and Methods), and the cryptic 5'ss at −38 is used very inefficiently. Three different β-thalassemia mutations (16) are shown below the sequence, with the position and nucleotide substitution indicated in each case. (**B**) *In vitro* splicing analysis of β-globin substrates. Labeled pre-mRNAs were spliced in HeLa cell nuclear extract, and the products were analyzed by electrophoresis on a denaturing polyacrylamide gel and autoradiography. Each construct is shown as a vertical diagram above each lane, with the exons as light-blue boxes and the intron as a line. The horizontal line across exon 1 shows the position of the main cryptic 5'ss at position −16. The small circles indicate the authentic (green) or mutant (red) 5'ss, and the open red 'do not' symbol denotes extensive mutation of the authentic 5'ss. The mobilities of the pre-mRNA and spliced mRNA bands are indicated by the diagrams on the left. Lane 1, the authentic 5'ss at +1 was mutated from CAG/GTTGGT to CAG/AACCCG; lane 2, the cryptic 5'ss at −16 was mutated to a duplicate copy of the authentic site at +1; lane 3, the authentic site at +1 was mutated to a duplicate copy of the cryptic 5'ss at −16; lane 4, the positions of the authentic site at +1 and the cryptic site at −16 were swapped; lane 5, wild-type

pre-mRNA; lane 6, IVS1-G1A thalassemia mutation.

