Widespread recognition of 5' splice sites by noncanonical base-pairing to U1 snRNA involving bulged nucleotides

Xavier Roca,^{1,2} Martin Akerman,¹ Hans Gaus,³ Andrés Berdeja,³ C. Frank Bennett,³ and Adrian R. Krainer^{1,4}

¹Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 11724, USA; ²School of Biological Sciences, Division of Molecular Genetics and Cell Biology, Nanyang Technological University, 637551 Singapore; ³Isis Pharmaceuticals, Carlsbad, California 92008, USA

An established paradigm in pre-mRNA splicing is the recognition of the 5' splice site (5'ss) by canonical base-pairing to the 5' end of U1 small nuclear RNA (snRNA). We recently reported that a small subset of 5'ss base-pair to U1 in an alternate register that is shifted by 1 nucleotide. Using genetic suppression experiments in human cells, we now demonstrate that many other 5'ss are recognized via noncanonical base-pairing registers involving bulged nucleotides on either the 5'ss or U1 RNA strand, which we term "bulge registers." By combining experimental evidence with transcriptome-wide free-energy calculations of 5'ss/U1 base-pairing, we estimate that 10,248 5'ss (~5% of human 5'ss) in 6577 genes use bulge registers. Several of these 5'ss occur in genes with mutations causing genetic diseases and are often associated with alternative splicing. These results call for a redefinition of an essential element for gene expression that incorporates these registers, with important implications for the molecular classification of splicing mutations and for alternative splicing.

[**Keywords:** 5' splice site; U1 small nuclear RNA; base-pairing register; bulged nucleotide; splicing mutation; alternative splicing]

Supplemental material is available for this article.

Received February 20, 2012; revised version accepted April 10, 2012.

Pre-mRNA splicing is an essential processing step for the expression of ~90% of protein-coding human genes and relies on conserved sequence elements at both ends of introns, termed splice sites (Sheth et al. 2006; Wahl et al. 2009). These elements are highly diverse, considering that thousands of different sequences act as naturally occurring splice sites in the human transcriptome (Sahashi et al. 2007; Roca and Krainer 2009). The characterization of these sequence elements and the factors that recognize them has been essential for predicting exons in new genes, for the study of alternative splicing (Nilsen and Graveley 2010), and for classifying mutations in these elements that cause human genetic diseases (Buratti et al. 2007).

Typically, the strength of a splice site (or splice site score) is estimated by algorithms that measure its concordance to matrices built using large collections of splice sites (Senapathy et al. 1990; Brunak et al. 1991; Yeo and Burge 2004; Sahashi et al. 2007; Hartmann et al. 2008). These methods implicitly assume that all of the sequences used to build the matrix are recognized by the same mechanism. However, there are cases in which the

splice site score does not reflect the strength of the splice site determined experimentally (Roca and Krainer 2009), highlighting the limitations of these tools. Furthermore, the recognition mechanisms for many splice sites predicted to be weak are poorly understood.

Splicing of >99% of pre-mRNA introns is catalyzed by the major spliceosome, a dynamic macromolecular machine composed of five small nuclear RNAs (snRNAs) and associated polypeptides, plus many other protein factors (Wahl et al. 2009). The U1 small nuclear ribonucleoprotein particle (snRNP), comprising the U1 snRNA and 10 polypeptides (Pomeranz Krummel et al. 2009), is the main component for early 5' splice site (5'ss) recognition by the major or U2-type spliceosome. The vast majority of such introns (>99%) belong to the GT-AG (or GU-AG) category, as defined by their intronic terminal dinucleotides (Sheth et al. 2006). For >30 years, it has been firmly established that 5'ss are recognized by base-pairing to the 5' end of U1 snRNA in a canonical register, defined as +1G at the 5'ss (the first intronic nucleotide) base-pairing to C8 of U1 (the eighth nucleotide of U1) (Fig. 1A, left; Lerner et al. 1980; Rogers and Wall 1980; Zhuang and Weiner 1986; Séraphin et al. 1988; Siliciano and Guthrie 1988). Thus, the 5'ss element spans the last 3 nucleotides (nt) of the exon and the first 8 nt of the intron,

⁴Corresponding author.
E-mail krainer@cshl.edu.

Article is online at <http://www.genesdev.org/cgi/doi/10.1101/gad.190173.112>.

Figure 1. Bulge 5'ss/U1 base-pairing. (A) Diagram of two base-pairing registers between consensus (left) and atypical (right) 5'ss and the 5' end of U1 snRNA. 5'ss positions are numbered; base-pairing at +7 and +8 can contribute to splicing (Hartmann et al. 2008), although these positions do not show conservation in 5'ss compilations (Sheth et al. 2006). Consensus nucleotides are shown in red in all figures. (Ψ) Pseudouridine; (dot) 2,2,7-trimethylguanosine cap at the 5' end of U1; (box) upstream exon; (line) intron. Base pairs in the canonical (C) or shifted (S) register are indicated by vertical lines. (B) Diagram of three base-pairing possibilities for the *ITPR2* intron 18 5'ss: C (6 bp), S (9 bp), and bulge +2/+3 (B) registers (11 bp). (C) Mutational analysis for the B register. (Top) Schematic of the *SMN1/2* minigenes, indicating the test 5'ss replacing the original 5'ss and the mutations introduced. Radioactive RT-PCR analysis is shown for *SMN1* (top) and *SMN2* (bottom) minigenes. The identity of the various spliced mRNAs is indicated on the left; from large to small, the bands correspond to exon 7 inclusion, exon 7 skipping, and exon 7 skipping with activation of a cryptic 5'ss at position -50 in exon 6. In all figures, the mean percentage and SD of exon 7 inclusion are shown below each lane. (D) Transition between shifted and bulge base-pairing. The nucleotides at positions -3 to -1 of the test 5'ss are indicated above each lane. (Lane 1) Atypical 5'ss (shifted register). (Lane 5) Bulge +2/+3 5'ss. Schematics for these 5'ss are in A and B. (Lanes 2–4) Intermediates between the atypical and the bulge +2/+3 5'ss. (E) Suppressor U1 experiments. The schematic shows, for the +5C and +6C mutant 5'ss, the suppressor U1s that restore base-pairing in the canonical (C; G4, or G3, respectively) or bulge (B; G5, or G4, respectively) register. The bulged nucleotide is depicted opposite to a gap in the bottom strand. The 5'ss mutation and suppressor U1s are indicated above each lane. (—) No suppressor.

establishing a maximum of 11 base pairs (bp) to U1—although the contribution of the seventh and eighth nucleotides in the intron, which are much more variable, appears to depend on the species (Staley and Guthrie 1999; Freund et al. 2005). Later in spliceosome assembly, U1 is replaced by U6 snRNA, which forms a few base pairs to the 5'ss and is likely involved in catalysis (Wassarman and Steitz 1992; Kandels-Lewis and Séraphin 1993; Lesser and Guthrie 1993). In a handful of documented cases, U1 base-pairs at some distance from the 5'ss, and the cleavage site depends on subsequent U6

base-pairing (Cohen et al. 1994; Hwang and Cohen 1996; Brackenridge et al. 2003). There is also an example of a natural human U2-type intron whose splicing appears to be U1 snRNA-independent (Fukumura et al. 2009).

Two minor categories of U2-type splice sites have been known for a long time: GC-AG 5'ss (0.9%) and very rare AT-AC 5'ss (only 15 introns in the human genome) (Sheth et al. 2006). These 5'ss conform to consensus motifs very similar to the major U2-type GT-AG 5'ss and are recognized by analogous mechanisms. We recently showed that restoration of base-pairing to both U1 and U6 is

essential to rescue recognition of a mutant AT 5'ss that causes aberrant splicing and myotonia (Kubota et al. 2011). U12-type introns are spliced by the minor spliceosome and are very rare as well (0.36%) (Will and Lührmann 2005).

We recently showed that a small subset of GT-AG 5'ss, which we termed atypical 5'ss, is recognized by a base-pairing register with U1 that is shifted by 1 nt (+1G base-pairs to U1 C9 instead of C8) without changing the actual exon–intron boundary or the sequence of the spliced mRNA (Fig. 1A, right; Roca and Krainer 2009). In budding yeast, mutational analysis led to the suggestion that the non-canonical *HOP2* 5'ss is recognized by a base-pairing register involving a bulged nucleotide (Leu and Roeder 1999). A bulge in a strand of RNA (or DNA) duplex is defined as a nucleotide (or more) that is not opposed by any nucleotide on the other strand. Here we present extensive experimental evidence for multiple base-pairing registers between human 5'ss and U1, with bulged nucleotides on either RNA strand, and estimate that ~5% of all 5'ss—present in ~40% of human genes—use one of these noncanonical registers.

