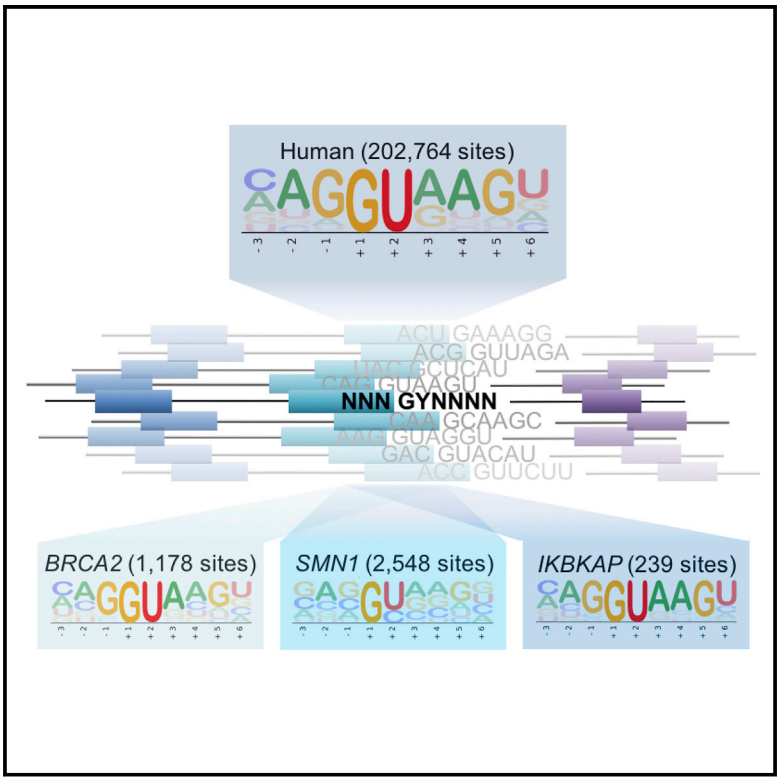
Article

Molecular Cell

Quantitative Activity Profile and Context Dependence of All Human 5′ Splice Sites

Graphical Abstract



Highlights

- Comprehensive measurement of 5′ss activity in three gene contexts
- A major determinant of 5′ss recognition stems from the nucleotide sequence
- Context can have a considerable influence on 5′ss usage
- Compiled 5′ss measurements help distinguish pathogenic from benign 5′ss mutations

Authors

Mandy S. Wong, Justin B. Kinney, Adrian R. Krainer

Correspondence

jkinney@cshl.edu (J.B.K.), krainer@cshl.edu (A.R.K.)

In Brief

To examine the complexity of 5′ splice site selection, Wong et al. established a method to comprehensively measure 5′ss activity in three gene contexts. Although context dependence was observed, the major determinant of 5′ss selection was found to be intrinsic to the 5′ss nucleotide sequence.

Wong et al., 2018, Molecular Cell *71*, 1–15
September 20, 2018
https://doi.org/10.1016/j.molcel.2018.07.033

CellPress

Please cite this article in press as: Wong et al., Quantitative Activity Profile and Context Dependence of All Human 5′ Splice Sites, Molecular Cell (2018), https://doi.org/10.1016/j.molcel.2018.07.033

Molecular Cell
Article

Quantitative Activity Profile and Context Dependence of All Human 5′ Splice Sites

Mandy S. Wong,[1] Justin B. Kinney,[1,2,*] and Adrian R. Krainer[1,2,3,*]
[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA
[2]These authors contributed equally
[3]Lead Contact
*Correspondence: jkinney@cshl.edu (J.B.K.), krainer@cshl.edu (A.R.K.)
https://doi.org/10.1016/j.molcel.2018.07.033

SUMMARY

Pre-mRNA splicing is an essential step in the expression of most human genes. Mutations at the 5′ splice site (5′ss) frequently cause defective splicing and disease due to interference with the initial recognition of the exon-intron boundary by U1 small nuclear ribonucleoprotein (snRNP), a component of the spliceosome. Here, we use a massively parallel splicing assay (MPSA) in human cells to quantify the activity of all 32,768 unique 5′ss sequences (NNN/GYNNNN) in three different gene contexts. Our results reveal that although splicing efficiency is mostly governed by the 5′ss sequence, there are substantial differences in this efficiency across gene contexts. Among other uses, these MPSA measurements facilitate the prediction of 5′ss sequence variants that are likely to cause aberrant splicing. This approach provides a framework to assess potential pathogenic variants in the human genome and streamline the development of splicing-corrective therapies.

INTRODUCTION

Pre-mRNA splicing is the process of joining exon sequences, with the concomitant removal of noncoding intron sequences, to generate mature mRNA in the nucleus. Alternative splicing affects ~95% of the genes in multicellular eukaryotes, allowing for the generation of over 100,000 proteins from ~23,000 protein-coding sequences, thus greatly expanding the coding capacity of eukaryotic genomes (Nilsen and Graveley, 2010). Splicing is a largely co-transcriptional and highly regulated process that involves the dynamic recruitment and assembly of many components, including the core spliceosome that comprises ~200 proteins and five small nuclear RNAs (snRNAs), which work in concert with remarkable precision (Will and Lührmann, 2011). Mutations in *cis*-elements essential for splicing or its regulation (i.e., the 5′ splice site, 3′ splice site, branchpoint sequence, and intronic or exonic enhancer and silencer elements) and deregulation of splicing-factor expression cause or contribute to the development of many human diseases. It is estimated that 14% of all disease-associated point mutations affect splice sites (Krawczak et al., 2000; Soemedi et al., 2017) and that as many as 50% of all mutations alter splicing, when accounting for mutations that affect enhancer and silencer elements (Cartegni et al., 2002). Such mutations cause aberrant splicing of relevant genes in cancer and in neuromuscular and other diseases (Krawczak et al., 1992; Srebrow and Kornblihtt, 2006).

The 5′ splice site (5′ss) is a 9-nt motif that demarcates the boundary between an exon and the intron that follows it. It comprises 3 nt at the end of the upstream exon (−3 to −1) and 6 nt at the beginning of the intron (+1 to +6). This 9-nt motif has a consensus sequence of CAG/GUAAGU, which is precisely complementary to a sequence at the 5′ end of U1 snRNA (Lerner et al., 1980; Rogers and Wall, 1980; Zhuang and Weiner, 1986). The great majority of introns are of the U2 type and are spliced by the major spliceosome (with U1, U2, U4, U5, and U6 snRNAs/small nuclear ribonucleoproteins [snRNPs]). Only a small subset of introns are of the U12 type and are spliced by the minor spliceosome (with U11, U12, U4atac, U6atac, and U5 snRNAs and snRNPs) (Tarn et al., 1995). Among U2-type introns, 98.8% have GU at the +1 and +2 positions of the 5′ss, whereas only 0.87% have GC (Sheth et al., 2006). The remaining 0.33% of U2-type introns have non-consensus sites, such as introns with 5′ AT and 3′ AC ends (Kubota et al., 2011).

Recent high-resolution crystal and cryo-electron microscopy (cryo-EM) structures of yeast and human spliceosomes and their components are improving our understanding of the mechanism of 5′ss recognition (Kondo et al., 2015; Bao et al., 2017; Bertram et al., 2017; Wan et al., 2017). It was definitively shown that the recognition of an RNA oligonucleotide with an AAG/GUAAGU 5′ss sequence by human U1 snRNP involves direct base-pairing with U1 snRNA and this interaction is stabilized by hydrogen bonds formed between U1-C polypeptide and the sugar-phosphate backbone of the pre-mRNA (i.e., without base-specific contacts) (Kondo et al., 2015). This additional stabilization by U1-C may support non-canonical base-pairing interactions (such as U-Ψ) (Tan et al., 2016), which may be especially important in higher eukaryotes, due to the more degenerate nature of their 5′ss motif, compared to that of budding yeast.

Many of the point mutations that affect splicing disrupt 5′ss recognition by U1 snRNP, the first step in spliceosome assembly. The 5′ss sequence is highly degenerate, but at least 6 Watson-Crick or wobble base pairs with U1 snRNA are thought to be necessary for splicing to occur (Zhuang and Weiner, 1986; Ketterling et al., 1999). This vague definition of what constitutes a

functional 5′ss sequence reflects current knowledge of the 5′ss element; this knowledge derives on one hand from extensive functional and structural studies involving the consensus sequence or a small number of model substrates and on the other hand from the alignment of 5′ss sequences from many natural introns. Aligning all the sequences in this manner, however, overlooks contextual influences and rare 5′ss sequences that may be biologically interesting.

In spite of the multiple base pairs formed with U1 snRNA, a single point mutation within the 5′ss can be sufficient to abolish the recognition of a 5′ss. Whereas mutations at the invariant +1G always cause aberrant splicing (due to its role in transesterification chemistry), the consequence of mutations at the other positions of the 5′ss are less predictable. Some mutations are causal for various human diseases, whereas others are neutral. Why certain 5′ss are more sensitive to mutations is mechanistically intriguing and has broad implications, yet there are insufficient data available for most disease-related genes to establish rules to accurately predict which mutations are pathogenic. Of 15,786 single-nucleotide polymorphisms (SNPs) identified at human 5′ss, only 3.3% are classified as pathogenic or likely pathogenic, whereas the remaining 96.7% have unknown functional consequences (Landrum et al., 2016). Without laborious experimental examination of the effects of individual point mutations, it has not been possible to reliably predict their effects on splicing. This gap in knowledge, in turn, hinders the development of splicing-corrective therapies.

Recent advances in massively parallel assays allow simultaneous surveying of the influence of many different mutations on splicing; this approach has been applied to examine, for example, the influence of hexamer sequence motifs on nearby alternative 5′ss sequences (Rosenberg et al., 2015) and the effects of systematic mutations within an exon (Singh et al., 2004; Julien et al., 2016; Ke et al., 2018). So far, such studies have focused on regulatory elements spread throughout the sequence, and revealed the high content of *cis*-regulatory elements within exons and introns. Here, we pursued a complementary approach, using a focused massively parallel splicing assay (MPSA) to empirically examine the effects of all possible variants of a single discrete element, the 9-nt 5′ss. We performed this assay in three different gene contexts: *BRCA2* intron 17, *SMN1* intron 7, and *IKBKAP* intron 20. We found that the 5′ss sequence alone is a major determinant of 5′ss recognition. In the *BRCA2* and *SMN1* contexts, the 5′ss sequence accounts for 68%–72% of the variation in 5′ss usage. Surprisingly, 5′ss selection was unusually stringent in the *IKBKAP* context, which we demonstrate is attributable to its weak upstream 3′ss; strengthening this 3′ss was sufficient to improve the recognition of selected 5′ss sequences to similar levels as in the *BRCA2* and *SMN1* contexts. Based on our results, we can predict the negative impact on splicing of ~90% of disease-associated 5′ss mutations. Therefore, this study establishes a quantitative assessment of the usage of all possible 5′ss and provides insights into how 5′ss mutations may alter splicing efficiency and cause disease. Knowing which mutations are pathogenic can help identify individuals at risk for various genetic diseases and should facilitate early detection and intervention.

RESULTS

5′ss Usage and the Effects of Mutations Are Recapitulated in Minigenes

To begin dissecting how a particular 5′ss sequence is recognized by U1 snRNP, we constructed a *BRCA2* (breast cancer 2, a tumor suppressor gene) minigene spanning exons 16–18. We introduced several single point mutations within the 5′ss of exon 17 to determine if its recognition is affected. We inserted the minigenes into the pcDNA5/Flp recombination target (FRT) vector and stably transfected them into a HeLa cell clone, which was selected for a single FRT site (introduced by the Flp-In system) to eliminate variability arising from random genomic integration.

The 5′ss of *BRCA2* intron 17 is a non-canonical GC 5′ss with the sequence CAG/GCAAGTTT, which adheres closely to the human consensus 5′ss generated from 202,764 natural 5′ss (Figure 1A) (Sheth et al., 2006), despite having a cytosine at the +2 position and a thymidine at the +7 position. Intervening sequence (IVS) mutations 17-1G > C, IVS17-1G > A, and IVS17+5G > A were previously observed in breast cancer samples (Hofmann et al., 2003; Landrum et al., 2016; Teng et al., 1996). In our minigene experiments, the G > A point mutation at the +5 position induced complete exon 17 skipping (Figure 1C, lane 3), suggesting that patients with these point mutations likely have splicing defects in *BRCA2*, and that the effects of such mutations can be accurately reproduced in our minigene assay. Strengthening the 5′ss by altering the GC to GT was sufficient to suppress the effect of the +5 mutation (lane 4).

5′ss recognition can sometimes be influenced by the next 2 nt in the intron (+7 and +8), which can extend the complementarity at the 5′ end of the U1 snRNA, even though the consensus motif has no nucleotide preference at these positions (Figure 1B) (Freund et al., 2005; Sheth et al., 2006; Hartmann et al., 2008). In this case, however, increasing the complementarity to U1 by changing +7T > A was insufficient to overcome the effect of the +5 mutation (lane 6). Mutations at almost every position (except for +7 and +8) also induced exon skipping, regardless of the actual nucleotide. A few exceptions arise when complementarity is maintained by forming a G:U (−2G:10U) or G:ψ (+3G:6ψ) wobble base pair with U1 snRNA (Figure 1D). Thus, we observed that many distinct single point mutations can strongly affect *BRCA2* exon 17 5′ss recognition.

To further evaluate the role of the nucleotide sequence in 5′ss recognition, we used the same 5′ss sequences to replace the natural 5′ss of intron 6 of another *BRCA2* minigene spanning exons 5 to 7 (Figure 1E). With two exceptions (−2C:10U and +6C:3A), the activities of these 5′ss in both *BRCA2* minigenes closely resembled each other, indicating that a given 5′ss behaves similarly when placed in these two different minigene contexts. This result suggests that although other *cis*-acting elements may influence the recognition of a 5′ss, the sequence of the 5′ss is a major determinant of its usage.

High-Throughput Analysis of the Activity of all 5′ss Sequences

To more thoroughly assess the effect of 5′ss variation, we developed an MPSA to comprehensively assay all 32,768 possible

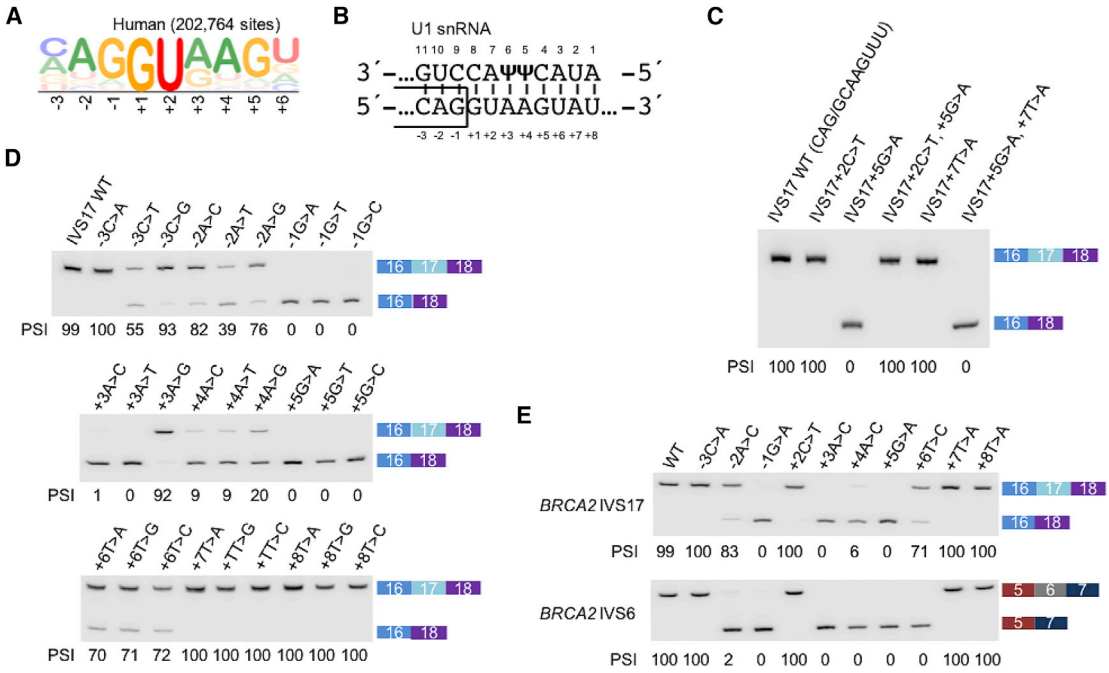


Figure 1. 5′ss Activity and the Effects of Mutations Can Be Recapitulated in Minigenes

(A) Sequence logo generated from 202,764 5′ss sequences in the human transcriptome.