Gene	Disease/defect	Mutation	Position of cryptic 5′ss[a]	Reference
ABCD1	X-linked adrenoleukodystrophy (X-ALD)	IVS1 G(–1)A	IVS1 +10	65
APOB	Homozygous hypobetalipoproteinemia	IVS24 T2C	IVS24 +41	66
AR	Androgen insensitivity	IVS4 G1T	Exon4 –123	67
ATM	Ataxia-telangiectasia	IVS45 G1A	IVS45 +72, +80	31
BRCA1	Breast cancer	IVS5 A(–2)G	Exon5 –22	68
		IVS16 T6C	IVS16 +70	69
CFTR	Cystic fibrosis	IVS20 G(–1)C	IVS20 +30	70
		IVS4 G1T	Exon4 –93	49
COL1A1	Severe type III osteogenesis imperfecta	Exon34 del[–3:IVS36+X]	Exon34 –8	71
		IVS8 G1A	IVS8 G +97/exon 8 –26	72
		IVS8 G5A	IVS8 +97	73
COL3A1	Ehlers-Danlos syndrome IV	IVS16 G1A	IVS16 +24	50
		IVS20 G1A	IVS20 +25	50
		IVS42 G1A	IVS42 +31	50
COL6A1	Mild Bethlem myopathy	IVS3 G1A	Exon3 –66	74
COL7A1	Recessive dystrophic epidermolysis bullosa	IVS3 A(–2)G	Exon 3 –104	46
		IVS95 G(–1)A	Exon95 –7	46
	Dominant dystrophic epidermolysis bullosa	Exon73 del[–98: –71]	Exon73 –62	40
C3	Hereditary C3 deficiency	IVS18 G1A	Exon18 –61	75
CYP19	Placental aromatase deficiency	IVS6 T2C	IVS6 +88	76
CYP27A1	Cerebrotendinous xanthomatosis	IVS6 C(–2)A/G(–1)A	Exon6 –89	77
DMD	Duchenne and Becker muscular dystrophy	IVS64 G5C	IVS64 +58	78
		IVS26 T2G	IVS26 +117	79
FAH	Chronic tyrosinemia type I	IVS12 G5A	IVS112 +106	80
FBN1	Marfan syndrome	IVS46 G1A	IVS46 +34	81
F5	Severe factor V deficiency	IVS10 G(–1)T	Exon10 –35	82
FGA	Common congenital afibrinogenemia	IVS4 G1T	IVS4 +5/exon4 –1, –66, –36	83
F7	FVII deficiency	IVS7 G5A/A7G/del[+3:+6]	IVS7 +38	51
GCK	Maturity onset diabetes of the young (MODY)	IVS4 del[+2:+16]	Exon4 –24	84
GHV	Mutation in placenta	IVS2 G1A	IVS2 +13	42
HBA2	Alpha-thalassemia	IVS1 del[+2:+6]	Exon1 –49	85
HBB	Beta-thalassemia	IVS1 G(–1)C/G1A/G1T/T2C/G5A/G5C/G5T/T6C	Exon1 –38, –16/IVS1 +13	16
		IVS2 del[+4:+5]	Exon2 –135/ IVS2 +48	32
HEXA	Tay-Sachs syndrome	IVS9 G1A	IVS9 +18	86