Results

A bulge 5'ss/U1 base-pairing register

Inspection of well-annotated 5'ss in the human transcriptome revealed a group of bona fide 5'ss sequences similar to the atypical 5'ss recognized by shifted base-pairing (Roca and Krainer 2009) but differing at exonic positions –3 to –1 by having consensus nucleotides (Fig. 1B, representative 5'ss sequence from intron 18 of the *ITPR2* gene). These 5'ss and U1 can establish 6 or 9 bp in the canonical or shifted registers, respectively. Because this sequence resembles the consensus 5'ss but with a U insertion at position +3, up to 11 bp could be formed by bulging out the U at either position +2 or +3 of the 5'ss. We tested whether this type of 5'ss is actually recognized by a “bulge +2/+3” register. Note that the proposed mechanism does not involve a shift in the sites of *trans*-esterification chemistry; i.e., the sequence of the spliced mRNA does not change.

We analyzed the test 5'ss sequences in the context of Survival of motor neuron minigenes (*SMN1/2*) comprising a human genomic fragment from exon 6 to exon 8. *SMN1* and *SMN2* paralog pre-mRNAs show different extents of exon 7 inclusion due to a single-nucleotide difference at its sixth position: Exon 7 is completely included in *SMN1* and predominantly skipped in *SMN2* (Lorson et al. 1999). We replaced the natural 5'ss of this exon in the minigenes by the *ITPR2* 5'ss to test the bulge +2/+3 hypothesis and introduced point mutations to disrupt different base pairs (Fig. 1C).

We transiently transfected HeLa cells with the different *SMN1/2* minigenes and analyzed the extent of exon 7 inclusion—reflecting recognition of the test 5'ss—by radioactive RT-PCR. The wild-type test 5'ss resulted in nearly complete inclusion of exon 7 in either *SMN1/2* context (Fig. 1C, lane 1, top and bottom). This observation indicates that the *ITPR2* 5'ss is stronger than the *SMN1/2* 5'ss. Point mutations on either side of the predicted bulged

nucleotide partially or completely disrupted exon inclusion in either context, consistent with loss of base pairs to U1 (Fig. 1C, lanes 2–4). Combinations of these mutations exacerbated the disruption of exon 7 inclusion (Fig. 1C, lanes 5,6). The tested mutations disrupted base pairs in different registers, yet the bulge +2/+3 register was the only one severely affected by all of the mutations that resulted in exon skipping (the –1C mutation disrupts a very weak G-U wobble terminal base pair in the shifted register). This observation suggests that the test 5'ss is recognized via this bulge register.

Next, we gradually converted an atypical 5'ss recognized by shifted base-pairing (from intron 8 of *GTF2H1*) (Roca and Krainer 2009) to a test 5'ss for the bulge +2/+3 register by introducing mutations at positions –3 to –1. Whereas the atypical and the bulge +2/+3 5'ss resulted in nearly complete *SMN2* exon 7 inclusion (Fig. 1D, lanes 1,5, bottom), two of the intermediate mutants showed reduced inclusion (Fig. 1D, lanes 3,4, bottom), suggesting that these 5'ss cannot base-pair efficiently in either register. Furthermore, direct sequencing of the RT-PCR products from the various constructs confirmed that splicing only occurred at the GU boundary and not at the noncanonical UU 1 nt downstream (Supplemental Fig. S1A). This is in contrast to what was reported for a similar mutant 5'ss in *FANCC* (AUG/UUAAGUAG, where “/” is the mapped exon/intron boundary) (Hartmann et al. 2010), and consistent with a mutation that creates a 5'ss in *SMARCD1* (ACU/GUUAAGUAC), associated with autosomal-dominant adermatoglyphia (lack of fingerprints) (Nousbeck et al. 2011).

We then used suppressor (or shift) U1 snRNA experiments to determine whether the test 5'ss is recognized via the bulge +2/+3 register (Fig. 1E). This powerful allele-specific suppression strategy has been used to prove base-pairing interactions between many 5'ss sequences and U1 in the canonical (Zhuang and Weiner 1986; Séraphin et al. 1988; Siliciano and Guthrie 1988) or shifted (Roca and Krainer 2009) register. We cotransfected the mutant *SMN1/2* minigenes along with U1 snRNA expression plasmids carrying compensatory mutations that restore base-pairing in either the canonical or the bulge +2/+3 register. In *SMN2*, the loss of exon 7 inclusion due to the +5C mutation (Fig. 1E, lane 3, bottom) was partially rescued by U1 snRNAs with the bulge (B, G5) but not the canonical (C, G4) suppressor mutation (Fig. 1E, lane 5 vs. 4). Similarly, the +6C mutation (Fig. 1E, lane 6) was weakly rescued by the bulge (B, G4) but not the canonical (C, G3) suppressor in *SMN1* (Fig. 1E, lane 8 vs. 7). Even though not all suppressor U1s are effective in such experiments, the suppressors in the bulge but not in the canonical register rescued recognition of the test 5'ss, thereby demonstrating that this 5'ss base-pairs to U1 in the bulge +2/+3 register.

Registers with bulges at other 5'ss positions

Next, we addressed the generality of bulge base-pairing; i.e., whether other natural 5'ss can be recognized by bulging out nucleotides at other positions. We found human authentic 5'ss sequences nearly identical to the

consensus, but with an insertion at position +4 or at +5 (Fig. 2A). Such 5'ss are predicted to have 6 or 7 bp to U1 in the canonical register, which can be extended to 11 bp if a nucleotide at +4 or +5 is bulged, respectively. Thus, the extra base pairs in the bulge register would provide an energetic advantage over the canonical register. For the experiments below, the 5'ss base-pair poorly to U1 in the shifted register (Supplemental Fig. S1B), so for simplicity, comparisons are only made between the canonical and bulge registers.

We tested the bulge +4 and bulge +5 hypotheses in the *SMN1/2* minigenes by mutational analysis and suppressor U1 experiments. The wild-type bulge +4 (from intron 6 of *MORC4*) and bulge +5 (from intron 8 of *PARD3*) 5'ss resulted in complete exon 7 inclusion with both minigenes (Fig. 2B, lanes 1,7). Among the mutations that disrupt base pairs on either side of the bulge, only the -1C mutants reduced exon 7 inclusion (Fig. 2B, lanes 2-4,8-10), indicating that these test 5'ss are quite strong. The +6C mutation, which disrupts a base pair only in the bulge registers, had no effect by itself (Fig. 2B, lanes 4,10). Nevertheless, +6C cooperated with -2C to reduce exon 7 inclusion and exacerbated the extent of skipping by the -1C mutation (Fig. 2B, lanes 6,11,12). The -2C mutation introduces a base pair in the shifted register, yet it is disruptive in combination with +6C, consistent with these 5'ss not being recognized by shifted base-pairing.

The results with the combined mutations suggest that the +6 nucleotide base-pairs to U1, consistent with the bulge +4 or bulge +5 registers.

Suppressor U1 experiments provided additional evidence for the bulge +4 or bulge +5 registers. The loss of exon 7 inclusion upon the -2C+6C mutation in the bulge +4 5'ss was rescued by the suppressor in the bulge but not in the canonical register (Fig. 2C,D [lane 9 vs. 7,8]). The rescue of the bulge +5 5'ss with the -2C+6C mutation was higher with the bulge than with the canonical suppressor in *SMN2* (Fig. 2D, lane 18 vs. 16,17). We conclude that the test 5'ss are preferentially if not exclusively recognized by the bulge +4 or bulge +5 registers, respectively.

We also tested a more complex case involving a 5'ss sequence predicted to base-pair to U1 by bulging out an adenosine at either position +3, +4, or +5 (from intron 2 of *JMJD6*). Mutational analyses and suppressor U1 experiments demonstrated that such 5'ss are indeed preferentially recognized by the bulge +3/+4/+5 register (Supplemental Fig. S2).

Registers with bulges at the 5' end of U1

Our data show that the 5'ss/U1 helix can tolerate bulges on the 5'ss strand to support productive splicing. Other 5'ss (example from intron 5 of *PRKD1*) can form 7 bp to U1 in the canonical register, which can be increased to

Figure 2. Base-pairing registers with bulges at positions +4 or +5. (A) Diagrams of the bulge +4 and bulge +5 base-pairing registers, as in Figure 1, A and B. The bulged nucleotide is numbered. (B) Mutational analyses in *SMN1* (top panels) and *SMN2* (bottom panels) minigene contexts. In B and D, RT-PCR products, mean exon inclusion, and SD are indicated as in Figure 1C. The additional bands are described in Supplemental Figure S2D. (C) Schematic of canonical (C; G3G10) and bulge (B; G4G10) suppressor U1s for the -2C+6C mutation in the bulge +4 register. In these experiments, suppressor U1s carry compensatory mutations for both 5'ss mutations. Other suppressors are shown in Supplemental Figure S1C. (D) Suppressor U1 snRNA experiments. The mutant test 5'ss and suppressor U1 are indicated at the top. Suppressor U1s in the canonical register were used in lanes 4, 8, 13, and 17. Suppressors in the bulge registers were used in lanes 5, 9, 14, and 18.