(B) Diagram showing the base-pairing between U1 snRNA and the consensus 5′ss sequence. Ψ represents pseudouridine, an isomer of uridine found at conserved positions in U1 snRNA.

(C) Systematic mutation of the exon 17 5′ss of a *BRCA2* minigene spanning exons 16–18. Gel image is representative of triplicates. Percent spliced in (PSI) is indicated below each lane.

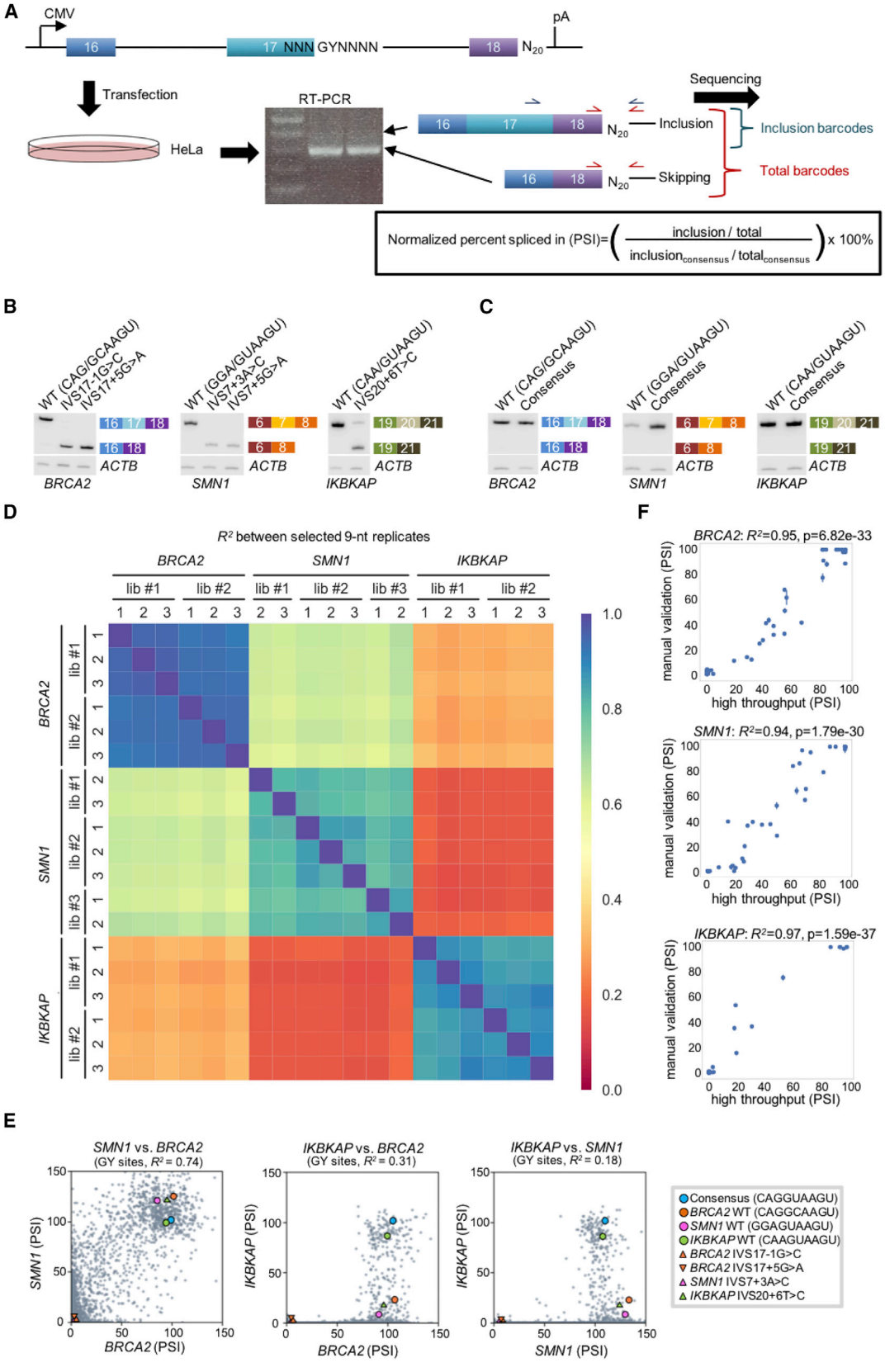
(D) Different nucleotides at the same position induce exon skipping to similar extents, except for a few exceptions when the mutated nucleotide maintains complementarity by forming a G:Ψ wobble base pair with U1 snRNA (e.g., +3A > G). Gel images are representative of triplicates.

(E) *BRCA2* intron 17 5′ss wild-type (WT) and mutant sequences replacing the 5′ss of *BRCA2* intron 6 in a *BRCA2* exons 5–7 minigene (bottom) show similar splicing efficiencies compared to the intron 17 context (top). Gel images are representative of triplicates.

9-nt GU and GC 5′ss sequences (Figures 2A and S1). For *BRCA2*, we first generated a partial minigene comprising exon 16, intron 16, and exon 17 with a randomized 9-nt 5′ss sequence replacing the natural exon 17 5′ss, and a randomized 20-nt barcode sequence at the 3′ end. We subjected this DNA library to next-generation sequencing, thus producing a "key" that associates each 20-nt barcode with a corresponding 5′ss. We then inserted intron 17 and exon 18 into restriction sites located between the randomized 5′ss and the 20-nt barcode to complete the minigene library. We transiently transfected the resulting minigene library into HeLa cells to assess splicing by RT-PCR. We separately amplified and deep sequenced the 20-nt barcodes of two samples: (1) the gel-purified exon-inclusion product and (2) the total transfected library. With the previously sequenced key, we used the barcodes to identify the extent to which individual 5′ss sequences resulted in exon inclusion versus skipping. Using this strategy, we determined and calculated the relative usage of each 5′ss sequence by quantifying the ratio of the number of exon-inclusion barcode reads to the number of barcode reads from total RNA.

We generated 5′ss libraries for three different disease-relevant contexts, namely the middle exon of minigenes *BRCA2* exons 16–18, *SMN1* exons 6–8, and *IKBKAP* exons 19–21, using the same strategy described above to determine the extent of context-specific effects. As mentioned above, the *BRCA2* IVS17+5G > A pathogenic mutation resulted in defective splicing in our minigene (Figure 2B). *SMN1* (survival of motor neuron 1) deletions or point mutations cause spinal muscular atrophy (Lefebvre et al., 1995). Besides a G > C mutation at the invariant +1 position, no other natural 5′ss mutations have been reported in *SMN1* intron 7 (Singh et al., 2017). We introduced an additional mutation (IVS7+24G > C) in the *SMN1* minigene to prevent the selection of a cryptic 5′ss (data not shown). Among several point mutations we introduced at this 5′ss of *SMN1* exon 7, mutations at the +3 and +5 positions induced exon 7 skipping (Figure 2B). Our choice of *IKBKAP* (inhibitor of kappa light polypeptide enhancer in B cells, kinase associated-protein) was motivated by the fact that a homozygous 5′ss mutation, IVS20+6T > C, is responsible for familial dysautonomia or Riley-Day syndrome (Anderson et al., 2001; Slaugenhaupt et al., 2001). The effect of this mutation was also recapitulated in our minigene assay (Figure 2B).

To ensure the reproducibility of our results, we generated two or three independently derived libraries, each having different barcode to 5′ss associations. We then assayed each library in three separate replicate experiments. Deep-sequencing analysis showed that each library covered at least 90% of the 32,768 possible 5′ss sequences (Table 1; Figures S2A and S2B). Only



(legend on next page)

Table 1. Summary of Sequencing Data

	Library	Number of Reads	Percentage of 5′ss Barcoded	Barcodes per 5′ss	Reads per BC-ss Association
BRCA2	1	1.89×10^7	100.0	28.84 ± 7.95	19.84 ± 16.96
	2	1.82×10^7	100.0	68.27 ± 14.99	8.10 ± 6.01
SMN1	1	5.70×10^6	91.4	10.83 ± 8.46	17.25 ± 21.01
	2	1.12×10^7	94.5	18.52 ± 14.59	19.37 ± 17.68
	3	1.13×10^7	92.9	12.47 ± 9.62	29.64 ± 31.50
IKBKAP	1	1.04×10^7	99.8	34.51 ± 21.86	9.02 ± 11.74
	2	8.5×10^6	99.7	26.44 ± 16.97	9.70 ± 12.36

Summary statistics for each independently derived minigene library for the three gene contexts is shown with SD when appropriate. Further information is provided in Figure S2. BC, barcode; ss, splice site.

8 of the 32,768 sequences (all eight beginning with GCGG) were not represented in any of the libraries. We normalized the inclusion ratio of each 5′ss sequence to that measured for the consensus 5′ss sequence (CAG/GUAAGU), since the consensus 5′ss resulted in complete exon inclusion in all three contexts (Figures 2C and S2C). We equated this normalized ratio to the percent spliced in (PSI) value measured for each 5′ss.

Within each minigene context, we observed good consistency in 5′ss usage among the independently derived libraries and replicates of each minigene, as indicated by the high coefficient of determination (R^2) between measured PSI values (Figures 2D and S2D). In addition to validating that each minigene library yielded consistent and reproducible 5′ss usage quantitation, this analysis demonstrated that the 20-nt randomized barcode sequence added to the 3′ end of each minigene generally did not detectably affect 5′ss selection (Figure S2E). The effects of the few outliers are negligible when averaging across the numerous barcodes associated with each 5′ss. By sequencing the exon-exon junction, we observed that GC 5′ss sequences with a GU dinucleotide at the −2 and −1 positions (NGU/GCNNNN) preferentially used the GU instead of the GC (Figure S3). To prevent our results from being skewed by these shifted junctions, we excluded these 1,024 5′ss sequences from this and subsequent analyses.

Measurements within each minigene context correlated substantially better with each other than between minigene contexts (Figure 2D). Averaging across replicates (Figure S2D), we find that PSI measurements for the two *BRCA2* libraries correlate at R^2 = 96%, measurements for the 3 *SMN1* libraries exhibit R^2 = 82%–87%, and measurements for the two *IKBKAP* libraries exhibit R^2 = 92%. There was a reduced correlation (R^2 = 68%–72%) between *BRCA2* and *SMN1* libraries. Finally, these *BRCA2* and *SMN1* measurements were far less well correlated with the *IKBKAP* measurements (R^2 = 30%–32% for *BRCA2* versus *IKBKAP* and R^2 = 18%–19% for *SMN1* versus *IKBKAP*). This result suggests a substantial difference in how 5′ss sequences are recognized by the spliceosome in the *IKBKAP* context relative to the *BRCA2* and *SMN1* contexts.

This context dependence is further illustrated in Figure 2E, where 5′ss selection in *IKBKAP* is seen to be more stringent than in *BRCA2* and *SMN1*. A large population of 5′ss that were used efficiently for splicing in both *BRCA2* and *SMN1* varied greatly in splicing activity in *IKBKAP*. All wild-type 5′ss, along with the splicing-deficient mutant 5′ss sequences of each minigene, behaved as expected in their respective contexts. Whereas all three wild-type sequences were efficiently used in both *BRCA2* and *SMN1*, *IKBKAP* could only tolerate its own natural 5′ss sequence. We manually validated 53 representative sequences in each context, to confirm the reliability of the next-generation sequencing results (Figures 2F and S4). Altogether, we empirically determined the usage of virtually all possible 9-nt 5′ss sequences in three different gene contexts.

Figure 2. MPSA Measurements for 5′ss Sequences

(A) Schematic of the MPSA used to assess all 5′ss sequences. Minigenes were inserted into the pcDNA5 expression vector, which has a cytomegalovirus (CMV) promoter and a bGH polyadenylation site (pA). See also Figure S1.

(B) Splicing of *BRCA2*, *SMN1*, and *IKBKAP* minigenes with wild-type and mutant 5′ss sequences. These measurements confirm that our minigene constructs can recapitulate the effects of known disease-associated mutations. *ACTB* was amplified in the same PCR reaction as a loading control. The gel image was divided for easier visualization. Gel images are representative of triplicates.

(C) Splicing of *BRCA2*, *SMN1*, and *IKBKAP* minigenes with wild-type and consensus (CAG/GUAAGU) 5′ss sequences. The consensus sequence gives 100 PSI in all three contexts, substantiating its use in normalizing PSI measurements. *ACTB* was amplified in the same PCR reaction as a loading control. The gel image was divided for easier visualization. Gel images are representative of triplicates.

(D) Heatmap reporting the squared Pearson correlation (R^2) of PSI values measured in 19 independent experiments. These correlations show that the replicate libraries within each context are more consistent with each other than with measurements made in heterologous contexts. Two low-quality datasets (*SMN1* library 1, replicate 1 and *SMN1* library 3, replicate 3) were not included in this and subsequent analyses (see Figures S2C and S2D).

(E) Scatterplots comparing PSI values for each pair of minigene contexts. The consensus and wild-type 5′ss sequences are marked by circles with the indicated colors, and the mutant sequences are marked by triangles.

(F) Comparison of high-throughput PSI measurements to manual measurements made in each context for the same 53 randomly selected 5′ss sequences for each context. Error bars indicate SD across triplicate transfections. Note that the high-throughput PSI measurements shown here are capped at 100. Figure S4 illustrates these measurements for each individual 5′ss assayed.

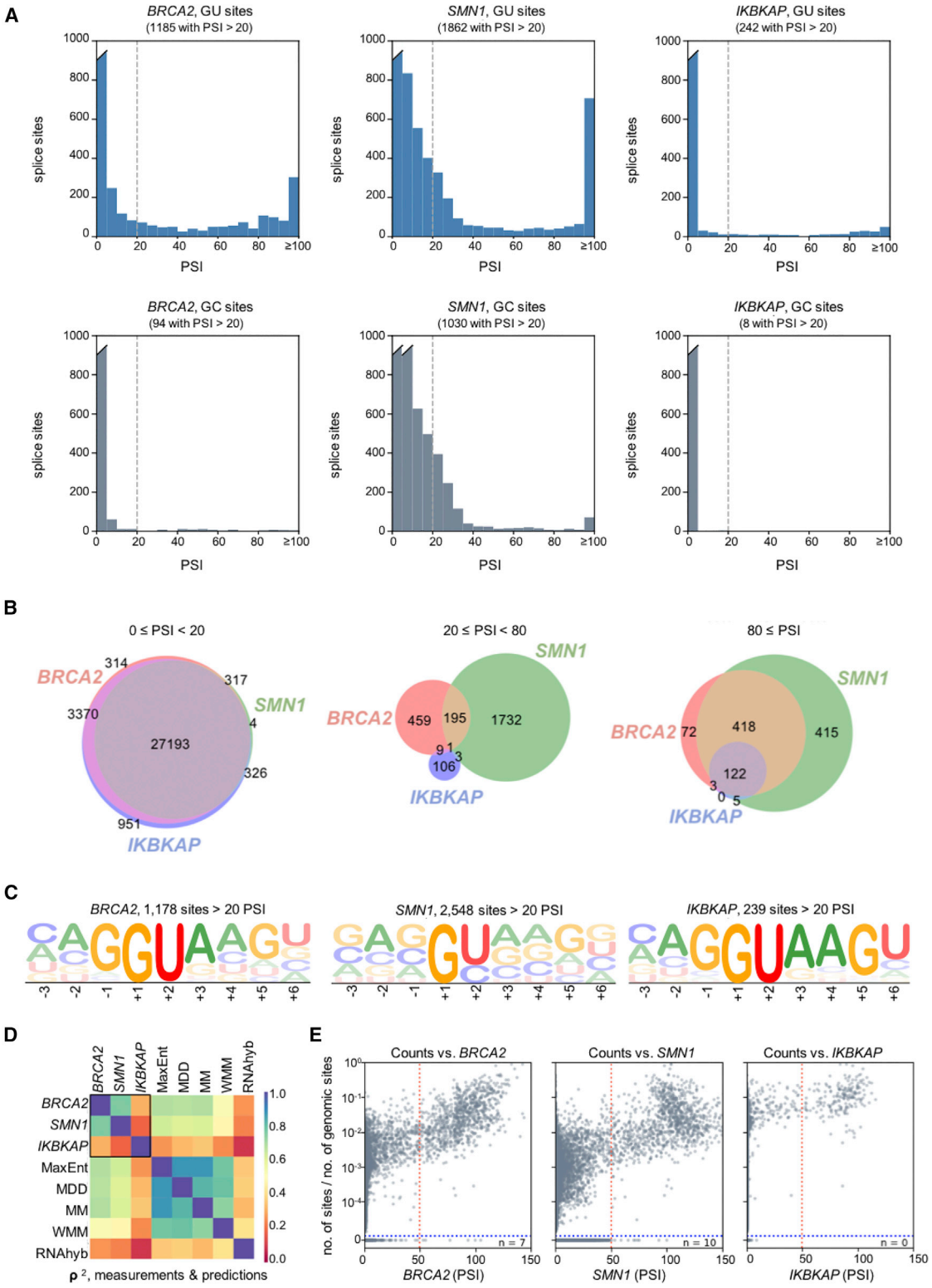


Figure 3. Further Comparisons of MPSA Measurements across Three Different Contexts

(A) Histograms showing the distribution of PSI measurements for all GU (top) and GC (bottom) 5′ss sequences in each of the three minigene contexts. The dashed line marks the 20% cutoff used to designate a 5′ss as active. The breaks in the leftmost bars (indicated by a slant mark) indicate values exceeding the upper limit on the y axis. PSI measurements above 100 were included in the rightmost bar in each plot.