HEXB	Sandhoff disease	IVS8 G5C	Exon8 –4	52
HMGCL	Hereditary HL deficiency	IVS 7	IVS7 +79	87
HPRT1	Somatic mutations in kidney tubular epithelial cells	IVS1 G5A/G5T/del[-2:+34]	IVS1 +50	41
	Lesch-Nyhan syndrome	IVS5 T2G/AA3:4GT/G5A/del[G1]	IVS5 +68	88
IDS	Mucopolysaccharidosis type II (Hunter syndrome)	IVS7 GG[–1:+1]TT	IVS7 +23	89
ITGB2	Leukocyte adhesion deficiency	IVS7 G1A	IVS7 +65,+299	90
ITGB3	Glanzmann thrombasthenia	IVS4 G1A	IVS4 +28	91
KRT5	Dowling-Meara epidermolysis bullosa simplex	IVS1 G1A	Exon1 –66	92
LDLR	Familial hypercholesterolemia	IVS12 T2C	IVS12 +12	53
LPL	Familial hypercholesterolemia	IVS2 G1A	Exon2 –18/IVS2 +43, +143, +247, +383	54
MTHFR	Severe deficiency of MTHFR	IVS4 G1A	Exon4 –57	93
NF1	Neurofibromatosis type I	IVS27b del[+1:+10]	Exon27b –69	30
		IVS28 G1A	Exon28 –54	30
NF2	Neurofibromatosis type II	IVS7 G5A/del[–3:+11]	Exon7 –23, –28	94
		IVS12 G1A/del[–14:+2]	Exon12 –38, –53	94
PBGD	Acute intermittent porphyria	IVS1 G1A/T2A/G5C/G3T	IVS1 +68	95
		IVS10 G(–1)T	Exon 10 –9	39
PGK1	Phosphoglycerate kinase deficiency	IVS4 G1T	IVS4 +31	96
PKD1	Polycystic kidney disease 1	IVS43	Exon43 –66	97
PTEN	Cowden syndrome	IVS7 G1A	IVS7 +76	98
		IVS4 T2C	IVS4 +5	98
PYGM	Myophosphorylase deficiency (McArdle disease)	IVS14 G1A	Exon14 –67	99
RPGR	X-linked retinitis pigmentosa (RP3)	IVS5 G1T	Exon5 –76	100
SERPINA1	Risk for emphysema	IVS2 G1T	Exon2 –84	101
TP53	Colorectal tumors	IVS5 G5C	Exon5 –46	102
	Squamous cell carcinoma	IVS6 G(–1)A/G1A	IVS6 +6	103
TSC2	Familial tuberous sclerosis	IVS37 ins[+2A]	Exon37 –29	104
UGT1A1	Crigler-Najjar syndrome type 1	IVS1 G1C	Exon1 –141	105
46 genes			76 cryptic 5′ss	

IVS, intervening sequence or intron.
[a]Position relative to the authentic 5′ss; positive numbers towards the downstream intron, negative numbers towards the upstream exon.