10 bp by bulging out one of the two consecutive pseudouridines (Ψ , a uridine isomer) at positions 5 and 6 of the 5' end of U1 (Fig. 3A). We tested the bulge Ψ hypothesis by mutational and suppressor U1 analyses. The wild-type test 5'ss supported efficient *SMN1/2* exon 7 inclusion (Fig. 3B, lane 1), which was partially or totally disrupted by point mutations on either side of the bulge ($-1C$ or $+5C$), independently or combined (Fig. 3B, lanes 2,3,6). The suppressor U1s in the bulge but not the canonical register rescued exon inclusion disrupted by the $+5C$ and $-1C+5C$ mutations (Fig. 3B, lane 4 vs. 3,5 in both *SMN1* and *SMN2*, and to a lesser extent, lane 9 vs. 6,10 in *SMN1*), demonstrating that these 5'ss base-pair to U1 in the bulge Ψ register.

Finally, we used U1-specific RNA decoys (Roca and Krainer 2009) to confirm that all of the tested 5'ss in the *SMN1/2* contexts are indeed recognized by U1 and not by other U1-like snRNAs (Kyriakopoulou et al. 2006). These vector-driven RNA decoys sequester endogenous U1, resulting in the loss of 5'ss recognition and, consequently, in *SMN1/2* exon 7 skipping. We detected exon 7 skipping upon cotransfection of the U1 decoy but not a control decoy with a mismatch, confirming that the recognition of the test 5'ss is mediated by U1 snRNA (Supplemental Fig. S3).

Bulge registers in their natural context

The above results demonstrate that bulge 5'ss/U1 base-pairing can occur in the context of mutant *SMN1/2* minigenes; next, we tested whether these registers are used in a natural context. We selected five representative examples from our genomic searches (see below) and constructed three-exon/two-intron minigenes with the test 5'ss as part of the middle exon. The registers tested included bulge +3 (*F5* minigene), bulge +4 (*RPGR* minigene), bulge +5 (*MLH1* minigene), bulge +3/+4/+5 (*RB1* minigene), and bulge Ψ (*MDM2* minigene).

We then performed mutational analyses and suppressor U1 experiments (Fig. 4; Supplemental Figs. S4, S5). 5'ss

point mutation alleles of *MLH1* and *RPGR*—associated with colorectal cancer and retinitis pigmentosa, respectively—resulted in reduced exon inclusion compared with wild-type minigenes (Fig. 4A [lane 2 vs. 1], B [lanes 2,3 vs. 1]). Mutations that disrupted the bulge also resulted in skipping of the middle exon (Fig. 4A [lanes 5,6], B [lanes 4,7,8,11]). Suppressor U1s in the bulge but not the canonical register rescued correct splicing (Fig. 4A [lane 8 vs. 6,7], B [lane 10 vs. 8,9]). Analogous experiments with the *F5*, *RB1*, and *MDM2* minigenes gave similar results (Supplemental Fig. S4). We conclude that the five representative 5'ss are recognized via the predicted bulge base-pairing registers.

Genomic analyses of bulge 5'ss/U1 base-pairing

Our experiments demonstrate that 5'ss positions +2 to +5 and the Ψ at U1 position 5 or 6 can be bulged in certain 5'ss/U1 RNA helices. We next addressed the following issues about bulge 5'ss/U1 base-pairing: (1) the number of authentic human 5'ss recognized via these noncanonical registers, (2) whether other bulge registers occur, (3) the energetic advantage of the bulge over the canonical register, (4) the implications for disease-causing mutations and single-nucleotide polymorphisms (SNPs) at 5'ss, and (5) the involvement in alternative splicing.

We used a bioinformatics approach to address these questions. We generated a data set of 201,541 human authentic (well-annotated) 5'ss sequences, including 15 nt on each side of the exon/intron junction, for both constitutive and alternative exons. For each sequence and the 5' end of U1, we estimated the base-pairing register and minimum free energy (ΔG_1 , in kilocalories per mole) using UNAFold hybrid (Markham and Zuker 2008). In a second run of UNAFold, we obtained the free energies for these 5'ss by forcing canonical base-pairing (ΔG_2) (see the Materials and Methods). UNAFold predicted a total of 10,248 5'ss (5.1%) that base-paired to U1 using a bulge register, which we termed "bulge 5'ss." The bulge 5'ss occurred in 6577 genes, amounting to 41% of the total of 15,894 genes in our data set. The energetic advantage of the bulge over the canonical register ($\Delta\Delta G$) was calculated as the difference between ΔG_1 and ΔG_2 and ranged from -0.1 to -4.9 kcal/mol. To identify the subset of bulge 5'ss in which the bulge register confers a substantial energetic advantage, we selected those cases with a $\Delta\Delta G \leq -1$, based on our SNP analysis (see the Results below and the Materials and Methods).

The bulge 5'ss set contains 6940 5'ss (3.4% of all 5'ss) using a base-pairing register with one bulged nucleotide (Table 1). This includes the registers for which we obtained the experimental evidence above and new ones, including a bulge at position -1 (see the Discussion), a bulge at either position +3/+4 or +4/+5, and a bulge at GC 5'ss, including the C at position +2.

In addition to single-nucleotide bulges, UNAFold also predicted many registers involving longer bulges at the 5'ss, ranging from 2 to 8 nt (Table 1). These registers have not been experimentally tested yet, but they would account for the recognition of 3294 5'ss (1.6%). The number of candidates and the $\Delta\Delta G$ became smaller as

Figure 3. The bulge Ψ base-pairing register. (A) Schematic for the bulge Ψ register and for suppressor U1s for the +5C mutation [other suppressors are shown in Supplemental Fig. S1D]. The U1 G4 and U1 G3 suppressors restore base-pairing in the canonical (C) or bulge (B) register, respectively. (B) Mutational analyses and suppressor U1 experiments demonstrate the bulge Ψ register. For the $-1C+5C$ mutation, the G3G9 and G4G9 suppressors restore 2 bp in the bulge and canonical registers, respectively. RT-PCR products, mean exon inclusion, and SD are indicated as in Figure 1C.

Figure 4. Bulge base-pairing registers in their natural context. (A) Bulge +4. (Top) Schematic of the *RPGR* minigene, indicating the bulge base-pairing register and some of the point mutations introduced. (Bottom) Mutational analysis and suppressor U1 experiments. The mutant 5'ss and suppressor U1 are indicated above each lane. RT-PCR products are indicated on the left and correspond to inclusion (1) and skipping (2) of the middle exon. All point mutations disrupted exon 4 inclusion (lanes 2–6) and the U1 suppressor for the +7C mutation in the bulge (G3) but not the canonical (G2) register rescued inclusion (lanes 6–8). (A,B) Mean exon inclusion values and SD are shown as in Figure 1C, and diagrams for the various U1 suppressors are shown in

Supplemental Figure S4A. (B) Bulge +5. Schematic of the *MLH1* minigene, with labels as in A. RT-PCR products as in A. All point mutations disrupted exon 10 inclusion (lanes 2–4,7,8), and the +7C mutation was rescued by the bulge (lane 10) and not by the canonical U1 suppressor (lane 9).

the bulge length increased (Supplemental Fig. S6). Finally, rare registers included a bulge of both pseudouridines in U1 (only conferring a marginal $\Delta\Delta G$ of -0.1), and two bulges at distant 5'ss positions.

By searching the splice site database SpliceRack (Sheth et al. 2006), examples of bona fide 5'ss recognized by our experimentally validated bulge registers were also found in *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Arabidopsis thaliana* (Supplemental Table S1). This conservation suggests that bulge 5'ss/U1 base-pairing is a general and phylogenetically widespread phenomenon.

To summarize our predictions, from the 10,248 bulge 5'ss (Table 1), a total of 3016 5'ss (1.5% of all 5'ss) use a single-nucleotide bulge register consistent with the experimental evidence in our model minigenes. Furthermore, 5877 5'ss (2.9%) have a bulge register that confers a substantial energetic advantage ($\Delta\Delta G \leq -1$) over the canonical register. We conclude that bulge 5'ss/U1 base-pairing is a frequent phenomenon, affecting the recognition of 1.5%–5.1% of all human 5'ss.