(legend continued on next page)

5′ss Usage Follows General Trends Modulated by Context-Dependent Effects

The 5′ss sequences tended to be used in a bimodal manner, such that the majority were either recognized for splicing with high efficiency or not recognized at all (Figure 3A). These bimodal distributions, however, varied substantially between gene contexts. The *SMN1* context had a large population of 5′ss sequences comprising both GU and GC that yielded moderate to low activity, but this was not the case in either the *BRCA2* or *IKBKAP* contexts. Even though our libraries covered at least 90% of the 32,768 possible 5′ss sequences, including both GU and GC, the latter were considerably weaker than the former in all three contexts. Paradoxically, though the natural *BRCA2* intron 17 5′ss is a GC 5′ss that is efficiently used for splicing, other GC 5′ss sequences were not tolerated for splicing in this context. Whereas 1279 5′ss sequences had PSI $\geq$20 in *BRCA2*, 2892 5′ss had PSI $\geq$20 in *SMN1*. Although the overlap in the list of 5′ss with activity in both contexts is high (1117 5′ss with PSI $\geq$20 in both *BRCA2* and *SMN1*), there is a subset of 5′ss with high activity in one context but little to no activity in the other (6 5′ss with PSI $\geq$80 in *BRCA2* and PSI <20 in *SMN1*; 70 5′ss with PSI <20 in *BRCA2* and PSI $\geq$80 in *SMN1*). In agreement with the stringent selection of 5′ss sequences in *IKBKAP* noted above, only 250 5′ss sequences had PSI $\geq$20 in *IKBKAP*. Collectively, our results show that context can have a considerable influence on 5′ss activity.

When the 5′ss in the bimodal distribution were separated by activity level, only 122 (0.4%) had high inclusion ratios (PSI $\geq$80) in all three contexts (Figures 3B and S5A). The great majority of 5′ss sequences (27,193, or 83%) had PSI <20 in all three contexts. This large proportion of seemingly non-functional sequences is consistent with the fact that only 9,574 (58.4%) of the 16,384 possible permutations of GU 5′ss, and 92 (0.56%) out of 16,384 possible GC 5′ss have been annotated as bona fide 5′ss sequences that occur at least once in the human transcriptome (Sheth et al., 2006). Interestingly, the sequences in the moderate-efficiency population with 20–80 PSI hardly overlapped among the three minigene contexts.

To better understand context-dependent sequence requirements, we generated sequence logos specific to each context (Figure 3C). Whereas the logo for *BRCA2* is very similar to the logo for all human 5′ss (Figure 1A), *SMN1* has much greater flexibility, showing minimal preference at any of the variable positions of the 5′ss (Figures 3C and S5). On the other hand, *IKBKAP* has a strong preference for the consensus sequence.

MPSA Measurements Predict 5′ss Activity Better than Computational Algorithms

Comparison between our empirically derived 5′ss data and several computational scoring matrices that predict 5′ss strength shows modest improvement by our method in predicting 5′ss usage (Figures 3D and S6). Using the measurements from the *BRCA2* libraries, we can predict 5′ss usage in the *SMN1* context with ~70% accuracy (as quantified by the squared Spearman correlation, ρ^2), and vice versa. Whereas some popular prediction algorithms, such as the maximum entropy model (MaxENT; Yeo and Burge, 2004), the maximum dependence decomposition model (MDD; Burge, 1998), and the first-order Markov model (MM; Krogh et al., 1994), can also predict 5′ss usage in our libraries with slightly lower accuracy, the weight matrix model (WMM) and the free energy of 5′ss/U1 base-pairing, as calculated by RNAhybrid (Krüger and Rehmsmeier, 2006), were the least predictive of the models we examined ($\rho^2 < 0.50$). However, none of the models could accurately predict 5′ss behavior in the *IKBKAP* context, nor could our *BRCA2* and *SMN1* data.

Natural Selection for Functional 5′ss Sequences

Next, we asked whether there are functional 5′ss sequences that exhibit high PSI but do not occur naturally in the human transcriptome. By using a PSI $\geq$50 cutoff, we disregarded the population of sequences that yielded low splicing activity in the *SMN1* context only (Figure 3E). We found only 7 sequences in the *BRCA2* context and 10 sequences in the *SMN1* context that had $\geq$50% activity and do not occur naturally (Table S1). However, many of these 5′ss have a secondary GU or GC embedded within the sequence (underlined in Table S1), possibly allowing for the use of an alternative 5′ss. Removal of those sequences left only 2 sequences in the *BRCA2* context and 4 sequences in the *SMN1* context. Remarkably, one of these sequences, ACG/GUAUCG, is shared between the two contexts, which we confirmed by RT-PCR to have high activity in both *BRCA2* and *SMN1* contexts (Figure S6C). This 5′ss can only base pair at five positions with the U1 snRNA. Considering the nucleotides flanking the 5′ss, it is unlikely that a shifted or bulged register to the U1 snRNA is used in either context (Roca and Krainer, 2009; Roca et al., 2012). Besides this rare exception, it would appear that natural selection of 5′ss sequences in the human transcriptome has already explored virtually the entire set of functional 9-nt 5′ss sequences.

In addition to base-pairing to the canonical 5′ss motif, we determined the number of 5′ss sequences in our libraries that

(B) Venn diagrams showing contextual overlap in the number of 5′ss with activities in the ranges 0–20, 20–80, or 80–100 PSI. A complete 9 × 9 table of such overlaps is provided in Figure S5A.

(C) Sequence logo generated from 5′ss sequences with PSI >20 in each context. Separate sequence logos for each independent-replicate library are shown in Figures S5B–S5D.

(D) Heatmap showing squared Spearman rank correlation values (ρ^2) between the PSI measurements in each minigene context, and the predictions of previously published models, including a maximum entropy model (MaxEnt; Yeo and Burge, 2004), a maximum dependence decomposition model (MDD; Burge, 1998), a first-order Markov model (MM; Krogh et al., 1994), a weight matrix model (WMM), and RNAhybrid predictions (RNAhyb; Krüger and Rehmsmeier, 2006). Scatterplots for MPSA versus model comparisons are shown in Figure S6A.

(E) Scatterplots comparing the occurrence of each 5′ss in the human transcriptome, normalized to the occurrence of the respective 9-mer in the genome, to our measured PSI values. Here "n" indicates the number of 5′ss sequences with >50 PSI (right of the red dotted line) that do not occur in the human transcriptome (below the blue dotted line). A cutoff of PSI >50 (marked by the red dash line) was chosen to disregard the population of 5′ss with low activity seen only in the *SMN1* context. See also Figure S6C.

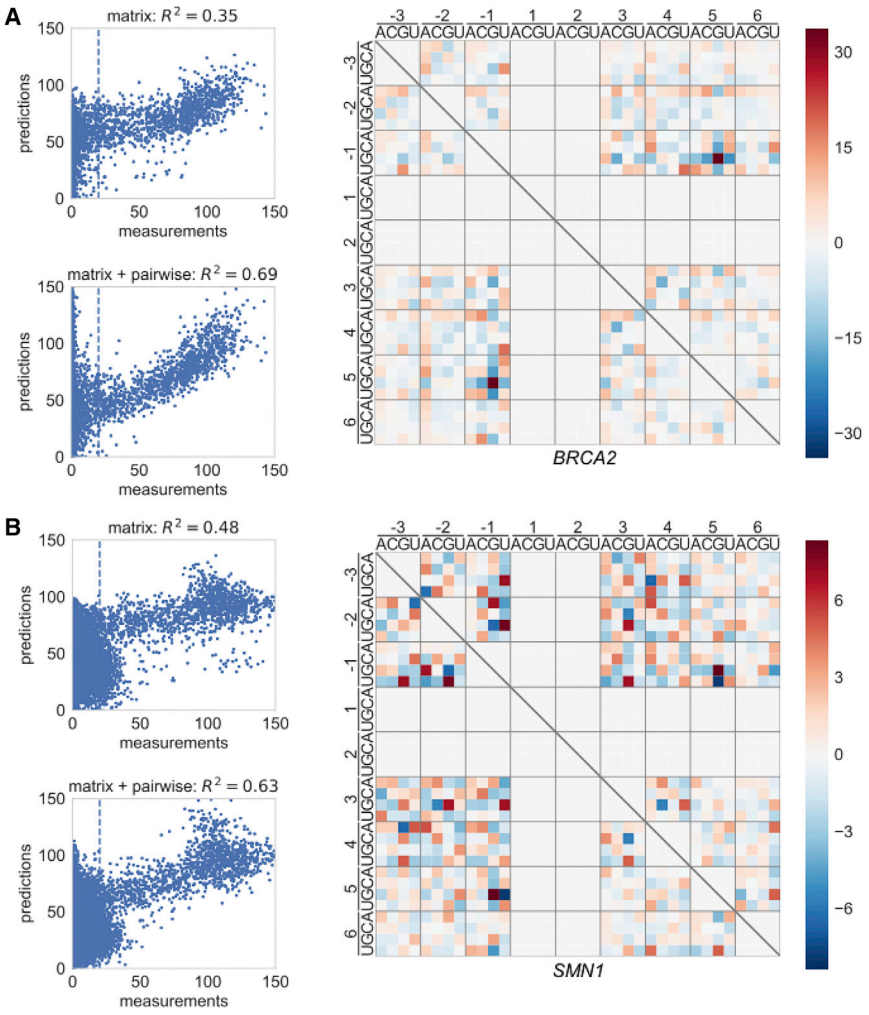


Figure 4. Epistatic Interactions in 5′ss Activity

(A) Scatterplots showing measured PSI values versus PSI values predicted by either the matrix model (top) or the matrix + pairwise model (bottom) for *BRCA2*. The heatmap shows the specific interactions present in the pairwise model. Red indicates a positive interaction, and blue indicates a negative interaction. Note that pairwise models were inferred only for GU splice sites. Analyses for each separate library are shown in Figure S7C.

(B) Same as (A), but for *SMN1*. See also Figure S7D.

can be recognized by a shifted register, in which 5′ss base-pairing with U1 snRNA is shifted by one position, so that +1G base-pairs to U1 C9 instead of C8 (Roca and Krainer, 2009; Figure 1B). We restricted this analysis to GU 5′ss only. Taking into account the fixed flanking sequences in each context, we analyzed 480 shifted-register 5′ss in *BRCA2*, 32 in *SMN1*, and 32 in *IKBKAP*. Only 5 out of 480 (1.0%) of the shifted-register 5′ss in *BRCA2* gave PSI ≥20. One of the productive shifted-register 5′ss was ACA/GUUAAG, previously validated as an atypical 5′ss sequence recognized by a shifted register (Roca and Krainer, 2009). None of the 32 shifted-register 5′ss in *SMN1* and *IKBKAP* were productively recognized for splicing. This analysis indicates that 5′ss are typically recognized via the normal register, although a shifted register is possible in rare cases.

We next asked whether 5′ss that occur more frequently in the human transcriptome tend to be stronger in our libraries. It was previously reported that accounting for the occurrence frequency of a 5′ss sequence can improve prediction accuracy (Sahashi et al., 2007). The *IKBKAP* context had the most stringent requirement for 5′ss selection, with only 250 5′ss having PSI ≥20. Almost all of these 250 5′ss had PSI ≥80 in both *BRCA2* and *SMN1*, suggesting that these are strong 5′ss sequences. However, we did not observe a strong correlation between the number of natural occurrences of these 250 5′ss sequences in the human transcriptome (ranging from 1 to 2,261 occurrences, with a median value of 280) and their inherent strengths.

Pairwise Nucleotide Association Influences 5′ss Strength

Pairwise nucleotide-position dependencies within the 5′ss were previously observed by comparative genomics (Burge and Karlin, 1997; Carmel et al., 2004; Roca et al., 2008). Identifying such interactions could potentially help predict the impact of 5′ss mutations on splicing.

To investigate the presence of such interactions in our MPSA data, we compared the explanatory power of two different models for PSI, as a function of 5′ss sequence: a "matrix model" and a "matrix+pairwise model." The matrix model assumes that each base at each position has an independent and additive effect on PSI (Equation 1, STAR Methods). By contrast, the matrix+pairwise model includes contributions to PSI from all pairs of nucleotide positions within the 5′ss (Equation 2, STAR Methods). (See STAR Methods for details on how these models were fit to the data.)

The matrix model was fit by regressing sequences against PSI measurements. The results of this model fitting largely recapitulated the motifs that we found by aligning active 5′ss. Only 35% and 48% of the variance in 5′ss PSI measurements in *BRCA2* and *SMN1*, respectively, can be attributed to the independent nucleotide contributions in the matrix model (Figures 4A, 4B, and S7). The pairwise component of the matrix+pariwise model was then fit to the residuals of this matrix model. Adding pairwise interactions increased R^2 to 69% and 63% in *BRCA2* and *SMN1*, respectively. The pairwise interactions in *BRCA2* were prominent. In particular, we identified a distinct and preferential association between −1G and +5G (Figure 4A). Remarkably, all three of the characterized disease-associated mutations in *BRCA2* intron 17 (−1G > C, −1G > A, and +5G > A) alter this −1G +5G interaction. The

pairwise interactions in *SMN1* were subtler, with preferences at several pairs of positions and nucleotides (Figure 4B). Due to the limited productive 5′ss usage in *IKBKAP* (only 250 5′ss with PSI ≥20), significant pairwise interactions could not be determined for this context. GC 5′ss were also excluded from this analysis for the same reason.

Notably, the strong positive association between −1G and +5G in *BRCA2* is also evident in the *SMN1* and *IKBKAP* contexts. We compared the usage of 5′ss with either a fixed −1G paired with +5U/C/A or a fixed +5G paired with −1U/C/A to their counterparts with −1G +5G. For this analysis, we only considered 5′ss that gave PSI ≥20 with −1G +5G, which resulted in 1,647 *BRCA2*, 1,818 *SMN1*, and 531 *IKBKAP* 5′ss sequences. When G was fixed at the −1 position, 87.1% of *BRCA2*, 85.8% of *SMN1*, and 97.6% of *IKBKAP* 5′ss sequences with +5U/C/A had a ≥20% reduction in PSI compared to +5G; when G was fixed at the +5 position, 94.7% of *BRCA2*, 86.5% of *SMN1*, and 94.9% of *IKBKAP* 5′ss sequences with −1U/C/A had a ≥20% reduction in PSI compared to −1G.

Indeed, the absence of a G at −1 and +5 positions is highly unfavorable (only 0.5% of NNH/GYNNHN 5′ss sequences in *BRCA2*, 5.7% in *SMN1*, and 0.03% in *IKBKAP* have PSI >20). Among the 5′ss sequences with either −1G or +5G (NNG/GYNNHN and NNH/GYNNGN), 5.6% in *BRCA2*, 10.6% in *SMN1*, and 0.7% in *IKBKAP* have PSI >20. 5′ss sequences with both −1G and +5G (NNG/GYNNGN) are the most favorable (24.6% in *BRCA2*, 25.8% in *SMN1*, and 7.7% in *IKBKAP* have PSI >20). It thus appears that a −1G +5G pairing is generally preferred and reflects a fundamental aspect of 5′ss recognition.