Table 1

Number of genes	46	
Number of cryptic 5′ss	76	
In exons	37 (49%)	
In introns	39 (51%)	
Reading frame (C relative to A)[a]		
0	32 (42%)	Chi-square (2) = 2.42,
I	24 (32%)	$P = 0.30$
II	20 (26%)	
Distance of C to A (nt)	Average	SD
All[b]	10.92	88.93
Absolute	62.58	63.73
Exonic C	53.05	35.92
Exonic C (internal)[c]	51.66	34.49
Intronic C	71.62	81.37
Intronic C (internal)[c]	75.86	84.55
Exon length A	200.03	217.44
Exon length A (internal)[c]	160.22	98.88
Exon length C	201.58	195.18
Exon length C (internal)[c]	174.21	111.56
Closest C (to A)	–4, +4	
Range	–141 to +398	

[a]C, cryptic 5′ss; A, authentic 5′ss.
[b]Considering positive distances for C within introns, and negative distances for C within exons.
[c]Associated with internal exons only.

Table 2

Category	Average	SD	
A	82.96	6.31	
M	65.59	8.71	
C	72.38	8.53	
P	68.10	4.49	
AS	78.26	9.08	
Comparisons by category[a]			
Pairs	Result	*P*	
A versus M	A > M	$<10^{-20}$	
A versus C	A > C	$<10^{-10}$	
C versus M	C > M	$<10^{-5}$	
C versus P	C > P	0.024	
A versus P	A > P	$<10^{-13}$	
A versus AS	A > AS	$<10^{-3}$	
AS versus C	AS > C	$<10^{-5}$	
Pairwise comparisons[b]			
Pairs	Average difference[c]	SD	*P*
A–M	16.40	9.20	$<10^{-4}$
A–C	11.22	11.78	$<10^{-3}$
C–M	7.98	13.60	$<10^{-2}$
C–P	4.28	10.71	<0.19
A–P	15.87	8.00	$<10^{-2}$

A, authentic 5′ss; C, cryptic 5′ss; M, mutant 5′ss; P, pseudo 5′ss; AS, alternative 5′ss. Gradation of 5′ss efficiency: A > AS > C > M, C ~ P.

[a]*P*-values obtained using the Mann–Whitney rank test.

[b]*P*-values obtained using the Wilcoxon signed rank test.

[c]Average of the differences between each pair of 5′ss.

Table 3

	MAXENT[a]	MDD[b]	MM[c]	S&S[d]	NN[e]	ΔG^f
Average (A)	8.16	12.45	7.78	82.96	0.84	–10.8
SD (A)	1.94	1.84	1.86	6.31	0.24	2
Average (C)	3.18	7.99	4.12	72.38	0.33	–8.6
SD (C)	4.89	3.65	3.45	8.53	0.36	2.9
Average (A–C)	4.98	4.46	3.66	11.22	0.52	–2.4
SD (A–C)	5.61	4.23	4.37	11.78	0.39	3.9
Exceptions[g,h]						
ABCD1 exon 1	+					+
AR exon 4						+
BRCA1 exon 5				+		
BRCA1 exon 16						+
COL1A1 exon 8 (1)	+	+	+	+	+	+
COL1A1 exon 8 (2)	+	+	+	+	+	+
COL3A1 exon 42						+
COL7A1 exon 3			+			
CYP19 exon 6	+	+	+	+		+
CYP27A1 exon 6				+		+
DMD exon 64			+	+		
FAH exon 12						+
GHV exon 2	+		+		+	+
HBA2 exon 1					+	
HBB exon 1 (2)				+		
HEXA exon 9	+	+		+		+
HPRT1 exon 1						+
HPRT1 exon 5	+	+	+	+	+	+
IDS exon 7	+	+	+	+		
ITGB2 exon 7 (1)						+
ITGB2 exon 7 (2)						+
ITGB3 exon 4				+		+
LPL exon 2 (3)						+
LPL exon 2 (4)						+
NF1 exon 28						+
NF2 exon 7 (2)					+	
PGK1 exon 4						+
PTEN exon 7			+	+		
PYGM exon 14				+		+
Total	8	6	9	13	6	21