Tm measurements support bulge base-pairing

We carried out oligonucleotide duplex melting experiments to further test the formation of bulge 5'ss/U1 helices and confirm the reliability of the UNAFold predictions. For the 5'ss strand, we designed a set of 45 oligoribonucleotides, which included sequences for calibration, and 18 test pairs for several bulge base-pairing registers (Table 2). The calibration oligonucleotides included two consensus 5'ss (consensus and consensus-long, with 9 and 11 bp to U1, respectively), five related sequences with mutations (consensus mut), and two sequences with very poor match to the consensus (low). The test pairs included a "bulge" 5'ss sequence (B) predicted to base-pair to U1 with bulged nucleotides, and a control sequence (C) with a single-nucleotide difference predicted to abolish the bulge (Supplemental Fig. S7). For the complementary U1 strand, we synthesized two U1 oligoribonucleotides: one with an

unmodified 11-nt 5' end, and the other with the two Ψs plus the 2'-O-methyl nucleotides at the first two positions, characteristic of U1 snRNA.

For each oligonucleotide duplex (5'ss and U1), we obtained melting curves and experimentally derived T_m and ΔG (ΔG_{exp}) values (Table 2; see the Materials and Methods). Overall, the ΔG_{exp} correlated very well with the predicted ΔG_1 ($R = 0.92$ for the modified U1 oligonucleotide). The T_m values for the helices with the modified U1 oligonucleotide were, on average, 3.4°C ($\pm 1.6^\circ\text{C}$) higher than those for the unmodified one, which is in the range of that previously described for a consensus 5'ss (2°C) (Hall and McLaughlin 1991). Analysis of a partially modified oligonucleotide with only the two Ψs indicated that ribose methylation did not contribute to the enhanced stability (data not shown). Many 5'ss oligonucleotides did not show a cooperative hyperchromic shift with increasing temperature, suggesting that these RNAs did not base-pair to U1 under these experimental conditions (Table 2; Supplemental Fig. S7).

Interestingly, the seven pairs of 5'ss RNAs that showed cooperative transitions had a consistent trend: The B oligonucleotide had a higher T_m and more stable ΔG_{exp} than the C oligonucleotide, such as for the bulge +2/+3 pair (Table 2; Supplemental Fig. S7). From these oligonucleotide pairs and modified U1, we derived a mean difference in ΔG_{exp} of -1.2 kcal/mol (± 0.8). In other cases, the B but not the C oligonucleotide showed a cooperative transition and ΔG_{exp} , such as for the bulge +5 pair. This trend was seen in 11 and eight pairs with the unmodified or modified U1 oligonucleotide, respectively. This finding also indicates that the bulge oligonucleotides base-paired more stably to U1 than the controls.

In summary, the 15 informative pairs showed that the bulge oligonucleotide bound to the modified U1 oligonucleotide with higher affinity than the control (Table 2). These thermodynamic data strongly support the notion that bulged nucleotides can occur in the context of 5'ss/U1 helices.

Table 1. Numbers and distribution of bulge 5'ss

	Position	Number	$\Delta\Delta G^a$	Number $\Delta\Delta G \leq -1^b$	Exp ^c
Total		10,248	-1.20	5877	
Bulge 1	All	6940	-1.29	4373	
	Ψ^d	616	-0.94	186	Yes
	-1	2913	-1.24	1744	No
	+2/+3 ^e	68	-1.92	48	Yes
	+2/+3/+4/+5 ^f	10	-1.17	4	No
	+3	51	-1.40	31	Yes
	+3/+4	348	-1.61	264	No ^g
	+3/+4/+5	631	-1.21	419	Yes
	+4	1,115	-1.46	753	Yes
	+4/+5	653	-1.48	456	No ^g
	+5	535	-1.25	468	Yes
Bulge 2	Ψ^h	6	-0.10	0	No
	5'ss ⁱ	1530	-1.15	810	No
Bulge 3 ⁱ		925	-1.06	507	No
Bulge 4		430	-0.76	130	No
Bulge 5		241	-0.53	31	No
Bulge 6		141	-0.41	12	No
Bulge 7		25	-0.47	6	No
Bulge 8		2	-0.40	0	No
Bulge multiple ^j		8	-3.31	8	No

^a($\Delta\Delta G$) $\Delta G_1 - \Delta G_2$; mean energetic advantage of the bulge over the canonical register.

^b($\Delta\Delta G \leq -1$) Cases in which the bulge register confers a substantial advantage over the canonical register.

^c(Exp) Experimental evidence for the register.

^dPseudouridine at either position 5 or 6 of the 5' end of U1 snRNA.

^e(+2/+3) Either 5'ss position +2 or +3 is bulged.

^fGC 5'ss bulging out one or more positions between the C at +2 and positions +3 to +5, not matching the other registers.

^gCases related to those tested because they involve the same 5'ss positions.

^hBulge of both Ψ s.

ⁱBulge 2 or Bulge 3, 2-nt or 3-nt bulge at the 5'ss.

^j5'ss with two bulges at two separate positions.

Implications for disease mutations and SNPs

We next asked whether the bulge 5'ss have previously documented mutations causing human genetic diseases or have SNPs. From a set of 581 disease-causing mutations at 5'ss (Roca et al. 2008), we found that 24 (4.1%) mapped to bulge 5'ss (Supplemental Table S2). As expected, the ΔG_1 values for the mutant 5'ss are substantially smaller than those for the corresponding wild-type 5'ss (mean difference of -1.7 kcal/mol). However, the mutation disrupted the bulge structure in only one case. Although a larger data set will be needed to establish statistical significance, these observations suggest that deleterious mutations at bulge 5'ss tend to disrupt a base pair not involved in the bulge.

From a set of 1,116 SNPs at human 5'ss (Roca et al. 2008), 57 (5.1%) mapped to bulge 5'ss (Supplemental Table S3). The ΔG_1 values for either 5'ss variant are similar (mean absolute difference [MAD] = -1.0 kcal/mol), consistent with a generally neutral effect of the SNP on 5'ss strength. In 34 cases, the bulge structure is

maintained by the SNP, resulting in small differences in ΔG_1 between the two 5'ss variants (MAD = -0.6 kcal/mol). In the remaining 23 cases, one allele base-pairs to U1 with bulged nucleotides, and the other allele base-pairs in a canonical register, resulting in a larger ΔG_1 difference (MAD = -1.7 kcal/mol). In eight out of 23 cases, the 5'ss variant with the bulge structure had a less stable ΔG_1 than the variant using the canonical register because the SNP introduces an extra base pair in the canonical register. For these eight SNPs, the difference in ΔG_1 between the two 5'ss variants would be -3.5 kcal/mol larger, on average, if the weaker allele did not base-pair in a bulge register, suggesting a compensatory effect of the bulge structure. These findings suggest that SNPs at bulge 5'ss have a low overall impact on 5'ss strength, even if the bulge structure is not conserved.

Involvement in alternative splicing

Finally, we investigated whether the bulge base-pairing registers tend to occur preferentially in 5'ss involved in alternative splicing. To this end, we subdivided the set of 5'ss into four categories: constitutive, alternative 5'ss (choice between at least two tandem 5'ss), alternative 3'ss (5'ss in exons with tandem 3'ss), and cassette exons (exons that are included or skipped) (Table 3). Compared with the canonical 5'ss, the bulge 5'ss were significantly enriched in the alternative 5'ss set (4.17% vs. 2.10%, $P < 10^{-28}$, Fisher's exact test) and cassette exon set (8.15% vs. 6.98%, $P < 10^{-5}$). These differences remained when we used the bulge 5'ss with a $\Delta\Delta G \leq -1$ (4.23% vs. 2.10% for alternative 5'ss, $P < 10^{-18}$).

Splice sites involved in alternative splicing are slightly weaker, on average, than constitutive splice sites (Itoh et al. 2004; Wang et al. 2005), and we confirmed this trend in our data set of canonical 5'ss (ΔG_1 for 5'ss in constitutive and alternative splicing are -7.88 and -7.53 kcal/mol, respectively, $P < 10^{-5}$, one-sample t -test). As the mean ΔG_2 for the bulge 5'ss is much less stable than that for the canonical 5'ss (-5.76 vs. -7.88 kcal/mol) (Table 3), the enrichment of the bulge 5'ss in alternative splicing events could be biased by the overall weakness of these 5'ss. Thus, we derived a subset of "weak canonical 5'ss" (see the Materials and Methods) with a mean ΔG_2 almost identical to that for the bulge 5'ss (-5.81 vs. -5.76). Comparison of the bulge 5'ss with the weak canonical 5'ss revealed that the enrichment for cassette exons essentially disappeared (8.15 vs. 7.90, $P < 0.14$, Fisher's exact test), but the enrichment for alternative 5'ss remained highly significant (4.17 vs. 2.62, $P < 10^{-12}$). This small yet statistically significant bias reveals that the bulge 5'ss, when compared with canonical 5'ss, are more frequently associated with alternative 5'ss events.