Other epistatic interactions were evident in both the *BRCA2* and *SMN1* contexts, including strong positive interactions between −3G −1U, −3G +4U, −1U +4U, −1G +5G, and +5G +6U, as well as strong negative interactions between −2G −1G, −2A +5G, and −1G +3C. We also observed context-dependent epistatic interactions exclusively in the *BRCA2* or *SMN1* context. For example, −2A in *BRCA2* prefers +4C but negatively interacts with +4A; in contrast, −2A in *SMN1* prefers +4A but negatively interacts with the other 3 nt. Similarly, +4U in *BRCA2* prefers +3U, but +4U in *SMN1* prefers +3G. The mechanisms underlying these general or context-dependent pairwise interactions are presently unknown.

There are multiple potential explanations for these observed epistatic interactions. They might result from mechanistic coupling between nucleotide pairs. Alternatively, they could result from nonlinearities in the relationship between PSI and some intermediate non-epistatic phenotype (e.g., U1 snRNP-5′ss binding energy), a phenomenon known as "global epistasis." More sophisticated quantitative modeling (e.g., along the lines of Otwinowski et al., 2018) might help distinguish between these possibilities in the future.

A Weak Upstream 3′ss Drives the Context Dependency of 5′ss Sequence Usage in *IKBKAP*

Recognition of the 5′ss in the *IKBKAP* context strikingly differs from that in the *BRCA2* and *SMN1* contexts. Many features can contribute to the overall context and influence the recognition of a 5′ss, such as 3′ss strength (Will and Lührmann, 2011), the presence of various exonic and intronic enhancers and/or silencers (Hastings and Krainer, 2001), and RNA secondary structure (Buratti and Baralle, 2004).

In the case of *IKBKAP* exon 20, the upstream 3′ss is predicted to be weaker than that of *BRCA2* or *SMN1* (Table S2), whereas the downstream 3′ss sequences are expected to have similar strengths in all three contexts as judged by the MaxEnt algorithm (Yeo and Burge, 2004). To examine the contribution of these 3′ss and other sequence elements to 5′ss selection, we selected 10 5′ss sequences from the 53 manually validated 5′ss (Figure 2F) that had PSI >80 in both *BRCA2* and *SMN1* but had PSI <50 in *IKBKAP*, presuming that the low efficiency of these 5′ss is due to the *IKBKAP* context. We then introduced various portions of *BRCA2* or *SMN1* into the *IKBKAP* minigene in place of the corresponding *IKBKAP* sequences (Figure 5A). To ensure that the expected splice junctions were used in the hybrid minigenes, we gel-purified and analyzed the splice products by Sanger sequencing. This analysis confirmed that only the expected splice junctions were used, and no usage of cryptic sites was observed. Notably, we found that replacing a 20-nt *IKBKAP* sequence comprising the upstream 3′ss, with either *BRCA2* or *SMN1* sequences strongly promoted exon inclusion with all 10 5′ss sequences examined (Figure 5B, construct 1). Whereas the *BRCA2* exonic sequence alone modestly increased exon inclusion with most of the tested 5′ss, *SMN1* exonic sequence resulted in further inhibition of exon inclusion (Figure 5B, construct 3). Using antisense oligonucleotides, it was previously determined that two silencer elements are present within *SMN1/2* exon 7, flanking an exonic splicing enhancer (Hua et al., 2007). The net effect of these opposing elements appears to be repressive in the context of the hybrid minigenes, promoting exon skipping.

Replacement of the downstream *IKBKAP* intron (intron 20) with the corresponding intron from *BRCA2* (intron 17) resulted in variable extents of exon 20 inclusion with a subset of the 5′ss, whereas replacement with the *SMN1* downstream intron (intron 7) resulted in predominantly unspliced products (not shown) (Figure 5B, construct 6). We conclude that context plays a considerable role in 5′ss recognition in *IKBKAP*, and strengthening the upstream 3′ss of the *IKBKAP* minigene was sufficient to increase 5′ss activity to similar levels as in *BRCA2* and *SMN1*. Taken together, our results suggest that 5′ss selection is relatively predictable, in the absence of strong contextual influences.

Library Results Help to Predict Pathogenic Mutations

To assess the effectiveness of our MPSA measurements in predicting the consequences of 5′ss mutations, we compiled a list of 122 pathogenic 5′ss mutations (excluding +1 mutations, as well as +2U > A, +2U > G, +2C > A, and +2C > G mutations, which are known to abolish splicing) and 103 5′ss mutations with unclassified or unknown consequences identified throughout the *BRCA1* and *BRCA2* transcripts (Landrum et al., 2016). We then compared the PSI values we measured for these 5′ss sequences in our libraries to the PSI values of their wild-type counterparts in each of the three minigene contexts (Figures 6A and 6B). We only analyzed the 5′ss mutations for which the corresponding wild-type 5′ss sequences had ≥20 measured PSI values (excluded data points are in the gray-shaded area). We thus examined

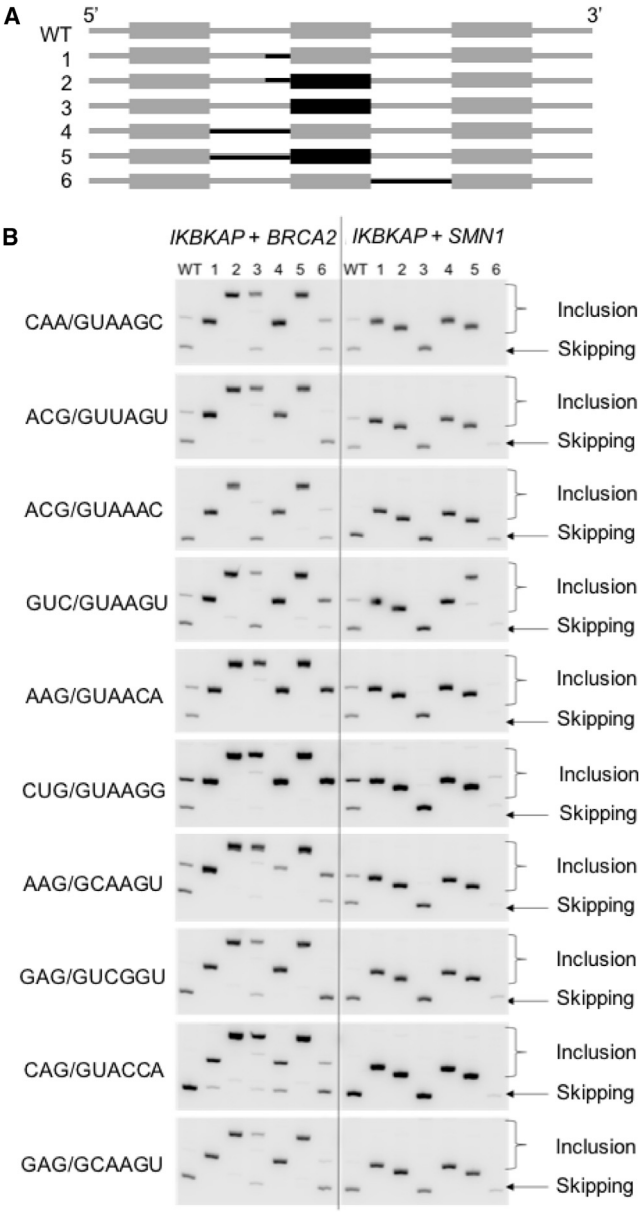


Figure 5. A Weak Upstream 3′ss Drives the Context Dependence of 5′ss Activity in *IKBKAP*

(A) Diagram of hybrid minigene constructs. *IKBKAP* minigene sequences are illustrated in gray. Black indicates either *BRCA2* or *SMN1* minigene sequences replacing the corresponding *IKBKAP* features.

(B) Splicing of the hybrid constructs is shown in the RT-PCR gels. Due to the size differences of the middle exon between constructs, the size of the inclusion band varies. Gel images are representative of triplicates.

122 *BRCA1* or *BRCA2* pathogenic 5′ss mutations in *BRCA2*, 109 in *SMN1*, and 76 in *IKBKAP*, along with 103 5′ss mutations with unknown consequences in *BRCA2*, 101 in *SMN1*, and 62 in *IKBKAP*. Among these, 86% of the pathogenic mutations in the *BRCA2*, 73% in the *SMN1*, and 93% in the *IKBKAP* context caused a $\geq$20% reduction in PSI, compared to the respective wild-type 5′ss sequences (Figure 6A). This is significantly different from the 40% of mutations with unknown consequences in the *BRCA2* ($p = 2.54 \times 10^{-13}$, Fisher's exact test), 38% in the *SMN1* ($p = 2.06 \times 10^{-7}$), and 57% in the *IKBKAP* context ($p = 2.87 \times 10^{-7}$) that caused a $\geq$20% reduction in PSI compared to the respective wild-type 5′ss sequences (Figure 6B). This analysis shows that we can use our library data to clearly distinguish known disease-associated mutations from unannotated mutations.

We further examined a selected set of 147 known disease-causing mutations across a broad range of genes and diseases (available at the DBASS5 online resource) (Figure 6C) (Buratti et al., 2007). In addition to the same exclusions as the previous analysis, we also excluded *BRCA1/2* mutations to avoid redundancy, and mutations that generated a *de novo* 5′ss. Among these 147 mutations, 128 in *BRCA2*, 128 in *SMN1*, and 65 in *IKBKAP* had corresponding wild-type 5′ss sequences with PSI $\geq$20. Consistently, our data show a $\geq$20% reduction in PSI in 93% of the mutations in the *BRCA2*, 91% in the *SMN1*, and 95% in the *IKBKAP* context. Finally, we compared the measured PSI value for major and minor variants of 625 common SNPs with a minor-allele frequency in the human population >10% (obtained from the ExAC database) (Lek et al., 2016) (Figure 6D). We only analyzed the 5′ss in which both the major and minor variants yielded $\geq$20 measured PSI values, which resulted in 515 in the *BRCA2*, 501 in the *SMN1*, and 229 in the *IKBKAP* context. Only 30% of the minor variants in the *BRCA2*, 29% in the *SMN1*, and 38% in the *IKBKAP* context deviated more than 20% in PSI from the respective major variants. Thus, our library data can be used to accurately predict the likely functional consequences of 5′ss mutations across different introns and genes.

DISCUSSION

To elucidate the mechanism of 5′ss recognition and determine the characteristics of 5′ss sequences that are prone to perturbation by point mutations, we measured the 5′ss recognition profile of nearly all 32,768 unique 9-nt GU or GC 5′ss sequences in three heterologous gene contexts (*BRCA2*, *SMN1*, and *IKBKAP*). Although high-throughput mutation studies have become increasingly common, reflecting advances in sequencing technology, the complexity of splicing regulation can confound the interpretation of such massive datasets. In particular, such data are frequently used for holistic modeling that integrates effects across multiple sequence elements, making it difficult to attribute an observed effect to a specific element. Due to the binding of splicing regulators to degenerate sequence motifs throughout the pre-mRNA, a mutation designed to abolish a regulator binding site may unintentionally create another. In addition, the presence of a binding motif is not necessarily indicative of productive binding by the cognate factor. For these and other reasons, it is difficult to accurately predict the effect of specific mutations when multiple random mutations are introduced sparsely throughout the pre-mRNA. These limitations motivated us to adopt a focused approach and directly measure the effects of all possible variations of one element, the 5′ss, which has a well-defined location and length. Our method allows for a transparent assessment of how individual 5′ss sequences affect splicing and how gene context can alter these effects.

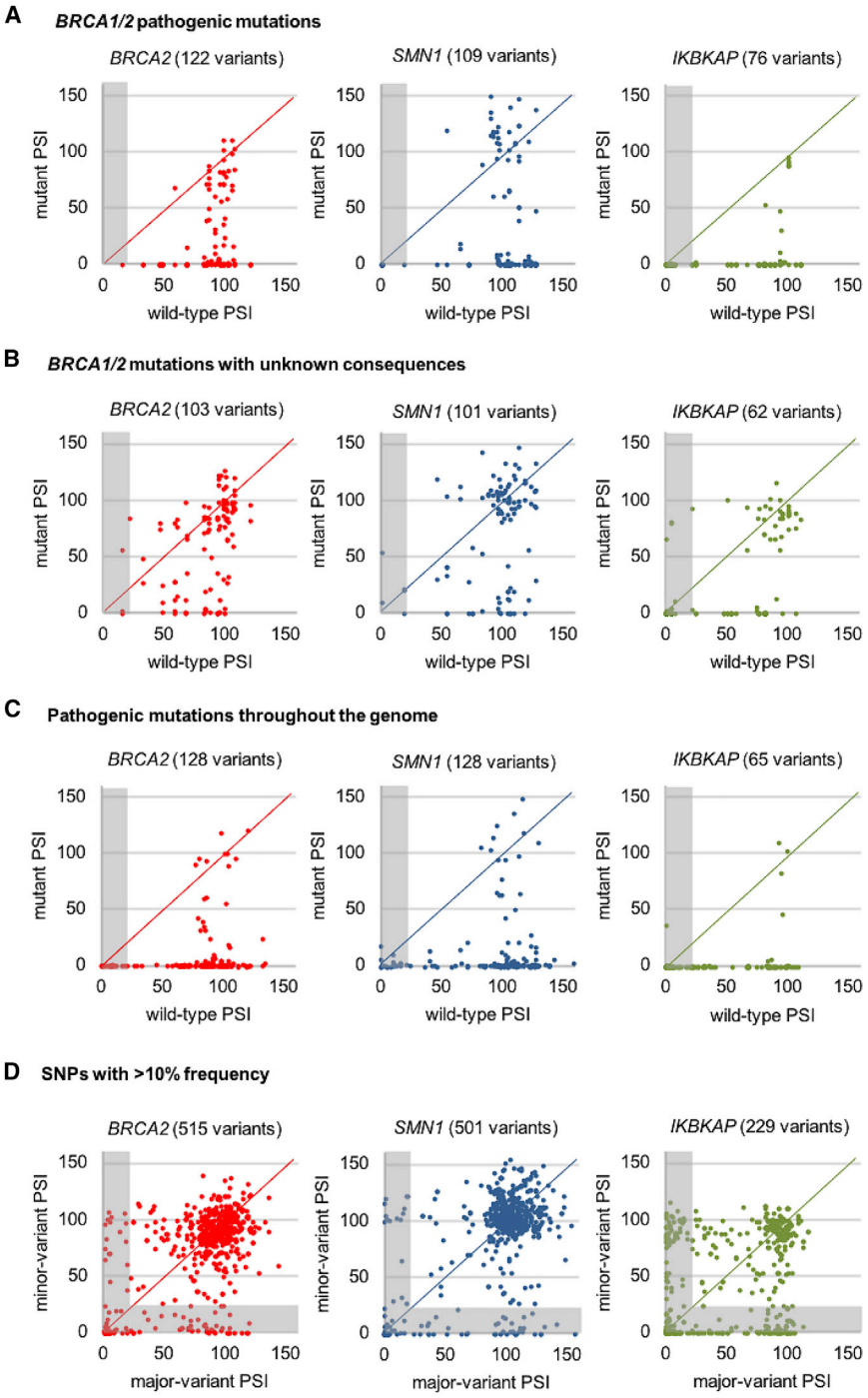


Figure 6. MPSA Measurements Help to Predict Pathogenic Mutations

(A) Scatterplots, corresponding to the three minigene contexts, comparing MPSA-measured PSI for multiple wild-type 5′ss sequences to mutant in *BRCA1* and *BRCA2* to mutant 5′ss variants thereof that are known to be pathogenic. Gray-shaded area indicates data points with wild-type PSI <20, which were excluded from the subsequent analysis. See also Figure S6B.

(B) Same as above, but for mutant *BRCA1* and *BRCA2* 5′ss sequences with unclassified or uncertain clinical significance. Gray-shaded area indicates data points with wild-type PSI <20, which were excluded from the analysis. See also Figure S6B.

(C) Same as above, but for known disease-causing mutations across a broad range of genes and diseases, available from the DBASS5 online resource (Buratti et al., 2007). Gray-shaded area indicates data points with wild-type PSI <20, which were excluded from the analysis. See also Figure S6B.

(D) Same as above, but for 5′ss SNPs with >10% frequency found in the human population, compiled from the ExAC database (Lek et al., 2016). Gray-shaded area indicates data points for which either the major or minor variants had PSI < 20, which were excluded from the analysis. See also Figure S6B.