A, authentic 5′ss; C, cryptic 5′ss.
[a]Maximum entropy model.
[b]Maximum dependence decomposition model.
[c]First-order Markov model.
[d]Shapiro and Senapathy matrix.
[e]Neural network.
[f]Stability of the RNA duplex between the 5′ss and the U1 snRNA 5′ terminus.
[g]The exceptions (+) are those in which the cryptic 5′ss has a higher score than the corresponding authentic 5′ss.
[h]If there is more than one cryptic 5′ss for a particular exon, it is indicated by the number in parenthesis, in the order shown in Table 1.

Table 4

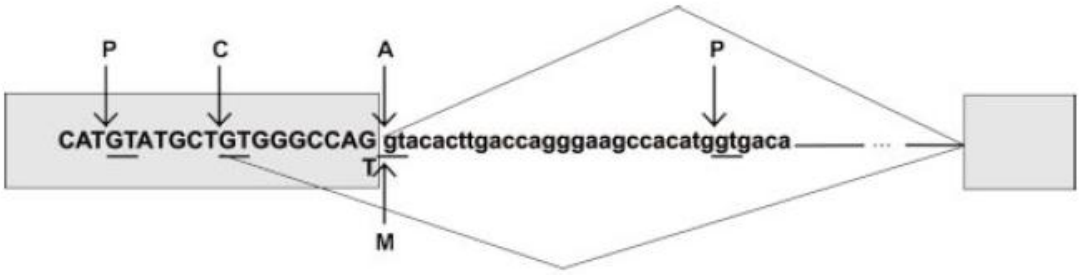
P
C
A
P
CATGTATGCTGTGGGCCAG gtacacttgaccagggaagccacatggtgaca
T
M

Fig. 1

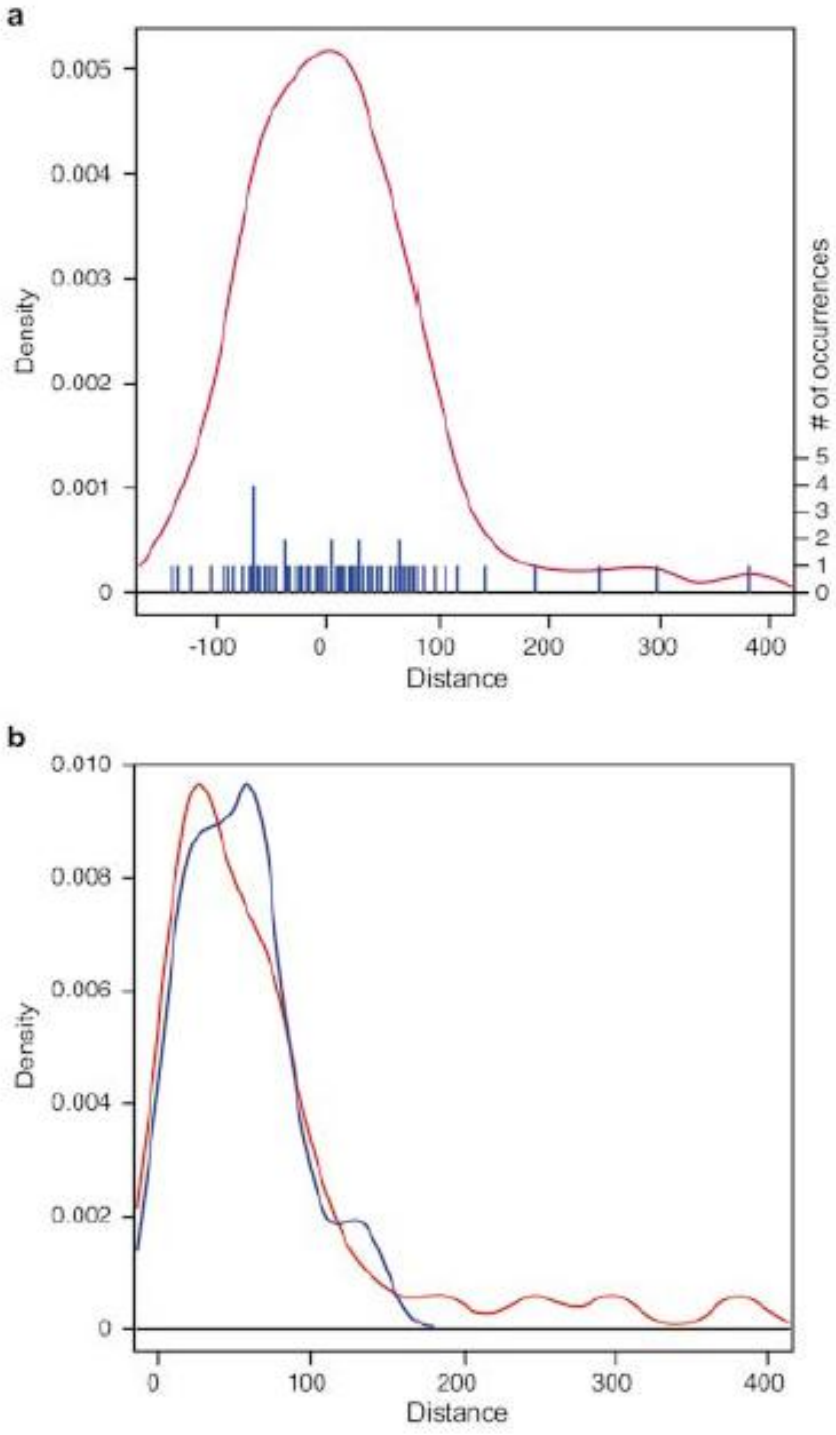
a
0.005
0.004
0.003
0.002
0.001
0
Density
# of occurrences
5
4
3
2
1
0
-100
0
100
200
300
400
Distance
b
0.010
0.008
0.006
0.004
0.002
0
Density
0
100
200
300
400
Distance