Discussion

We experimentally demonstrated that many 5'ss are recognized by U1 snRNA via base-pairing registers involving bulged nucleotides. We also estimated that 1.5%–

Table 2. Summary of oligonucleotide melting experiments

Oligonucleotide name	Sequence (5' to 3')	ΔG_1	Tm1 ^a	Tm2 ^b	ΔG_{exp} ^c
U1 unmodified	AUACUUACCUG				
U1 modified	A _m U _m ACΨΨACCUG				
Consensus Long	AACAGGUAAGUAUAA ^d	−13.4	51.0	55.4	−14.4
Consensus LongMut ^e	AACAGGUAACUAUAA	−8.8	33.4	37.5	−8.4
Consensus	AACAGGUAAGUCCAA	−11.0	46.9	49.8	−12.1
Consensus Mut1	AACACGUAAGUCCAA	−5.6	— ^f	—	—
Consensus Mut2	AAAAGGUAAGUCCAA	−9.7	39.5	46.4	−10.4
Consensus Mut3	AACAGAUAAAGUCCAA	−5.5	27.3	30.5	−6.7
Consensus Mut4	AACAGGCAAGUCCAA	−7.0	26.0	29.0	−6.4
Low 1 ^g	AAUUUGUCAGCCAAA	−0.6	—	—	—
Low 2	AAACCGUAAUUUCAA	−1.8	—	—	—
Bulge 1 5,6Ψ.B ^h	AAAAGGUAGUAUCAA	−6.7	29.1	32.3	−7.1
Bulge 1 5,6Ψ.C ⁱ	AAAAGGUAGU <u>C</u> UCAA ⁱ	−6.0	—	—	—
Bulge 2 Ψ.B ^j	AAAAGGUGUAUAAAA	−5.3	—	—	—
Bulge 2 Ψ.C	AAAAGGUGU <u>C</u> UAAAA	−5.2	—	—	—
Bulge 1 −1.B ^k	ACAGAGUAAGUCAAA	−7.1	30.5	32.6	−7.0
Bulge 1 −1.C	ACA <u>U</u> AGUAAGUCAAA	−5.1	—	27.0	−6.6
Bulge 1 +2/+3.B	AUCAGGUUAAGAAAA	−5.9	29.4	31.8	−7.0
Bulge 1 +2/+3.C	AUCAGGUUAA <u>U</u> AAAA	−4.8	23.4	25.1	−5.4
Bulge 1 +3.B	AACAGGUCAAGCAAA	−5.6	29.2	31.7	−7.1
Bulge 1 +3.C	AACAGGUCA <u>U</u> GCAAA	−5.2	—	28.7	−6.7
Bulge 1 +3/+4/+5.B	AAAAGGUAAAGUACA	−7.4	32.6	36.9	−8.1
Bulge 1 +3/+4/+5.C	AAAAGGUAAA <u>C</u> UACA	−6.5	—	28.7	−6.6
Bulge 1 +3/+4/+5.2.B	AAAAGGUUUUGUAAA	−6.8	23.9	27.8	−6.1
Bulge 1 +3/+4/+5.2.C	AAAAGGUUUU <u>C</u> UAAA	−3.9	—	—	—
Bulge 1 +3/+4.B	AGCAGGUUUUGUAUU	−9.3	31.1	35.6	−7.6
Bulge 1 +3/+4.C	AGCAGGUUUUG <u>C</u> UAUU	−4.9	—	—	—
Bulge 1 +4/+5.B	AAAAGGUAAUUGUAUG	−8.6	30.2	35.8	−7.8
Bulge 1 +4/+5.C	AAAAGGUAAUU <u>U</u> UAUG	−5.5	—	26.8	−6.2
Bulge 1 +4.B	AAAAGGUACAGUACA	−7.4	32.7	37.1	−8.1
Bulge 1 +4.C	AAAAGGUACA <u>U</u> UACA	−5.4	—	24.1	−5.8
Bulge 1 +5.B	AAAAGGUAAACGUACA	−7.4	30.6	34.8	−7.6
Bulge 1 +5.C	AAAAGGUAAAC <u>C</u> UACA	−6.2	—	—	—
Bulge 2.B ^l	AAAAGGUUUUGAGUUA	−5.7	—	28.1	−6.0
Bulge 2.C	AAAAGGUUUUGA <u>C</u> UUA	−3.9	—	—	—
Bulge 3.B	AAAAGGUUUUGAAGUU	−5.1	23.2	25.7	−5.6
Bulge 3.C	AAAAGGUUUUGAA <u>A</u> UU	−3.9	—	25.3	−5.2
Bulge 4.B	AAAAGGUCCCUGAGUAG	−6.5	26.9	26.1	−5.9
Bulge 4.C	AAAAGGUCCCUGA <u>C</u> UAG	−3.9	—	—	—
Bulge 5.B	AAAGGGUUUUAAAAGUAUC	−6.2	24.9	27.4	−6.1
Bulge 5.C	AAAGGGUUUUAAA <u>A</u> CUAUC	−4.5	—	—	—
Bulge 6.B	AACAGGUCUAACAGAGUGA	−6.6	—	34.1	−7.7
Bulge 6.C	AACAGGUCUAACAGA <u>C</u> UGA	−5.2	—	—	—
Bulge 7.B	AUCAGGUUUUGCCCUAGUGG	−6.0	—	—	—
Bulge 7.C	AUCAGGUUUUGCCCUGA <u>C</u> UGG	−4.8	—	—	—
Bulge 8.B	UUAAGGUUCAUUUUAGAGUG	−4.2	—	—	—
Bulge 8.C	UUAAGGUUCAUUUUAGA <u>C</u> UG	−3.9	—	—	—

^a(Tm1) Tm for the unmodified U1 oligo.^b(Tm2) Tm for the modified U1 oligo.^c(ΔG_{exp}) Experimental free energy for the modified U1 oligo.^dConsensus nucleotides are in bold.^e(mut) Mutated.^f(—) No value, no cooperative transition.^g(Low) Low number of consensus nucleotides.^h(Bulge 1 5,6Ψ.B) Single-nucleotide bulge, the pseudouridine at position 5 or 6 of U1.ⁱ(C) Control oligo, with 1 nt difference (underlined) that disrupts the bulge.^j(Bulge 2 Ψ.B) Bulge of both Ψs.^k(Bulge 1 −1.B) Single-nucleotide bulge at position −1 of the 5'ss.^l(Bulge 2.B) Two-nucleotide bulge.

Table 3. Bulge base-pairing and alternative splicing

	Canonical		Bulge	
	All	$\Delta G_2 < -7$	All	$\Delta \Delta G \leq -1$
Total	133,247	49,762	8084	4703
Constitutive	88.51%	87.35%	85.71%	85.29%
Cassette	6.98%	7.90%	8.15%	8.38%
Alternative 3' ss	2.41%	2.11%	1.97%	2.11%
Alternative 5' ss	2.10%	2.62%	4.17%	4.23%
ΔG_2	-7.88	-5.81	-5.76	-5.64

5.1% of all naturally occurring human 5'ss are recognized by these noncanonical registers. Indeed, 6577 genes have at least one bulge 5'ss, which represents 41% of the total number of genes in our data set. These predictions strongly suggest that bulge base-pairing is much more prevalent than shifted base-pairing (only 59 cases) (Roca and Krainer 2009) and that bulge 5'ss are considerably more abundant than noncanonical or GC 5'ss (0.9%) and minor spliceosome (U12-type) 5'ss (0.36%) (Sheth et al. 2006). Bulge 5'ss/U1 base-pairing appears to occur in a wide range of species, likely including one example in budding yeast (Leu and Roeder 1999), which has a small number of U2-type introns with a very highly conserved 5'ss consensus motif and lacks U12-type introns. Furthermore, we estimate that 2.9% of all human 5'ss use a bulge register that confers a substantially lower free energy than the canonical register ($\Delta \Delta G \leq -1$). On the other hand, a small $\Delta \Delta G$ indicates that the bulge helix is roughly as stable as the nonbulge helix. Thus, many 5'ss can be recognized by either canonical or bulge base-pairing at relative levels that can be estimated by $\Delta \Delta G$ and a partition function (Huang et al. 2009).