Our study confirmed that the major determinant of 5′ss activity is largely inherent to its nucleotide sequence—as suggested by the high similarity in 5′ss strength between *BRCA2* and *SMN1*—but also revealed substantial context-dependent differences in 5′ss activity. Much of this context-dependent variation is likely contributed by 3′ss strength and the presence of various exonic and intronic enhancers and/or silencers. In particular, we determined that a weak upstream 3′ss in *IKBKAP* muddled exon definition and prevented the usage of the majority of 5′ss sequences. Strengthening the *IKBKAP* upstream 3′ss recovered 5′ss recognition to a similar extent as in *BRCA2* and *SMN1*. The inherent nature of 5′ss sequences was also evident in our finding that there were virtually no functional sequences used in our libraries that do not occur naturally as bona fide 5′ss sequences in the human transcriptome. In a dataset of 202,764 human authentic 5′ss sequences (Roca et al., 2012), there are 4,141 naturally occurring 9-nt human GU (25.3%) and 49 GC (0.3%) 5′ss sequences that are used at least three times in the human transcriptome out of all possible unique GU or GC 5′ss sequences (16,384 each). In our *BRCA2* experiments, we identified 9.97% (1,634/16,384) of GU and 1.1% (177/16,384) of GC 5′ss sequences with at least 5% splicing activity. With just one exception, all of the active 5′ss sequences in our *BRCA2* context also occur naturally. This observation suggests that our results can closely simulate the usage of naturally occurring 5′ss sequences.

Intriguingly, however, many natural 5′ss sequences were not active in our experiments. Out of 4,141 authentic 5′ss sequences in the human transcriptome, 2,181 (53.7%) had measured PSI values lower than 5% in all three contexts we examined. The

efficient use of these 5′ss sequences in their natural contexts may require features associated with specific cell types, physiological conditions, cell-cycle status, or differentiation state. Splicing regulation in general relies on combinatorial interactions involving multiple *trans*-acting splicing activators and repressors that bind to their corresponding *cis*-acting enhancer and silencer elements, respectively. Each of the above conditions could result in altered expression or activity of various splicing regulators, the net effect of which would lead to efficient use of 5′ss sequences that showed little to no activity in HeLa cells. The precise requirements for efficient use of these 5′ss sequences warrants further investigation (e.g., by comparing their activities in additional gene contexts, different cell types, and physiological conditions).

In addition to the inherent contribution of the 5′ss nucleotide sequence, gene context can play a considerable role in determining 5′ss activity. Although 5′ss activity was similar between the *BRCA2* and *SMN1* contexts, key contextual differences were evident in the large fraction of GC 5′ss with low to moderate activity in *SMN1* that were not active in *BRCA2*. The evidently permissive nature of *SMN1* to a broad spectrum of 5′ss sequences, especially GC 5′ss, warrants further investigation. The usage of 5′ss sequences in *IKBKAP* further demonstrated substantial context dependency. Notably, the three contexts we examined are constitutive exons in their natural pre-mRNAs. It was previously determined that the 5′ss strengths of alternative exons are only marginally weaker than those of constitutive exons (Wang et al., 2005; Roca et al., 2012). Rather, context may play a more crucial role in the strength of a given 5′ss sequence in alternative exons. In this study, we observed that the rank order of the strength of a 5′ss is largely intrinsic to its sequence, whereas the particular context imposes an activity threshold. It will be of interest in future studies to generalize whether 5′ss recognition relies more on the nucleotide sequence and is fine-tuned by context (as in *BRCA2* and *SMN1*) or whether 5′ss usage that is strongly driven by context (as in *IKBKAP*) is the more prevalent situation in the human transcriptome.

Even though our MPSA experiments comprehensively characterized both GU and GC 5′ss sequences, only a minor subset of possible 9-nt GC 5′ss sequences were recognized as functional 5′ss. After U1 snRNP is displaced from the 5′ss (Staley and Guthrie, 1999), +2U of the 5′ss base-pairs with A51 of U6 snRNA in the spliceosomal C* complex (Fica et al., 2017; Sontheimer and Steitz, 1993). Mutations at both +2U of the 5′ss and A51 of U6 snRNA prevent exon ligation in yeast (Collins and Guthrie, 2001; Siatecka et al., 1999). In addition to base-pairing to U1 snRNA, this role of +2U in U6 binding, though not strictly required, may explain why GC 5′ss are suboptimal.

Based on the epistasis analysis we performed using our *BRCA2* and *SMN1* measurements, we expect that mutations that disrupt positive pairwise interactions (such as the strong −1G+5G interaction) may have a negative effect on splicing efficiency. The interaction between positions −1 and +5 was previously described on the basis of comparative genomics between human and mouse (Burge and Karlin, 1997; Carmel et al., 2004; Roca et al., 2008). Previously, a ‘‘seesaw linkage’’ pattern was observed, whereby −1G permits any nucleotide at position +5, and conversely, +5G permits any nucleotide at −1; in our study, we observed a strong positive interaction between −1G and +5G, such that a nucleotide change at either position results in a ≥20% reduction in PSI compared to the respective −1G and +5G counterpart. Whereas previous reports of such pairwise interactions were based on the statistics of aligned 5′ss sequences from across the genome, we defined couplings on the basis of direct functional measurements of PSI. Our comprehensive approach allowed us to refine this relationship between −1 and +5 positions, in which having Gs at both −1 and +5 positions is highly preferential, having a G at either position results in a reduction in PSI, and having a G at neither position is highly unfavorable. The mechanistic reason for the almost exclusive positive interaction between −1G and +5G is yet to be determined. We speculate that the disruption of the more stable G-C base-pairing at the −1 and +5 positions may contribute to the strong dependency observed. Additionally, this interaction may reflect structural constraints that could be revealed by ongoing structural studies of the spliceosome (Fica et al., 2017). Though some additional interactions were context dependent, we observed other shared pairwise interaction patterns between *BRCA2* and *SMN1*. The precise interactions that may apply to a wide variety of 5′ss sequences will need to be elucidated in future work, with larger datasets. Finally, we observed that 5′ss efficiency largely follows a bimodal distribution, such that the majority of 5′ss were used with either high activity or little to no activity. This finding helps to explain why splicing mutations can have such detrimental effects and underscores the need to characterize SNPs that may strongly alter splicing and thus cause or contribute to disease development.

Besides improving our understanding of the mechanism of 5′ss recognition and splicing, the present findings also have translational relevance. As genetic screening and whole-genome sequencing emerge as common practice, there is a great need to determine which SNPs contribute to disease development. One of the prime examples is the genetic screening for mutations in the tumor-suppressor genes *BRCA1* and *BRCA2*, which allows preventive action based on the inherited risk of developing breast or ovarian cancer. However, even for these heavily studied genes, many of the 1,277 and 1,331 point mutations, respectively, that have been identified to date, remain as ‘‘unclassified variants of unknown significance’’ (University of Utah Department of Pathology and ARUP Laboratories, 2014). For most disease-related genes, the currently available data are insufficient to assess the significance of SNPs discovered by genetic screening. Using the data we collected for *BRCA2*, we can predict that 86.1% of 122 5′ss mutations annotated as pathogenic throughout the *BRCA1* and *BRCA2* transcripts, and 93% of 147 known disease-causing mutations across a broad range of genes and diseases do indeed result in a reduction of PSI compared to the respective wild-type sequence. In addition, we can clearly segregate another 103 SNPs at 5′ss with unclassified or unknown consequences into likely benign and likely pathogenic mutations. By contrast, we observed little to no deviation in the measured PSI value in the majority of major and minor variants of 515 common 5′ss SNPs present at >10% in the human population. This analysis suggests that despite observing context-dependent effects on 5′ss recognition, our data can nonetheless help to predict the likely

functional consequences of 5′ss mutations throughout the human transcriptome.

Our measurements of 5′ss activity slightly outperformed various computational scoring matrices in predicting 5′ss usage. Although some of these computational methods can similarly segregate pathogenic from likely nonpathogenic mutations (Figure S6B), one key distinctive feature of our method is the direct quantification of PSI. Indeed, our data often revealed orders of magnitude of change in PSI, which are not reflected in the scores assigned by existing computational methods. Notwithstanding the limited contexts examined, our quantitative measurements of 5′ss activity will likely have clinical applications.

Promising therapeutic approaches for reversing certain splicing defects are being pursued. Small-molecule enhancers of splicing stabilize the binding of U1 snRNP to *SMN2* pre-mRNA, increasing full-length SMN mRNA and protein (Palacino et al., 2015; Sivaramakrishnan et al., 2017). Nusinersen (Spinraza) is an antisense oligonucleotide that increases splicing of full-length *SMN2* mRNA in patients lacking functional *SMN1* and is the first and so far only treatment for spinal muscular atrophy that has been approved by the US Food and Drug Administration and the European Medicines Agency (Finkel et al., 2017; Hua et al., 2008). Our comprehensive analysis of 5′ss improves our ability to predict which mutations are likely to affect splicing and are therefore potentially amenable to similar splicing-corrective therapies. To carry forward the momentum of developing successful therapies, it is essential to precisely and efficiently identify disease-associated mutations and SNPs on which to focus therapeutic efforts. Therefore, it will be of interest to expand the systematic method we have established in this study to other splicing regulatory elements, such as the 3′ss. The cumulative data on the usage of every possible 5′ss is an important step toward elucidating the "splicing code" (Wang and Burge, 2008) and will facilitate predictions of the outcome of 5′ss mutations and SNPs for risk assessment and development of targeted therapeutics.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Construction and transfection of minigene plasmids
  - RT-PCR and splicing analysis
  - Construction and sequencing of libraries
  - *In silico* analysis of shifted-register 5′ss
  - Primers
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Read parsing
  - 5′ss-barcode association
  - PSI quantification
  - Junction quantification
  - Pairwise dependency
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and three tables and can be found with this article online at https://doi.org/10.1016/j.molcel.2018.07.033.

ACKNOWLEDGMENTS

The HeLa FRT clone was a generous gift from Dr. Woodring Wright and Dr. Jerry Shay (UT Southwestern Medical Center, TX). We sincerely thank Dr. Xavier Roca (Nanyang Technological University, Singapore) and Dr. David McCandlish (Cold Spring Harbor Laboratory, NY) for comments on the manuscript. We also thank Dr. Sara Ballouz (CSHL, NY) for assistance in parsing the ExAC database. This work was supported by the NIH-NIGMS (grants 5F32GM116372-03 and R37GM42699). J.B.K. and A.R.K. are members of the CSHL Cancer Center, which is supported by NIH Cancer Center Support Grant 5P30CA045508.

AUTHOR CONTRIBUTIONS

M.S.W., J.B.K., and A.R.K. conceived and designed the experiments. M.S.W. performed the experiments under the guidance of J.B.K. and A.R.K. J.B.K. and M.S.W. performed the computational analysis. M.S.W. wrote the manuscript, with input and edits on successive drafts from J.B.K. and A.R.K.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 23, 2018
Revised: June 18, 2018
Accepted: July 23, 2018
Published: August 30, 2018

REFERENCES

Anderson, S.L., Coli, R., Daly, I.W., Kichula, E.A., Rork, M.J., Volpi, S.A., Ekstein, J., and Rubin, B.Y. (2001). Familial dysautonomia is caused by mutations of the IKAP gene. Am. J. Hum. Genet. *68*, 753–758.

Bao, P., Hobartner, C., Hartmuth, K., and Lührmann, R. (2017). Yeast Prp2 liberates the 5′ splice site and the branch site adenosine for catalysis of pre-mRNA splicing. RNA *23*, 1770–1779.

Bertram, K., Agafonov, D.E., Dybkov, O., Haselbach, D., Leelaram, M.N., Will, C.L., Urlaub, H., Kastner, B., Lührmann, R., and Stark, H. (2017). Cryo-EM structure of a pre-catalytic human spliceosome primed for activation. Cell *170*, 701–713.e11.

Buratti, E., and Baralle, F.E. (2004). Influence of RNA secondary structure on the pre-mRNA splicing process. Mol. Cell. Biol. *24*, 10505–10514.

Buratti, E., Chivers, M., Královicová, J., Romano, M., Baralle, M., Krainer, A.R., and 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.

Burge, C. (1998). Modeling dependencies in pre-mRNA splicing signals, in Computational Methods in Molecular Biology, S.L. Salzberg, D.B. Searls, and S. Kasif, eds. (Elsevier Science), pp. 129–164.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. *268*, 78–94.

Carmel, I., Tal, S., Vig, I., and Ast, G. (2004). Comparative analysis detects dependencies among the 5′ splice-site positions. RNA *10*, 828–840.

Cartegni, L., Chew, S.L., and Krainer, A.R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat. Rev. Genet. *3*, 285–298.

Collins, C.A., and Guthrie, C. (2001). Genetic interactions between the 5′ and 3′ splice site consensus sequences and U6 snRNA during the second catalytic step of pre-mRNA splicing. RNA *7*, 1845–1854.

Fica, S.M., Oubridge, C., Galej, W.P., Wilkinson, M.E., Bai, X.C., Newman, A.J., and Nagai, K. (2017). Structure of a spliceosome remodelled for exon ligation. Nature *542*, 377–380.

Finkel, R.S., Mercuri, E., Darras, B.T., Connolly, A.M., Kuntz, N.L., Kirschner, J., Chiriboga, C.A., Saito, K., Servais, L., Tizzano, E., et al.; ENDEAR Study Group (2017). Nusinersen versus sham control in infantile-onset spinal muscular atrophy. N. Engl. J. Med. *377*, 1723–1732.

Freund, M., Hicks, M.J., Konermann, C., Otte, M., Hertel, K.J., and 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.

Hartmann, L., Theiss, S., Niederacher, D., and Schaal, H. (2008). Diagnostics of pathogenic splicing mutations: does bioinformatics cover all bases? Front. Biosci. *13*, 3252–3272.

Hastie, T., Tibshirani, R., and Friedman, J. (2011). The Elements of Statistical Learning, Second Edition (Springer).

Hastings, M.L., and Krainer, A.R. (2001). Pre-mRNA splicing in the new millennium. Curr. Opin. Cell Biol. *13*, 302–309.

Hofmann, W., Horn, D., Hüttner, C., Classen, E., and Scherneck, S. (2003). The *BRCA2* variant 8204G>A is a splicing mutation and results in an in frame deletion of the gene. J. Med. Genet. *40*, e23.

Hua, Y., Vickers, T.A., Baker, B.F., Bennett, C.F., and Krainer, A.R. (2007). Enhancement of SMN2 exon 7 inclusion by antisense oligonucleotides targeting the exon. PLoS Biol. *5*, e73.

Hua, Y., Vickers, T.A., Okunola, H.L., Bennett, C.F., and Krainer, A.R. (2008). Antisense masking of an hnRNP A1/A2 intronic splicing silencer corrects SMN2 splicing in transgenic mice. Am. J. Hum. Genet. *82*, 834–848.

Julien, P., Miñana, B., Baeza-Centurion, P., Valcárcel, J., and Lehner, B. (2016). The complete local genotype-phenotype landscape for the alternative splicing of a human exon. Nat. Commun. *7*, 11558.

Ke, S., Anquetil, V., Zamalloa, J.R., Maity, A., Yang, A., Arias, M.A., Kalachikov, S., Russo, J.J., Ju, J., and Chasin, L.A. (2018). Saturation mutagenesis reveals manifold determinants of exon definition. Genome Res. *28*, 11–24.

Ketterling, R.P., Drost, J.B., Scaringe, W.A., Liao, D.Z., Liu, J.Z., Kasper, C.K., and Sommer, S.S. (1999). Reported in vivo splice-site mutations in the factor IX gene: severity of splicing defects and a hypothesis for predicting deleterious splice donor mutations. Hum. Mutat. *13*, 221–231.

Kondo, Y., Oubridge, C., van Roon, A.M., and Nagai, K. (2015). Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5′ splice site recognition. eLife *4*, e04986.

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.

Krawczak, M., Ball, E.V., Fenton, I., Stenson, P.D., Abeysinghe, S., Thomas, N., and Cooper, D.N. (2000). Human gene mutation database-a biomedical information and research resource. Hum. Mutat. *15*, 45–51.

Krogh, A., Brown, M., Mian, I.S., Sjölander, K., and Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. J. Mol. Biol. *235*, 1501–1531.

Krüger, J., and Rehmsmeier, M. (2006). RNAhybrid: microRNA target prediction easy, fast and flexible. Nucleic Acids Res. *34*, W451–W454.

Kubota, T., Roca, X., Kimura, T., Kokunai, Y., Nishino, I., Sakoda, S., Krainer, A.R., and Takahashi, M.P. (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.

Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. *44* (D1), D862–D868.