Fig. 2

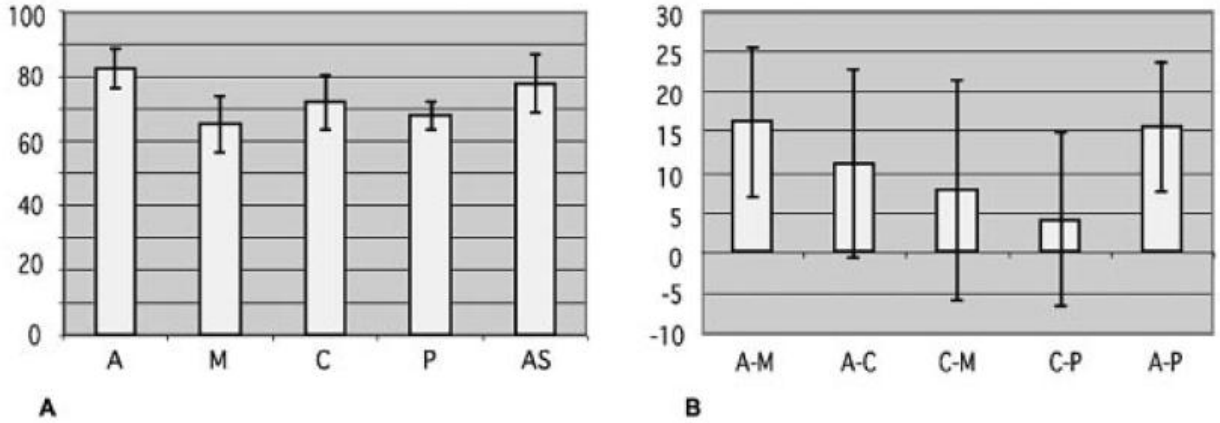

Fig. 3

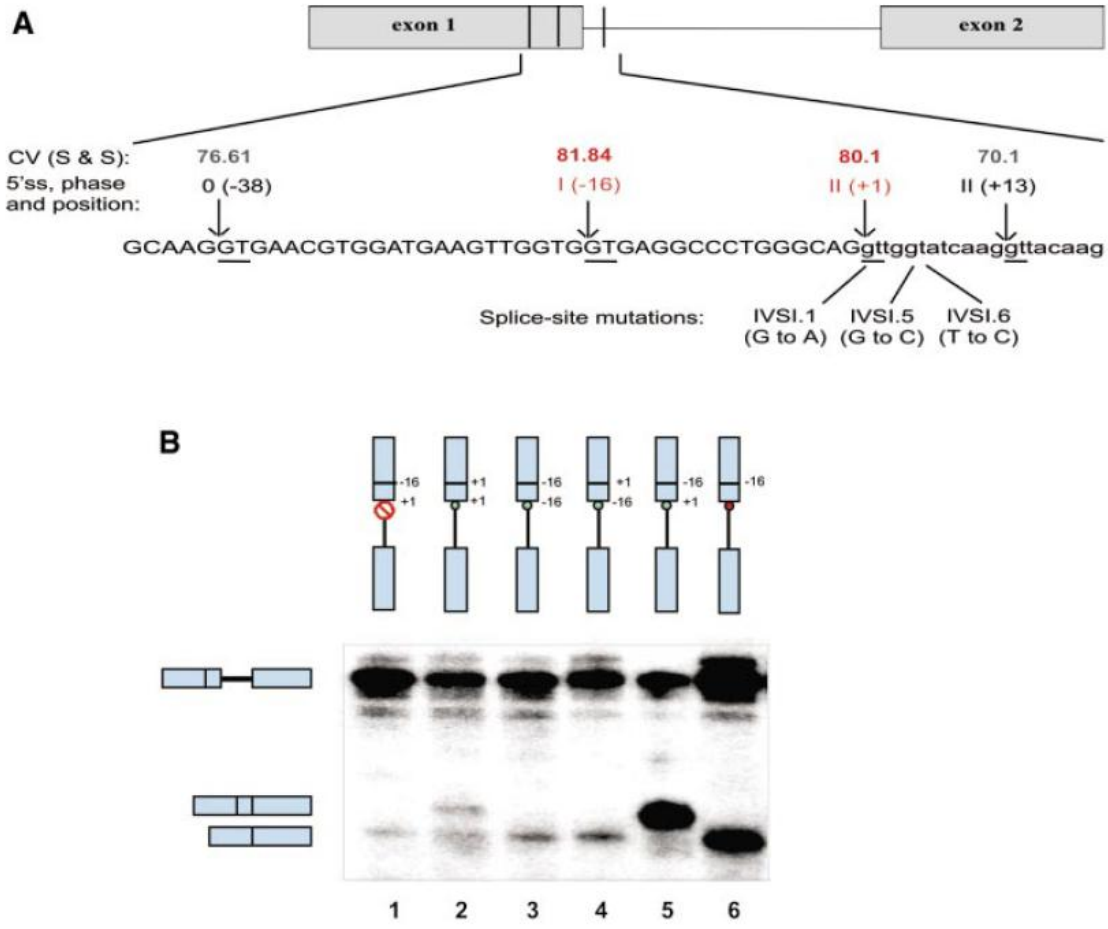
A
exon 1
exon 2
CV (S & S):
5'ss, phase and position:
76.61
0 (-38)
81.84
I (-16)
80.1
II (+1)
70.1
II (+13)
GCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGgttggtatcaaggttacaag
Splice-site mutations:
IVSI.1 (G to A)
IVSI.5 (G to C)
IVSI.6 (T to C)
B
-16 +1
+1 +1
-16 -16
+1 -16
-16 +1
-16
1
2
3
4
5
6

Fig. 4