The formation of 5'ss/U1 helices with single-nucleotide bulges is strongly supported by experimental and computational methods (Table 1). The only exception is the bulge -1 register (2913 candidates), which did not pass the mutational analysis test in *SMN1/2* (data not shown). Nevertheless, 1744 candidate 5'ss for the bulge -1 register have a $\Delta \Delta G \leq -1$, and the T_m data for bulge -1 was consistent with bulge base-pairing (Table 2). Furthermore, even if the bulge at position -1 does not form in *SMN1/2*, we cannot exclude the possibility that it does so in its natural contexts. For these reasons, we kept this register in our predictions.

For single-nucleotide bulges, all of the bulged positions at the 5'ss (+2 to +5, and perhaps -1) or at the 5' end of U1 (only the Ψ at 5 or 6) are limited to the middle of the helix. This observation reflects an energetic requirement for the bulged nucleotide to be flanked by a sufficient number of base pairs. Furthermore, most of the 5'ss positions that are bulged out in the helix are clustered opposite or close to the two Ψ s of U1. Ψ s in RNA helices establish an additional water bridge with the phosphate backbone and stabilize base-stacking in general (Arnez and Steitz 1994; Davis 1995) and specifically in the context of consensus 5'ss/U1 helices (Hall and McLaughlin 1991). Our T_m data also showed that Ψ s strengthen 5'ss/U1 helices, but we could not detect an additional role in bulge structures (Table 2). The Ψ in the U2

snRNA/branch point sequence helix stabilizes base-stacking around the bulged adenosine in the pre-mRNA in addition to placing the bulge in an extrahelical conformation (Lin and Kielkopf 2008). Thus, it is possible that these modified nucleotides contribute to the stability of the 5'ss/U1 bulge structure.

Our predictions also include registers with longer bulges, ranging from 2 to 8 nt. These registers usually confer a smaller energetic advantage over the canonical register (Supplemental Fig. S6), likely reflecting distortions of the RNA helix. These distortions include a kink, with the bending angle increasing with the length of the bulge (Bhattacharyya et al. 1990; Gohlke et al. 1994), and the loss of base-stacking interactions with bulges longer than 1 nt (Znosko et al. 2002). In any case, a total of 1496 such cases have a $\Delta \Delta G \leq -1$, arguing that bulges longer than 1 nt may occur in certain 5'ss/U1 helices.

We found that disease-causing mutations do not affect the bulge structure, but SNPs usually do, yet the overall 5'ss strength is roughly conserved. This suggests that the disruption of base pairs introduced by the bulge register has a less severe impact on 5'ss strength than the disruption of base pairs in common with the canonical register. Furthermore, the bulge 5'ss are proportionately more often involved in alternative splicing than the canonical 5'ss. The selection of alternative 5'ss is influenced by their relative strengths; other *cis*-elements, such as exonic or intronic splicing enhancers or silencers; and *trans*-acting protein factors (Eperon et al. 2000; Yu et al. 2008). Our data suggest that, in general, the additional base pairs enabled by bulging result in a subtle increase in 5'ss strength, which might be important to fine-tune alternative splicing.

An important implication of bulge base-pairing is that the length of the 5'ss motif increases with the length of the bulge, such that some 5'ss span >11 nt. Most of the current 5'ss scoring methods only consider 9 nt or, in some cases, 11 nt (Senapathy et al. 1990; Yeo and Burge 2004; Sahashi et al. 2007; Hartmann et al. 2008), thereby omitting important information that can contribute to 5'ss strength. Furthermore, the longer the 5'ss motif, the more likely it is that U1 base-pairing can compete with proteins binding at overlapping sites and/or with pre-mRNA secondary structures (Warf and Berglund 2010). Such competition scenarios could limit or regulate the formation of bulge structures in cells. Finally, it is also possible that the 5'ss/U1 bulge structures are positively or negatively influenced by RNA-binding proteins bound at nonoverlapping exonic/intronic splicing enhancers/silencers.

In addition to U1 snRNA, there are proteins that bind the 5'ss and influence splicing, such as U1C (a component of U1 snRNP), which can bind the 5'ss even in the absence of the U1 5' end (Du and Rosbash 2002). We found that bulge 5'ss/U1 base-pairing resulted in productive splicing, suggesting that binding of these proteins—if it is indeed necessary—can tolerate the distortion of the helix induced by the bulge. The crystal structure of the U1 snRNP (Pomeranz Krummel et al. 2009) revealed that U1C contacts the minor groove of the base pairs established by U1 nucleotides C8 and A7, which are maintained

in almost all of the bulge registers we describe (except GC 5'ss). The diversity of footprint patterns for the U1 snRNP onto 5'ss sequences (Yu et al. 2008) is also consistent with a dynamic and flexible interaction that can presumably accommodate small distortions in the RNA double helix, such as a bulge. Furthermore, U6 snRNA could potentially base-pair to our test 5'ss in bulge registers as well. U6 and the human consensus 5'ss can only form five Watson-Crick base pairs, suggesting modest energetic requirements for the 5'ss/U6 helix. Since this helix is at the catalytic core of the spliceosome (Rhode et al. 2006), changes in the U6 base-pairing register could affect the selection of the correct *trans*-esterification site during the first step of splicing. It is also possible that U6 binds to the bulge 5'ss in a canonical register, as suggested for the atypical 5'ss that are recognized by U1 in a shifted register (Roca and Krainer 2009).

The multiple 5'ss/U1 base-pairing registers help explain the efficient recognition of many authentic 5'ss otherwise predicted to be weak. In turn, these additional registers increase the number of pseudo-5'ss in exons and especially introns. Pseudo-5'ss are sequences that conform to the 5'ss motif but are not used for splicing (Sun and Chasin 2000), although at least some of them can have functional roles. First, large introns might be spliced in a stepwise manner using internal splice sites, as shown in flies (Hatton et al. 1998) and, in one case, humans (Parra et al. 2008). Second, U1 binding at intronic sites can repress the inclusion of pseudoxons—intronic fragments resembling exons—in the mRNA (Buratti and Baralle 2010). Third, the recently discovered role of U1 in preventing premature polyadenylation at intronic cryptic poly(A) sites (Kaida et al. 2010; Vorlová et al. 2011) might further increase the need for the recognition of intronic pseudo-5'ss by U1. As a corollary, the diversity of 5'ss/U1 registers presumably increases the importance of *cis*-elements and *trans*-factors other than U1 for proper discrimination between authentic and pseudo-5'ss.

We showed that a substantial fraction of 5'ss are recognized by U1 using noncanonical base-pairing registers involving bulged nucleotides and that the use of these registers increases 5'ss strength. These findings further highlight the flexibility of the interaction between the 5'ss and the 5' end of U1, allowing for many base-pairing registers to support efficient splicing. Importantly, these registers are not considered by any of the current splice site scoring methods. Thus, the characterization of these registers should allow the development of more accurate—albeit more complex—prediction tools, with important implications for the molecular classification of splicing mutations and SNPs and for the study of alternative splicing.

Materials and methods

Minigene cloning

We used the U1 expression plasmids and decoys as well as the *SMN1/2* minigenes in the pCI vector as described (Roca and Krainer 2009). We amplified the *MLH1*, *RPGR*, *F5*, *MDM2*, and

RB1 fragments from human genomic DNA and subcloned them into the pcDNA3.1⁺ vector (Invitrogen). We internally deleted introns 3 and 4 of *RPGR*, intron 11 of *MDM2*, and introns 3 and 4 of *RB1* to leave only 225 nt at each end. Likewise, we deleted introns 9 and 10 of *MLH1* and introns 22 and 23 of *F5* to leave 250 nt and 200 nt at each end, respectively. We incorporated the designed mutations into the minigenes by PCR mutagenesis.

Minigene transfection

We cultured and transfected HeLa cells in six-well plates using FuGENE 6 (Roche Diagnostics) as described (Roca and Krainer 2009). The transfection mixture included 80 ng of EGFP-N1 (Clontech), 80 ng of the splicing minigene, and 840 ng of control (pcDNA3.1⁺ or pUC19) or suppressor U1 plasmid.