Lefebvre, S., Bürglen, L., Reboullet, S., Clermont, O., Burlet, P., Viollet, L., Benichou, B., Cruaud, C., Millasseau, P., Zeviani, M., et al. (1995). Identification and characterization of a spinal muscular atrophy-determining gene. Cell *80*, 155–165.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Lerner, M.R., Boyle, J.A., Mount, S.M., Wolin, S.L., and Steitz, J.A. (1980). Are snRNPs involved in splicing? Nature *283*, 220–224.

Nilsen, T.W., and Graveley, B.R. (2010). Expansion of the eukaryotic proteome by alternative splicing. Nature *463*, 457–463.

Otwinowski, J., McCandlish, D.M., and Plotkin, J.B. (2018). Inferring the shape of global epistasis. Proc. Natl. Acad. Sci. U S A. Published online July 23, 2018. https://doi.org/10.1073/pnas.1804015115.

Palacino, J., Swalley, S.E., Song, C., Cheung, A.K., Shu, L., Zhang, X., Van Hoosear, M., Shin, Y., Chin, D.N., Keller, C.G., et al. (2015). SMN2 splice modulators enhance U1-pre-mRNA association and rescue SMA mice. Nat. Chem. Biol. *11*, 511–517.

Roca, X., and Krainer, A.R. (2009). Recognition of atypical 5′ splice sites by shifted base-pairing to U1 snRNA. Nat. Struct. Mol. Biol. *16*, 176–182.

Roca, X., Olson, A.J., Rao, A.R., Enerly, E., Kristensen, V.N., Børresen-Dale, A.L., Andresen, B.S., Krainer, A.R., and Sachidanandam, R. (2008). Features of 5′-splice-site efficiency derived from disease-causing mutations and comparative genomics. Genome Res. *18*, 77–87.

Roca, X., Akerman, M., Gaus, H., Berdeja, A., Bennett, C.F., and Krainer, A.R. (2012). Widespread recognition of 5′ splice sites by noncanonical base-pairing to U1 snRNA involving bulged nucleotides. Genes Dev. *26*, 1098–1109.

Rogers, J., and Wall, R. (1980). A mechanism for RNA splicing. Proc. Natl. Acad. Sci. USA *77*, 1877–1879.

Rosenberg, A.B., Patwardhan, R.P., Shendure, J., and Seelig, G. (2015). Learning the sequence determinants of alternative splicing from millions of random sequences. Cell *163*, 698–711.

Sahashi, K., Masuda, A., Matsuura, T., Shinmi, J., Zhang, Z., Takeshima, Y., Matsuo, M., Sobue, G., and 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.

Sheth, N., Roca, X., Hastings, M.L., Roeder, T., Krainer, A.R., and Sachidanandam, R. (2006). Comprehensive splice-site analysis using comparative genomics. Nucleic Acids Res. *34*, 3955–3967.

Siatecka, M., Reyes, J.L., and Konarska, M.M. (1999). Functional interactions of Prp8 with both splice sites at the spliceosomal catalytic center. Genes Dev. *13*, 1983–1993.

Singh, N.N., Androphy, E.J., and Singh, R.N. (2004). In vivo selection reveals combinatorial controls that define a critical exon in the spinal muscular atrophy genes. RNA *10*, 1291–1305.

Singh, N.N., Del Rio-Malewski, J.B., Luo, D., Ottesen, E.W., Howell, M.D., and Singh, R.N. (2017). Activation of a cryptic 5′ splice site reverses the impact of pathogenic splice site mutations in the spinal muscular atrophy gene. Nucleic Acids Res. *45*, 12214–12240.

Sivaramakrishnan, M., McCarthy, K.D., Campagne, S., Huber, S., Meier, S., Augustin, A., Heckel, T., Meistermann, H., Hug, M.N., Birrer, P., et al. (2017). Binding to SMN2 pre-mRNA-protein complex elicits specificity for small molecule splicing modifiers. Nat. Commun. *8*, 1476.

Slaugenhaupt, S.A., Blumenfeld, A., Gill, S.P., Leyne, M., Mull, J., Cuajungco, M.P., Liebert, C.B., Chadwick, B., Idelson, M., Reznik, L., et al. (2001). Tissue-specific expression of a splicing mutation in the IKBKAP gene causes familial dysautonomia. Am. J. Hum. Genet. *68*, 598–605.

Soemedi, R., Cygan, K.J., Rhine, C.L., Wang, J., Bulacan, C., Yang, J., Bayrak-Toydemir, P., McDonald, J., and Fairbrother, W.G. (2017). Pathogenic variants that alter protein code often disrupt splicing. Nat. Genet. *49*, 848–855.

Sontheimer, E.J., and Steitz, J.A. (1993). The U5 and U6 small nuclear RNAs as active site components of the spliceosome. Science *262*, 1989–1996.

Srebrow, A., and Kornblihtt, A.R. (2006). The connection between splicing and cancer. J. Cell Sci. *119*, 2635–2641.

Staley, J.P., and Guthrie, C. (1999). An RNA switch at the 5′ splice site requires ATP and the DEAD box protein Prp28p. Mol. Cell *3*, 55–64.

Tan, J., Ho, J.X., Zhong, Z., Luo, S., Chen, G., and Roca, X. (2016). Noncanonical registers and base pairs in human 5′ splice-site selection. Nucleic Acids Res. *44*, 3908–3921.

Tarn, W.Y., Yario, T.A., and Steitz, J.A. (1995). U12 snRNA in vertebrates: evolutionary conservation of 5′ sequences implicated in splicing of pre-mRNAs containing a minor class of introns. RNA *1*, 644–656.

Teng, D.H., Bogden, R., Mitchell, J., Baumgard, M., Bell, R., Berry, S., Davis, T., Ha, P.C., Kehrer, R., Jammulapati, S., et al. (1996). Low incidence of BRCA2 mutations in breast carcinoma and other cancers. Nat. Genet. *13*, 241–244.

University of Utah Department of Pathology and ARUP Laboratories. (2014). BRCA mutation database. http://arup.utah.edu/database/BRCA/.

Wan, R., Yan, C., Bai, R., Lei, J., and Shi, Y. (2017). Structure of an intron lariat spliceosome from Saccharomyces cerevisiae. Cell *171*, 120–132.e12.

Wang, Z., and Burge, C.B. (2008). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA *14*, 802–813.

Wang, J., Smith, P.J., Krainer, A.R., and Zhang, M.Q. (2005). Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes. Nucleic Acids Res. *33*, 5053–5062.

Will, C.L., and Lührmann, R. (2011). Spliceosome structure and function. Cold Spring Harb. Perspect. Biol. *3*, a003707.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. *11*, 377–394.

Zhuang, Y., and Weiner, A.M. (1986). A compensatory base change in U1 snRNA suppresses a 5′ splice site mutation. Cell *46*, 827–835.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
Lipofectamine 2000 Transfection Reagent	Invitrogen	Cat#11668019
Hygromycin B	Invitrogen	Cat#10687010
Gibson Assembly Master Mix	NEB	Cat#E2611L
Phusion High-Fidelity DNA Polymerase	NEB	Cat#M0530L
TRIzol	Life Technologies	Cat#15596018
Improm-II Reverse Transcription System	Promega	Cat#A3800
Deposited Data		
Custom Python scripts	This paper	https://github.com/jbkinney/15_splicing
Original data in Medeley Data	This paper	https://doi.org/10.17632/z25p7f4zvt.1
Illumina sequencing data	This paper	SRA:SRP135892
Experimental Models: Cell Lines		
HeLa	ATCC	N/A
DH5-α, MegaX DH10B T1 Electrocomp Cells	ThermoFisher	Cat#C640003
Oligonucleotides		
See Table S3 for primer information		
Recombinant DNA		
pcDNA5/FRT	Invitrogen	
BRCA2 ex16-18 minigene	This paper	N/A
SMN1 ex6-8 minigene with IVS24G > C	This paper	N/A
IKBKAP ex19-21 minigene	This paper	N/A
BRCA2 ex5-7 minigene	This paper	N/A
Software and Algorithms		
ImageJ	NIH	N/A

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by Adrian Krainer (krainer@cshl.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

HeLa cervical carcinoma cells were cultured at 37°C in 5% CO_2 in Dulbecco's Modification of Eagle's Medium (DMEM, Corning, Manassas, VA), containing 10% fetal bovine serum (Seradigm, Radnor, PA).

METHOD DETAILS

Construction and transfection of minigene plasmids

A HeLa FRT clone selected for a single FRT integration site was a generous gift from Dr. Woodring Wright and Dr. Jerry Shay (UT Southwestern Medical Center at Dallas, TX). Human sequences were inserted into the pcDNA5/FRT expression vector (Invitrogen) using a variety of restriction sites. The sequences of the minigenes are available at https://github.com/jbkinney/15_splicing. For stable integration, 0.4 μg of minigene plasmid DNA and 3.6 μg of pOG44 (Invitrogen) were co-transfected into 10^6 HeLa FRT cells using Lipofectamine 2000 (Life Technologies). After 48 hr, transfected cells were selected with 200 μg/ml hygromycin B (Invitrogen). For transient transfection, 1 μg of minigene plasmid was transfected into HeLa cells. Cells were collected after 48 hr and RNA analyzed by radioactive RT-PCR.

RT-PCR and splicing analysis

RNA was isolated from minigene-expressing HeLa cells using Trizol (Life Technologies). cDNA was made using Improm-II Reverse Transcription System (Promega), following the manufacturer's instructions. For splicing analysis, a minigene-specific forward primer (FRT F) was used in conjunction with the appropriate reverse primer (*BRCA2* 18R; *BCA2* 7R; *SMN1* R; *IKBKAP* R) in the presence of [^{32}P]-dCTP to amplify the splicing isoforms using Phusion High-Fidelity DNA Polymerase (New England Biolabs), following the manufacturer's instructions. The reaction was initially denatured at 98°C for 2 min, then denatured at 98°C for 15 s, annealed at 58°C for 30 s, and extended at 72°C for 1 min for 26 cycles, with a final extension at 72°C for 10 min. The PCR products were resolved on a 5.5% non-denaturing polyacrylamide gel and were detected with a Typhoon FLA7000 phosphorimager. Quantification of the isoforms was done using ImageJ (NIH).

Construction and sequencing of libraries

A single-stranded DNA fragment containing a randomized 9-nt 5′ss sequence at the end of the middle exon (*BRCA2* Bsu36I ss top; *SMN1* BseRI ss top; *IKBKAP* BseRI ss top) and a separate fragment containing a randomized 20-nt barcode sequence at the end of the last exon (*BRCA2* NotI bc bot; *SMN1* Not bc bot; *IKBKAP* XhoI bc bot) were synthesized. Equal molar ratios of the two strands were annealed at the 20-nt complementary region using 1x annealing buffer (10 mM Tris pH 8.0, 50 mM NaCl, 1 mM EDTA) with Phusion High-Fidelity DNA Polymerase (New England Biolabs), heated to 95°C for 5 min, and extended at 72°C for 30 min to generate the ss-barcode double-stranded fragment. Using the respective restriction enzymes, the ss-barcode fragment was ligated to the pcDNA5 vector comprising the first two exons. The ligated DNA was purified by drop dialysis using a 0.025 μm membrane filter (Millipore) for at least 2 hr, and electroporated into MegaX DH10B T1 Electrocomp Cells (ThermoFisher) using a 0.1-cm cuvette at 2.0 kV, 200 Ω, 25 μF in a BioRad Gene Pulser. The same restriction sites were used to extract the ss-barcode fragment for Illumina sequencing, after ligating the fragment to Illumina-compatible ends generated by annealing PE1 top with PE1 bot, and PE2 top with PE2 bot, respectively (by heating to 95°C for 5 min and then gradually cooling to room temperature over the course of 1 hr). The ligation product was separated on a 2% agarose gel and purified using the QIAquick Gel Extraction Kit (QIAGEN). The N_{10} tract within the primer sequences represents a 7-10 nucleotide library identifier sequence. This library was sequenced using an Illumina HiSeq 2500 PE100, and acted as the "key" to determine the identities of the 5′ss associated with the barcodes.

The remaining intronic and exonic sequences were amplified by the respective insert F and insert R primers, and inserted seamlessly into the library using aaRI, a type IIS restriction endonuclease. The library (6 μg) was then transfected into 5x10^6 HeLa cells using Lipofectamine 2000. Transfected cells were collected after 48 hr. RNA was isolated using Trizol, and analyzed by RT-PCR using Improm-II Reverse Transcription System and Phusion High-Fidelity DNA Polymerase. The exon inclusion product was amplified by PCR, first using a forward primer in the middle exon (BRCA2 17F; SMN17F; IKBKAP 20F) and a minigene-specific reverse primer immediately after the barcode sequence (BC R). Then, the barcode was isolated using the respective forward primer just before the barcode sequence (*BRCA2* 18F; *SMN1* 8F; *IKBKAP* 21F) and the same reverse primer (barcode R). The same pair of primers was used to amplify the total barcode sequences from the transfected library. The barcodes in the exon inclusion product and the total barcodes were amplified first with the respective barcode-LID F and barcode-LID R to add a library-specific identifier represented by the N_{10} within the primer sequences. A second round of amplification using PE1_v4 and PE2_v4 added Illumina-compatible ends for sequencing. Using the previously sequenced "key," the barcode was used to identify the 5′ss sequences that resulted in exon inclusion or skipping. For each gene, two or three independently derived libraries were made, and each library was transfected in triplicate into HeLa cells to ensure the reproducibility of the results. The inclusion ratio of each 5′ss sequence was normalized to that of the consensus 5′ss sequence (CAG/GUAAGU) for each gene context. Two low-quality datasets (*SMN1* library 1, replicate 1 and *SMN1* library 3, replicate 3) were removed from further analysis (Figure S2A).

To ensure that the proper 5′ss was used, a junction analysis was performed using the respective forward (*BRCA2* 17F; *SMN1* 7F; *IKBKAP* 20F) and reverse primers (*BRCA2* 18R; *SMN1* 8R; *IKBKAP* 21R) that flank the exon-exon junction. The fragment was amplified first with the respective JUNCT F and JUNCT R primers to add a library-specific identifier, represented by the N_{10} within the primer sequences. A second round of amplification using PE1_v4 and PE2_v4 added Illumina-compatible ends for sequencing.

In silico analysis of shifted-register 5′ss

We identified shifted-register 5′ss using the following query sequences: NNHGTYRAGT, NYGGTYRAGT, NYAGTRRAGT, NYAGTYYAGT, NYAGTYRBGT, NYAGTYRAHT, and NYAGTYRAGV, where N = A, G, C or T; Y = C or T; R = A or G; H = A, T or C; B = G, C or T; V = G, A or C.

Primers

Primers are listed in Table S3. All primers were purchased from Sigma-Aldrich.

QUANTIFICATION AND STATISTICAL ANALYSIS

Read parsing

FASTQ files were parsed as follows. First, reads were organized according to sample, based on the 7-10 nt sample barcode, which was then removed from the read. Features of interest, including 5′ss sequences, barcodes, and junction sequences, were then

parsed from these reads using regular expressions matching anchor sequences to the side of each feature of interest. These features were then tallied and stored in tab-delimited text files for further processing.

5′ss-barcode association

As described above, 5′ss-barcode fragments were cut from their host plasmids, ligated to Illumina adaptors, and submitted for sequencing. This direct ligation protocol avoids PCR-mediated recombination between 5′ss and barcodes, which we found to be a major problem in preliminary experiments. Upon parsing the resulting sequence data, we found that, by and large, each 20-nt barcode was associated with one unique 5′ss sequence. Specifically, for each barcode, we called an associated 5′ss if at least 2 reads linked that 5′ss to the barcode in question, and if this number of reads was at least 4 times as large as the total number of reads linking the same barcode to other 5′ss sequences.

PSI quantification

As described above, barcodes were amplified from either total RNA or inclusion RNA using an RT-PCR reaction that added Illumina adaptors. After sequencing, each barcode was computationally associated with its corresponding 5′ss, and the total number of reads for each 5′ss was tallied. We denote these quantities for a given 5′ss as n_{tot} and n_{inc}, respectively. A "relative splicing ratio" that quantifies the relative amount of splicing, independent of the sequencing depth of each sample, was then computed as $r = (n_{inc}/n_{tot}) / (N_{inc}/N_{tot})$, where N_{inc} and N_{tot} denote the total number of reads in the inclusion RNA and total RNA samples, respectively. From this ratio, the "percent spliced in" was computed as $PSI = 100*r/r_{con}$, where r_{con} is the relative splicing ratio of the consensus 5′ss CAG/GUAAGU. The PSI values reported for each library are the median PSI values across replicates for that library, whereas the PSI values reported for each minigene context are the median PSI values across all replicates in all libraries for that context.