RNA extraction, reverse transcription, and PCR

We performed RT-PCR analyses as described (Roca and Krainer 2009). Briefly, 48 h post-transfection, we extracted total RNA from cells using TRIzol (Invitrogen). We treated the RNA with RQ1 DNase (Promega), and reverse-transcribed it using SuperScript II RT (Invitrogen) and oligo-dT. We amplified the resulting cDNAs by PCR using vector-specific primers. We 5'-end-labeled one of the PCR primers using T4 polynucleotide kinase (New England Biolabs) and γ -³²P-ATP. We performed 23 cycles of PCR, ensuring exponential amplification. We separated the PCR products by 6% native PAGE and quantified the band intensity by PhosphorImage analysis. We performed RT-PCRs from three independent transfections to derive the mean percentage of inclusion for each experiment. In all cases, the standard deviations were $\leq 5\%$, allowing comparison of exon inclusion percentage values between experiments. We determined the identity of each PCR product by subcloning agarose gel-purified bands with an Original TA Cloning kit (Invitrogen) followed by sequencing. We also directly sequenced the RT-PCR products in Figure 1D to test for potential splicing at a UU 5'ss dinucleotide, as reported for 5'ss similar to the atypical 5'ss (Hartmann et al. 2010).

T_m analysis

The RNA oligonucleotides were synthesized at IDT (Integrated DNA Technologies), purified by RNase-free HPLC, and confirmed by mass spectrometry. Before use, we checked the oligonucleotides by LC-MS under denaturing conditions (tributylamine in 70% [v/v] acetonitrile). We diluted the oligonucleotides in T_m buffer containing 100 mM NaCl, 10 mM NaPO₄, and 0.1 mM EDTA (at pH 7). We mixed each oligonucleotide (5'ss and U1) at a final concentration of 8 μ M, based on the absorbance at 85°C and extinction coefficients (Puglisi and Tinoco 1989). We heated each oligonucleotide pair for 5 min at 95°C and allowed it to cool for 2 h to room temperature. We measured the absorbance at 260 nm as a function of increasing temperature using a Cary 100 Bio UV-Visible spectrophotometer, heating each duplex at a rate of 0.5°C/min in 1-cm path length cells, followed by cooling to confirm reversibility and lack of evaporation. We obtained the T_m and ΔG_{exp} from the absorbance versus temperature curves using CaryWinUV software (Agilent Technologies).

In silico analyses

We compiled an updated database of naturally occurring, well-annotated 5'ss in the human genome, spanning 30 nt with the exon-intron junction in the middle (nucleotides 15/16). We assembled this 5'ss collection using different databases, including SpliceRack (Sheth et al. 2006), dbCASE (Zhang et al. 2007),

and RefSeq (Pruitt et al. 2009). We removed redundant sequences, noncanonical 5'ss, and U12-type 5'ss to obtain all U2-type GT-AG and GC-AG 5'ss.

We used these 5'ss sequences and the 5' end of U1 as inputs for the UNAFold hybrid tool (Markham and Zuker 2008), which calculated the most stable intermolecular base-pairing structure and minimum free energy for each 5'ss. UNAFold predictions are based on empirical thermodynamic parameters, known as nearest-neighbor energy rules (Mathews et al. 1999). These rules include energetic parameters for bulged nucleotides but not for Ψs. The first UNAFold run—without restrictions—returned the most stable structure and ΔG_1 (in kilocalories per mole). From these predictions, we only considered those sequences with registers for 5'ss +1G (G16 on the 30-nt sequence) base-paired to C8 of U1 snRNA. We obtained the ΔG_2 values for canonical base-pairing by forcing +1G of the 5'ss to base-pair to C8 of U1, and not allowing bulges, using the maximum asymmetry option (maxas = 0). The energetic advantage of the bulge over the canonical register was estimated as $\Delta\Delta G = \Delta G_1 - \Delta G_2$. We selected a $\Delta\Delta G$ cutoff of ≤ -1 to identify the bulge 5'ss in which the bulge structure confers a substantial advantage. This value was derived from the mean absolute difference between bulge 5'ss variants carrying SNPs, as in general, both 5'ss variants should be equally functional. Finally, we classified the various bulge 5'ss into different categories, depending on the length of the bulge and the position of the bulged nucleotides.

We cross-analyzed the bulge 5'ss with updated data sets of disease-causing mutations and SNPs at human 5'ss that fall outside of the invariant dinucleotide at positions +1/+2 (Roca et al. 2008). We assessed the conservation of the bulge structure by running UNAFold for the 5'ss with either the mutated nucleotide or the SNP allele that is not in the reference genome sequence. For each disease-causing mutation, we derived the ΔG_1 difference as $\Delta G_1(\text{wild-type}) - \Delta G_1(\text{mutant})$. For SNPs, because both alleles are presumably functional, we calculated the MAD (in negative numbers) as $-(\Delta G_1(\text{allele 1}) - \Delta G_1(\text{allele 2}))$. We then calculated means and standard deviations.

Using the annotations in dbCASE (Zhang et al. 2007), we derived the frequency of each alternative splicing event for internal exons in the canonical 5'ss and bulge 5'ss data sets. We defined a “weak canonical 5'ss” data set as the subset of “canonical 5'ss” with $\Delta G \leq -7$ (49,762 5'ss). We obtained P-values using Fisher's exact test or one-sample t-test.

Acknowledgments

We thank Zuo Zhang, Fred Allain, Pierre Barraud, David Horowitz, and Fedor Karginov for advice; Chaolin Zhang for providing the 5'ss data in dbCASE; and Michael Zuker for implementing the maximum asymmetry option in UNAFold. X.R., M.A., and A.R.K. acknowledge support from NIH grant GM42699.

References

- Arnez JG, Steitz TA. 1994. Crystal structure of unmodified tRNA(Gln) complexed with glutamyl-tRNA synthetase and ATP suggests a possible role for pseudo-uridines in stabilization of RNA structure. *Biochemistry* **33**: 7560–7567.
- Bhattacharyya A, Murchie AI, Lilley DM. 1990. RNA bulges and the helical periodicity of double-stranded RNA. *Nature* **343**: 484–487.
- Brackenridge S, Wilkie AO, Screaton GR. 2003. Efficient use of a ‘dead-end’ GA 5' splice site in the human fibroblast growth factor receptor genes. *EMBO J* **22**: 1620–1631.
- Brunak S, Engelbrecht J, Knudsen S. 1991. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J Mol Biol* **220**: 49–65.
- Buratti E, Baralle D. 2010. Novel roles of U1 snRNP in alternative splicing regulation. *RNA Biol* **7**: 412–419.
- Buratti E, Chivers M, Královicová J, Romano M, Baralle M, Krainer AR, Vorechovsky I. 2007. Aberrant 5' splice sites in human disease genes: Mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. *Nucleic Acids Res* **35**: 4250–4263.
- Cohen JB, Snow JE, Spencer SD, Levinson AD. 1994. Suppression of mammalian 5' splice-site defects by U1 small nuclear RNAs from a distance. *Proc Natl Acad Sci* **91**: 10470–10474.
- Davis DR. 1995. Stabilization of RNA stacking by pseudouridine. *Nucleic Acids Res* **23**: 5020–5026.
- Du H, Rosbash M. 2002. The U1 snRNP protein U1C recognizes the 5' splice site in the absence of base pairing. *Nature* **419**: 86–90.
- Eperon IC, Makarova OV, Mayeda A, Munroe SH, Cáceres JF, Hayward DG, Krainer AR. 2000. Selection of alternative 5' splice sites: Role of U1 snRNP and models for the antagonistic effects of SF2/ASF and hnRNP A1. *Mol Cell Biol* **20**: 8303–8318.
- Freund M, Hicks MJ, Konermann C, Otte M, Hertel KJ, Schaal H. 2005. Extended base pair complementarity between U1 snRNA and the 5' splice site does not inhibit splicing in higher eukaryotes, but rather increases 5' splice site recognition. *Nucleic Acids Res* **33**: 5112–5119.
- Fukumura K, Taniguchi I, Sakamoto H, Ohno M, Inoue K. 2009. U1-independent pre-mRNA splicing contributes to the regulation of alternative splicing. *Nucleic Acids Res* **37**: 1907–1914.
- Gohlke C, Murchie AI, Lilley DM, Clegg RM. 1994. Kinking of DNA and RNA helices by bulged nucleotides observed by fluorescence resonance energy transfer. *Proc Natl Acad Sci* **91**: 11660–11664.
- Hall KB, McLaughlin LW. 1991. Properties of a U1/mRNA 5' splice site duplex containing pseudouridine as measured by thermodynamic and NMR methods. *Biochemistry* **30**: 1795–1801.
- Hartmann L, Theiss S, Niederacher D, Schaal H. 2008. Diagnostics of pathogenic splicing mutations: Does bioinformatics cover all bases? *Front Biosci* **13**: 3252–3272.
- Hartmann L, Neveling K, Borkens S, Schneider H, Freund M, Grassman E, Theiss S, Wawer A, Burdach S, Auerbach AD, et al. 2010. Correct mRNA processing at a mutant TT splice donor in FANCC ameliorates the clinical phenotype in patients and is enhanced by delivery of suppressor U1 snRNAs. *Am J Hum Genet* **87**: 480–493.
- Hatton AR, Subramaniam V, Lopez AJ. 1998. Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon–exon junctions. *Mol Cell* **2**: 787–796.
- Huang FW, Qin J, Reidys CM, Stadler PF. 2009. Partition function and base pairing probabilities for RNA–RNA interaction prediction. *Bioinformatics* **25**: 2646–2654.
- Hwang DY, Cohen JB. 1996. U1 snRNA promotes the selection of nearby 5' splice sites by U6 snRNA in mammalian cells. *Genes Dev* **10**: 338–350.
- Itoh H, Washio T, Tomita M. 2004. Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes. *RNA* **10**: 1005–1018.
- Kaida D, Berg MG, Younis I, Kasim M, Singh LN, Wan L, Dreyfuss G. 2010. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468**: 664–668.
- Kandels-Lewis S, Séraphin B. 1993. Involvement of U6 snRNA in 5' splice site selection. *Science* **262**: 2035–2039.

- Kubota T, Roca X, Kimura T, Kokunai Y, Nishino I, Sakoda S, Krainer AR, Takahashi MP. 2011. A mutation in a rare type of intron in a sodium-channel gene results in aberrant splicing and causes myotonia. *Hum Mutat* **32**: 773–782.
- Kyriakopoulou C, Larsson P, Liu L, Schuster J, Söderbom F, Kirsebom LA, Virtanen A. 2006. U1-like snRNAs lacking complementarity to canonical 5' splice sites. *RNA* **12**: 1603–1611.
- Lerner MR, Boyle JA, Mount SM, Wolin SL, Steitz JA. 1980. Are snRNPs involved in splicing? *Nature* **283**: 220–224.
- Lesser CF, Guthrie C. 1993. Mutations in U6 snRNA that alter splice site specificity: Implications for the active site. *Science* **261**: 1982–1988.
- Leu JY, Roeder GS. 1999. Splicing of the meiosis-specific HOP2 transcript utilizes a unique 5' splice site. *Mol Cell Biol* **19**: 7933–7943.
- Lin Y, Kielkopf CL. 2008. X-ray structures of U2 snRNA-branchpoint duplexes containing conserved pseudouridines. *Biochemistry* **47**: 5503–5514.
- Lorson CL, Hahnen E, Androphy EJ, Wirth B. 1999. A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. *Proc Natl Acad Sci* **96**: 6307–6311.
- Markham NR, Zuker M. 2008. UNAFold: Software for nucleic acid folding and hybridization. *Methods Mol Biol* **453**: 3–31.
- Mathews DH, Sabina J, Zuker M, Turner DH. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288**: 911–940.
- Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**: 457–463.
- Nousbeck J, Burger B, Fuchs-Telem D, Pavlovsky M, Fenig S, Sarig O, Itin P, Sprecher E. 2011. A mutation in a skin-specific isoform of SMARCA1 causes autosomal-dominant adermatoglyphia. *Am J Hum Genet* **89**: 302–307.
- Parra MK, Tan JS, Mohandas N, Conboy JG. 2008. Intrasplicing coordinates alternative first exons with alternative splicing in the protein 4.1R gene. *EMBO J* **27**: 122–131.
- Pomeranz Krummel DA, Oubridge C, Leung AK, Li J, Nagai K. 2009. Crystal structure of human spliceosomal U1 snRNP at 5.5 Å resolution. *Nature* **458**: 475–480.
- Pruitt KD, Tatusova T, Klimke W, Maglott DR. 2009. NCBI reference sequences: Current status, policy and new initiatives. *Nucleic Acids Res* **37**: D32–D36. doi: 10.1093/nar/gkn721.
- Puglisi JD, Tinoco I Jr. 1989. Absorbance melting curves of RNA. *Methods Enzymol* **180**: 304–325.
- Rhode BM, Harmuth K, Westhof E, Lührmann R. 2006. Proximity of conserved U6 and U2 snRNA elements to the 5' splice site region in activated spliceosomes. *EMBO J* **25**: 2475–2486.
- Roca X, Krainer AR. 2009. Recognition of atypical 5' splice sites by shifted base-pairing to U1 snRNA. *Nat Struct Mol Biol* **16**: 176–182.
- Roca X, Olson AJ, Rao AR, Enerly E, Kristensen VN, Børresen-Dale AL, Andresen BS, Krainer AR, Sachidanandam R. 2008. Features of 5'-splice-site efficiency derived from disease-causing mutations and comparative genomics. *Genome Res* **18**: 77–87.
- Rogers J, Wall R. 1980. A mechanism for RNA splicing. *Proc Natl Acad Sci* **77**: 1877–1879.
- Sahashi K, Masuda A, Matsuura T, Shinmi J, Zhang Z, Takeshima Y, Matsuo M, Sobue G, Ohno K. 2007. In vitro and in silico analysis reveals an efficient algorithm to predict the splicing consequences of mutations at the 5' splice sites. *Nucleic Acids Res* **35**: 5995–6003.
- Senapathy P, Shapiro MB, Harris NL. 1990. Splice junctions, branch point sites, and exons: Sequence statistics, identification, and applications to genome project. *Methods Enzymol* **183**: 252–278.
- Séraphin B, Kretzner L, Rosbash MH. 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.
- Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R. 2006. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res* **34**: 3955–3967.
- Siliciano PG, Guthrie C. 1988. 5' splice site selection in yeast: Genetic alterations in base-pairing with U1 reveal additional requirements. *Genes Dev* **2**: 1258–1267.
- Staley JP, Guthrie C. 1999. An RNA switch at the 5' splice site requires ATP and the DEAD box protein Prp28p. *Mol Cell* **3**: 55–64.
- Sun H, Chasin LA. 2000. Multiple splicing defects in an intronic false exon. *Mol Cell Biol* **20**: 6414–6425.
- Vorlová S, Rocco G, Lefave CV, Jodelka FM, Hess K, Hastings ML, Henke E, Cartegni L. 2011. Induction of antagonistic soluble decoy receptor tyrosine kinases by intronic polyA activation. *Mol Cell* **43**: 927–939.
- Wahl MC, Will CL, Lührmann R. 2009. The spliceosome: Design principles of a dynamic RNP machine. *Cell* **136**: 701–718.
- Wang J, Smith PJ, Krainer AR, Zhang MQ. 2005. Distribution of SR protein exonic splicing enhancer proteins in human protein-coding genes. *Nucleic Acids Res* **33**: 5053–5062.
- Warf MB, Berglund JA. 2010. Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem Sci* **35**: 169–178.
- Wassarman DA, Steitz JA. 1992. Interactions of small nuclear RNAs with precursor messenger RNA during in vitro splicing. *Science* **257**: 1918–1925.
- Will CL, Lührmann R. 2005. Splicing of a rare class of introns by the U12-dependent spliceosome. *Biol Chem* **386**: 713–724.
- Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**: 377–394.
- Yu Y, Maroney PA, Denker JA, Zhang XH, Dybkov O, Lührmann R, Jankowsky E, Chasin LA, Nilsen TW. 2008. Dynamic regulation of alternative splicing by silencers that modulate 5' splice site competition. *Cell* **135**: 1224–1236.
- Zhang C, Hastings ML, Krainer AR, Zhang MQ. 2007. Dual-specificity splice sites function alternatively as 5' and 3' splice sites. *Proc Natl Acad Sci* **104**: 15028–15033.
- Zhuang Y, Weiner AM. 1986. A compensatory base change in U1 snRNA suppresses a 5' splice site mutation. *Cell* **46**: 827–835.
- Znosko BM, Silvestri SB, Volkman H, Boswell B, Serra MJ. 2002. Thermodynamic parameters for an expanded nearest-neighbor model for the formation of RNA duplexes with single nucleotide bulges. *Biochemistry* **41**: 10406–10417.

Widespread recognition of 5' splice sites by noncanonical base-pairing to U1 snRNA involving bulged nucleotides

Xavier Roca, Martin Akerman, Hans Gaus, et al.

Genes Dev. 2012, **26**:

Access the most recent version at doi:[10.1101/gad.190173.112](https://doi.org/10.1101/gad.190173.112)

Supplemental Material

<http://genesdev.cshlp.org/content/suppl/2012/05/14/26.10.1098.DC1>

References

This article cites 59 articles, 15 of which can be accessed free at:
<http://genesdev.cshlp.org/content/26/10/1098.full.html#ref-list-1>

License

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).