Junction quantification

RT-PCR was used to amplify exon junctions and add Illumina adaptors. Junction reads were parsed by regular expression matching to anchor sequences upstream and downstream of the variable 5′ss and splice junction with ~30 nt in between. The number of junctions with each observed length was then tallied, and junctions of exactly the expected length were deemed "correct." We noted that a number of junction sequences were missing positions −2 and −1 of the 5′ss, which is likely due to the occurrence of 'GU' at these positions. We therefore excluded 5′ss with 'GU' at positions −2 and −1 from our analysis.

Pairwise dependency

We now describe the regression procedure used to infer pairwise dependencies within the 5′ss. In what follows, indices i and j are used to denote the seven variable positions within each GU 5′ss, i.e., {−2, −1, 3, 4, 5, 6, 7}. Indices b and c are used to denote the four possible RNA bases, {A, C, G, U}. Each sequence s is represented using a 7 × 4 matrix with binary elements s_{ib}, where $s_{ib} = 1$ if base b occurs at position i and $s_{ib} = 0$ otherwise. To infer pairwise dependencies, we fit two different models: a "matrix" model, which accounts for the independent effects on PSI of each possible base at each position, and a "matrix + pairwise" model, which additionally accounts for effects on PSI missed by the matrix model. Mathematically, the matrix model is given by

$$f(s) = \sum_i \sum_b A_{ib} s_{ib} \quad \text{(Equation 1)}$$

where A_{ib} quantifies the contribution of base b at position i. The matrix + pairwise model extends this matrix model and is defined by

$$g(s) = f(s) + \sum_i \sum_{j<i} \sum_{b,c} B_{ijbc} s_{ib} s_{jc} \quad \text{(Equation 2)}$$

where B_{ijbc} denotes the cooperative contribution of having base b at position i together with base c appearing at position j. Matrix model parameters were fit to PSI measurements (only those $\geq$ 20%) using ridge regression, with a regularization parameter chosen using generalized cross-validation (Hastie et al., 2011). The additional pairwise model parameters, B_{ijbc}, were then fit to the residuals using the same ridge-regression procedure. These resulting values for B_{ijbc} are plotted in Figure 4.

DATA AND SOFTWARE AVAILABILITY

Illumina sequencing data has been deposited on the NCBI Sequence Read Archive under accession number SRA:SRP135892 (BioProject:PRJNA420342). Computational analyses were performed using custom Python scripts, which are available at https://github.com/jbkinney/15_splicing. Original gel images have been deposited at Mendeley Data, https://doi.org/10.17632/z25p7f4zvt.1.

Molecular Cell, Volume *71*

Supplemental Information

Quantitative Activity Profile and Context Dependence of All Human 5′ Splice Sites

Mandy S. Wong, Justin B. Kinney, and Adrian R. Krainer

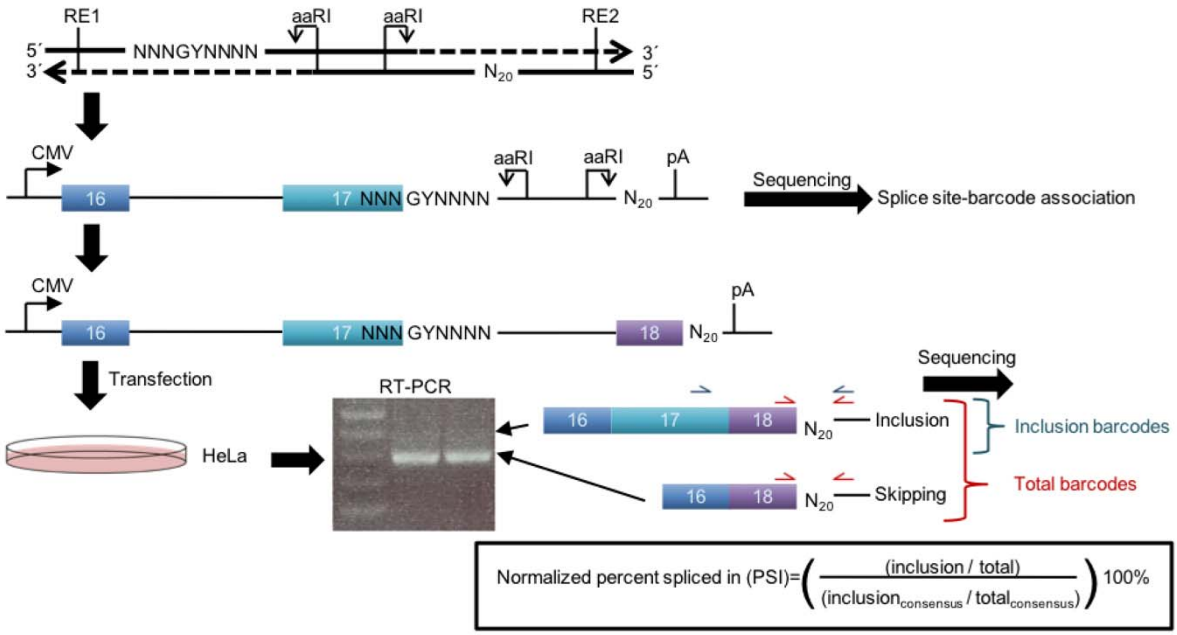


Figure S1. Detailed diagram of high-throughput method used to assess all 5′ss sequences. Related to Figure 2A.
RE1 and RE2 represent restriction enzyme sites used to insert the minigenes into the pcDNA5 expression vector, which has a cytomegalovirus (CMV) promoter and a bGH polyadenylation site (pA).

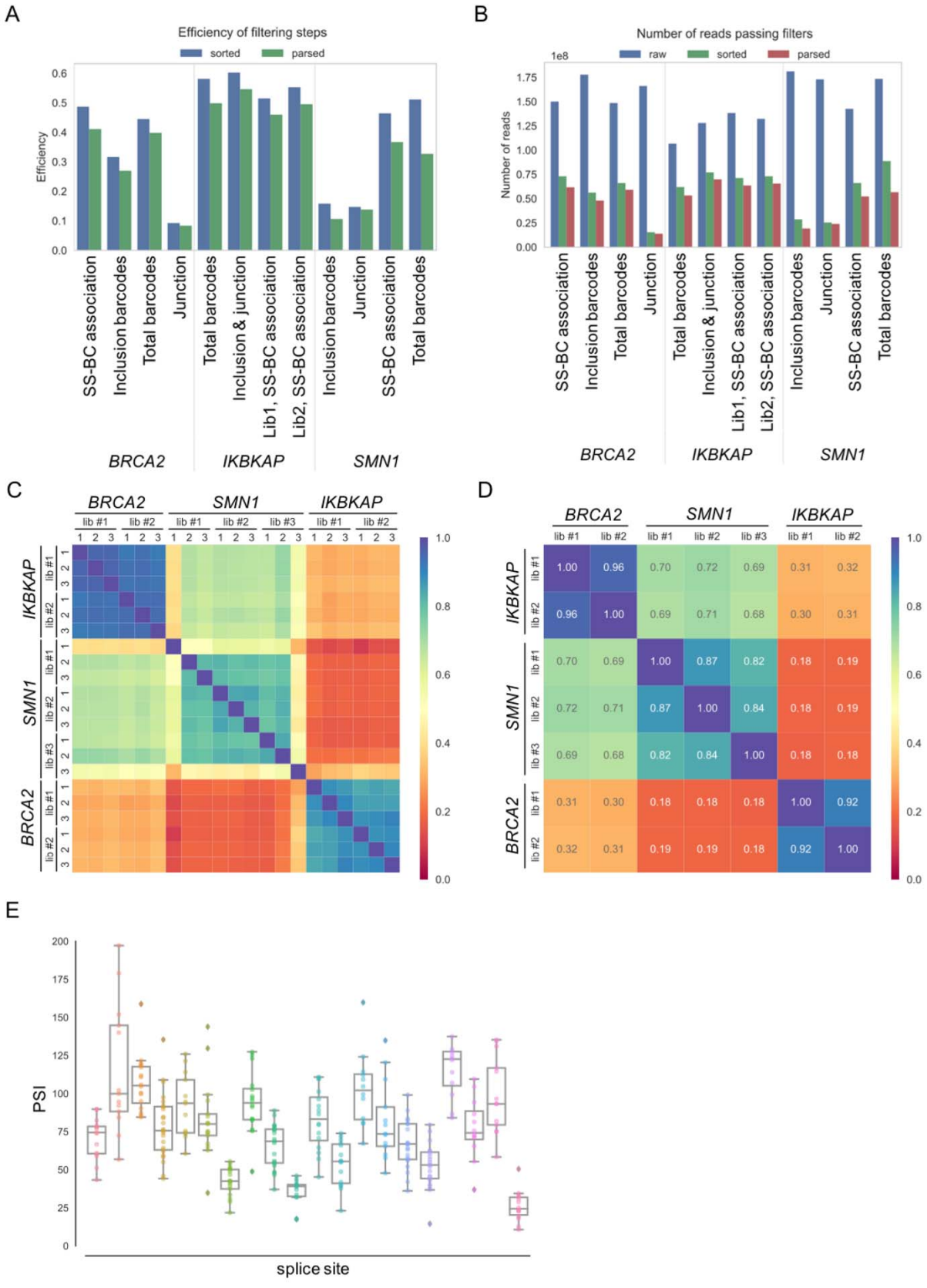

A
Efficiency of filtering steps
sorted
parsed
Efficiency
SS-BC association
Inclusion barcodes
Total barcodes
Junction
Total barcodes
Inclusion & junction
Lib1, SS-BC association
Lib2, SS-BC association
Inclusion barcodes
Junction
SS-BC association
Total barcodes
BRCA2
IKBKAP
SMN1
B
Number of reads passing filters
1e8
raw
sorted
parsed
Number of reads
C
BRCA2
SMN1
IKBKAP
lib #1
lib #2
lib #3
D
1.00 0.96 0.70 0.72 0.69 0.31 0.32
0.96 1.00 0.69 0.71 0.68 0.30 0.31
0.70 0.69 1.00 0.87 0.82 0.18 0.19
0.72 0.71 0.87 1.00 0.84 0.18 0.19
0.69 0.68 0.82 0.84 1.00 0.18 0.18
0.31 0.30 0.18 0.18 0.18 1.00 0.92
0.32 0.31 0.19 0.19 0.18 0.92 1.00
E
PSI
splice site

Figure S2. Parsing of raw high-throughput results and association matrix of the coefficient of determination comparing libraries. Related to Table and Figure 2D.

(A) The efficiency of filtering of raw sequencing data through the bioinformatics pipeline.

(B) Number of reads for analysis after filtering.

(C) Two low-quality datasets (*SMN1* library 1, replicate 1, and *SMN1* library 3, replicate 3) were removed from subsequent analyses.

(D) Coefficient of determination values show that the independently derived libraries highly correlate with each other within a context.

(E) Box plot of the PSI of individual barcodes for 20 randomly selected 5′ss. For this analysis, each barcode is required to have at least 10 counts in the total RNA sample in order to accurately calculate a PSI. The median PSI of the 5′ss is required to be ≥20. For each 5′ss, a minimum of 10 associated barcodes is necessary for the 5′ss to be included in the analysis. The central rectangle spans the first to the third quartile, with the median line segment. The vertical line presents the maximum and minimum.

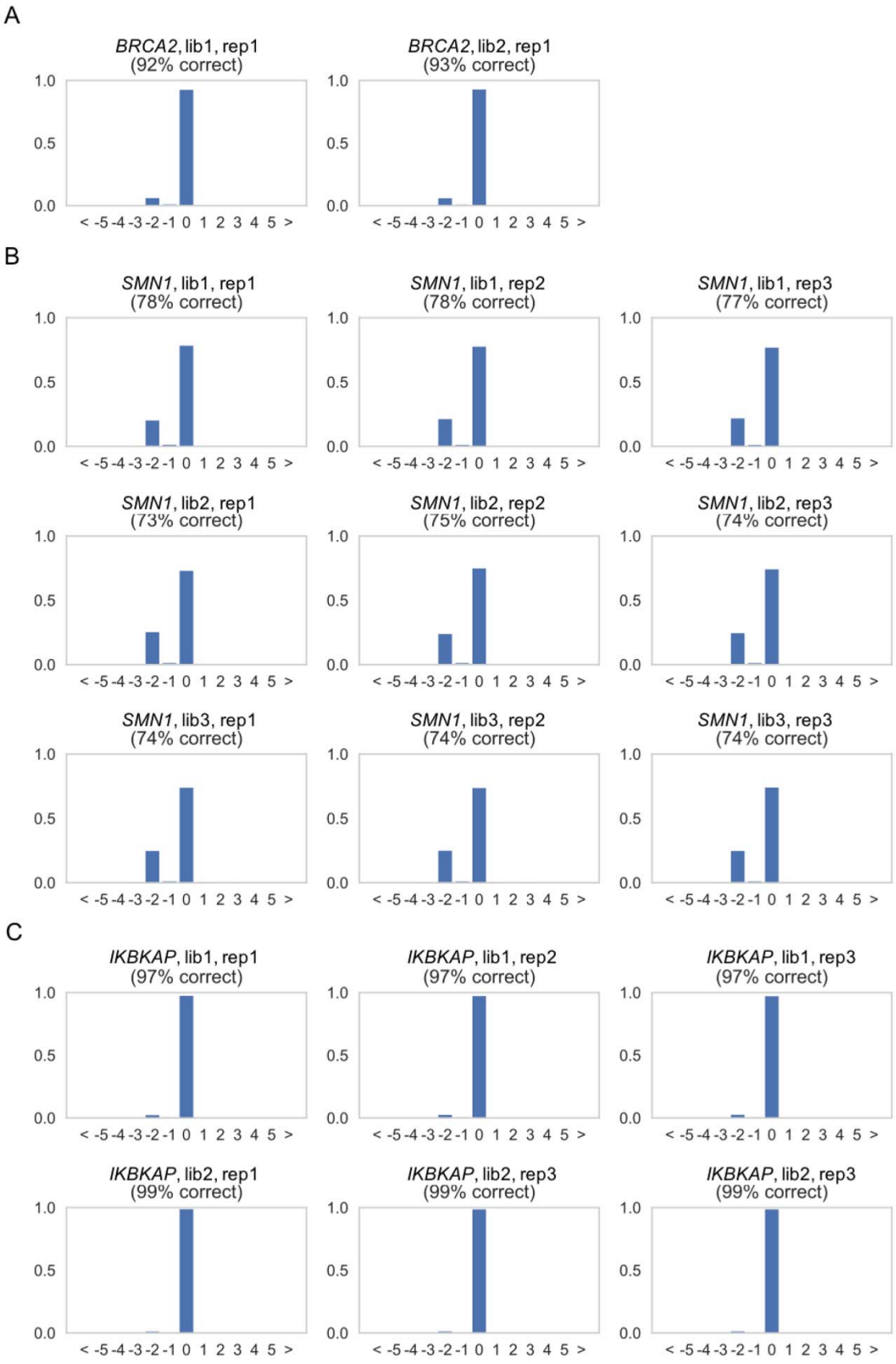

A
BRCA2, lib1, rep1
(92% correct)
BRCA2, lib2, rep1
(93% correct)
B
SMN1, lib1, rep1
(78% correct)
SMN1, lib1, rep2
(78% correct)
SMN1, lib1, rep3
(77% correct)
SMN1, lib2, rep1
(73% correct)
SMN1, lib2, rep2
(75% correct)
SMN1, lib2, rep3
(74% correct)
SMN1, lib3, rep1
(74% correct)
SMN1, lib3, rep2
(74% correct)
SMN1, lib3, rep3
(74% correct)
C
IKBKAP, lib1, rep1
(97% correct)
IKBKAP, lib1, rep2
(97% correct)
IKBKAP, lib1, rep3
(97% correct)
IKBKAP, lib2, rep1
(99% correct)
IKBKAP, lib2, rep3
(99% correct)
IKBKAP, lib2, rep3
(99% correct)
1.0
0.5
0.0
< -5 -4 -3 -2 -1 0 1 2 3 4 5 >

Figure S3. Exon-exon junction analysis of libraries. Related to Figure 2 and “high-throughput analysis of the activity of all 5′ss sequences”.

(A) Sequencing results for the exon-exon junction reveal that a secondary GU at the -2 and -1 positions is preferentially used when the GU or GC at the +1 and +2 positions escapes recognition. 5′ss sequences with the secondary -2G-1U were removed from further analysis.

(B) Same as A, but in *SMN1*.

(C) Same as A, but in *IKBKAP*.

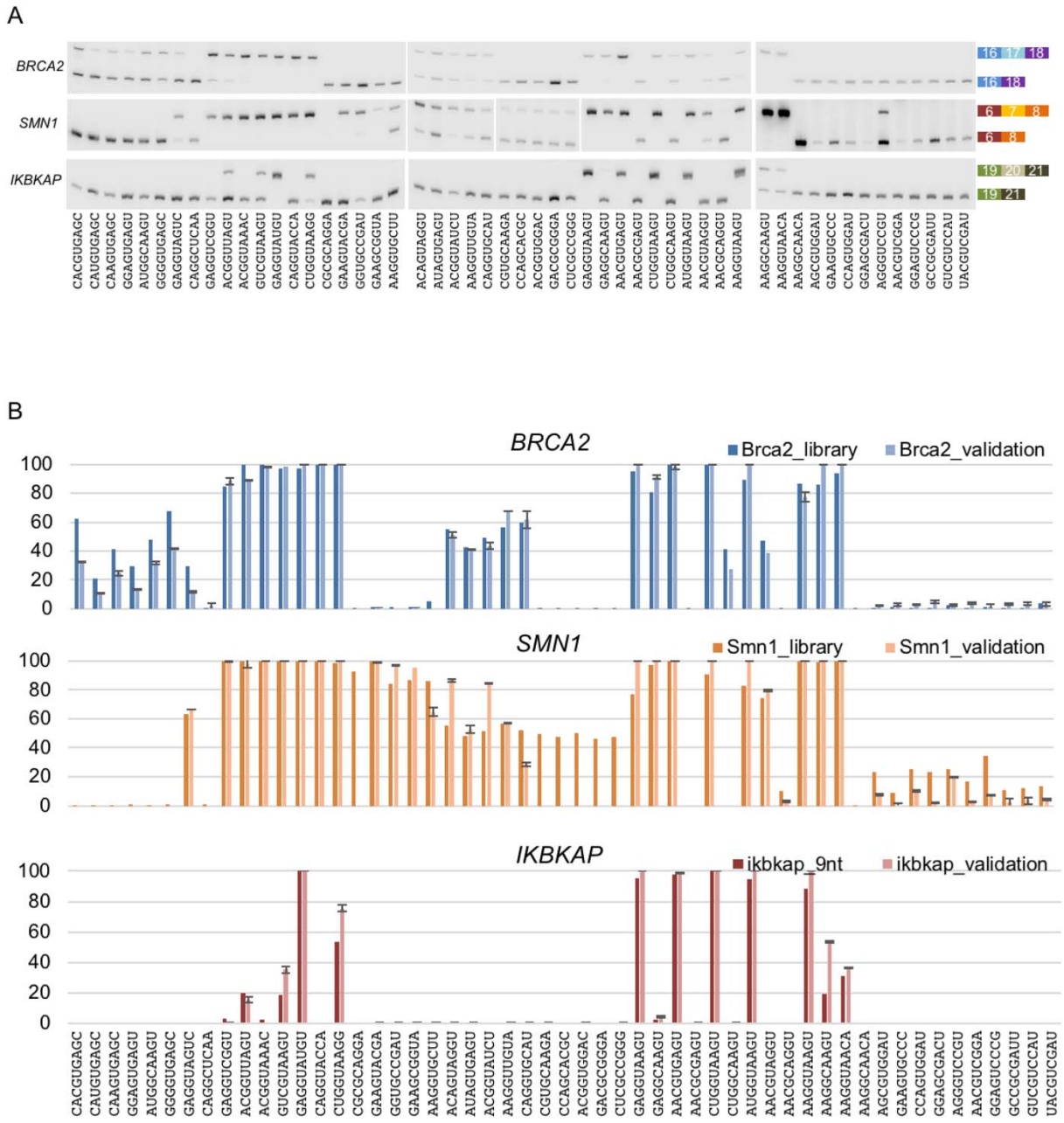


Figure S4. Manual validations of randomly-selected 5′ss in the three contexts. Related to Figure 2F.

(A) Gel showing the splicing results of the same 53 randomly-selected 5′ss in the *BRCA2*, *SMN1*, and *IKBKAP* minigenes. Gel images are representatives of triplicates. Vertical gaps indicate samples run on different gels.

(B) Graphs comparing the percent spliced in (PSI) of 5′ss derived from the library results and manual validations. Standard deviation is represented by error bars.

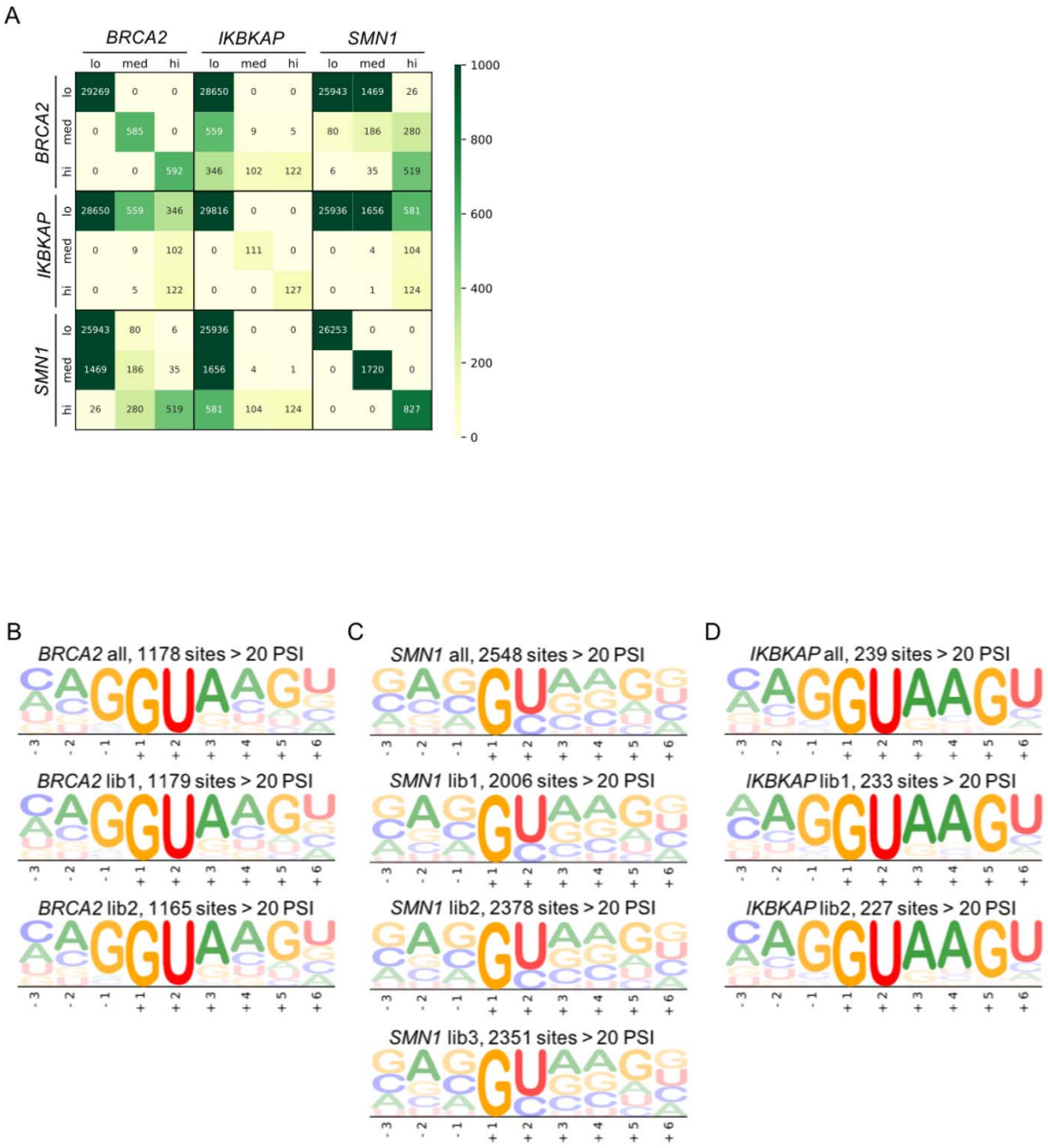


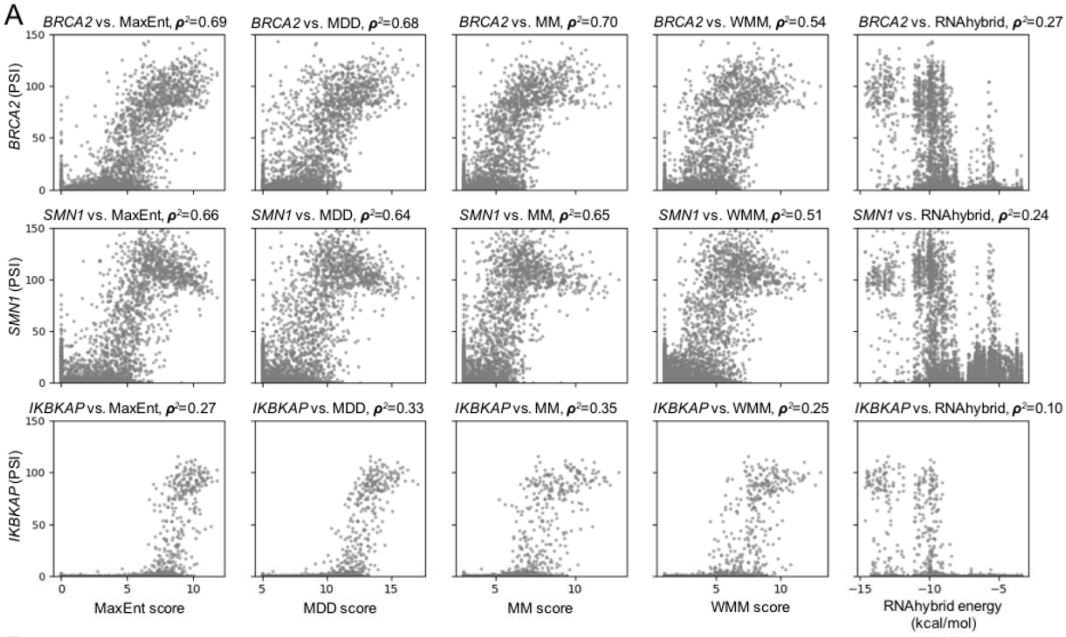
Figure S5. Sequence logo of 5′ss in each context. Related to Figure 3B and 3C.

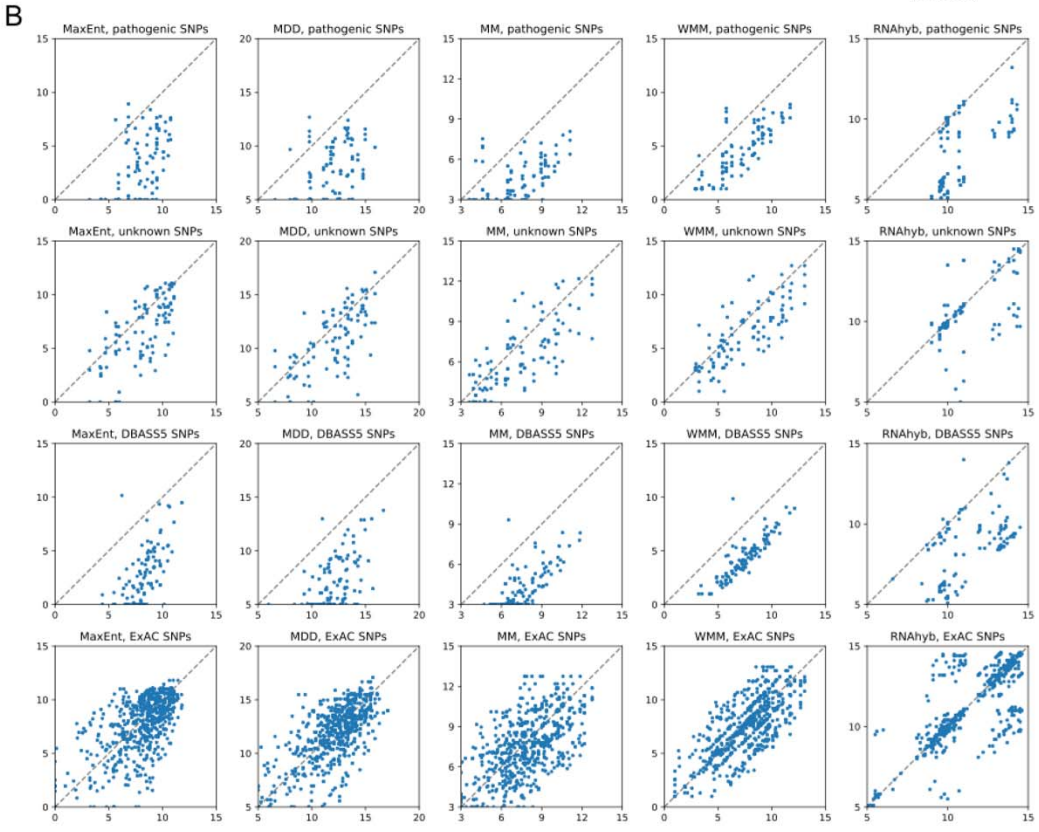
(A) Heat map illustrating the overlaps between 5′ss with lo (PSI < 20), med (20≤PSI<80) or hi (80≤PSI) activity levels among the three gene contexts.

(B) Sequence logo for all 5′ss compiled or separated by each independently-derived library in *BRCA2*.

(C) Same as A, but in *SMN1*.

(D) Same as A, but in *IKBKAP*.


A
BRCA2 vs. MaxEnt, ρ²=0.69
BRCA2 vs. MDD, ρ²=0.68
BRCA2 vs. MM, ρ²=0.70
BRCA2 vs. WMM, ρ²=0.54
BRCA2 vs. RNAhybrid, ρ²=0.27
BRCA2 (PSI)
SMN1 vs. MaxEnt, ρ²=0.66
SMN1 vs. MDD, ρ²=0.64
SMN1 vs. MM, ρ²=0.65
SMN1 vs. WMM, ρ²=0.51
SMN1 vs. RNAhybrid, ρ²=0.24
SMN1 (PSI)
IKBKAP vs. MaxEnt, ρ²=0.27
IKBKAP vs. MDD, ρ²=0.33
IKBKAP vs. MM, ρ²=0.35
IKBKAP vs. WMM, ρ²=0.25
IKBKAP vs. RNAhybrid, ρ²=0.10
IKBKAP (PSI)
MaxEnt score
MDD score
MM score
WMM score
RNAhybrid energy (kcal/mol)


B
MaxEnt, pathogenic SNPs
MDD, pathogenic SNPs
MM, pathogenic SNPs
WMM, pathogenic SNPs
RNAhyb, pathogenic SNPs
MaxEnt, unknown SNPs
MDD, unknown SNPs
MM, unknown SNPs
WMM, unknown SNPs
RNAhyb, unknown SNPs
MaxEnt, DBASS5 SNPs
MDD, DBASS5 SNPs
MM, DBASS5 SNPs
WMM, DBASS5 SNPs
RNAhyb, DBASS5 SNPs
MaxEnt, ExAC SNPs
MDD, ExAC SNPs
MM, ExAC SNPs
WMM, ExAC SNPs
RNAhyb, ExAC SNPs

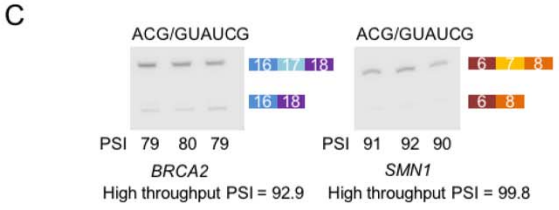

C
ACG/GUAUCG
16 17 18
16 18
PSI 79 80 79
BRCA2
High throughput PSI = 92.9
ACG/GUAUCG
6 7 8
6 8
PSI 91 92 90
SMN1
High throughput PSI = 99.8

Figure S6. Comparison between library results and several conventional prediction algorithms. Related to Figure 3D, 3E, and 6.

(A) Scatter plots comparing the predicted values for each 5′ss using maximum entropy (MaxENT; Yeo et al., 2004), maximum dependence decomposition (MDD; Burge et al., 1998), first-order Markov model (MM; Krogh et al., 1994), weight matrix model (WMM), and RNAhybrid (RNAhyb; Kruger et al., 2006) versus the experimentally derived library results.

(B) Scatter plots, corresponding to the computational scoring matrices used above, comparing the predicted PSI of the WT 5′ss sequences to mutant 5′ss sequences known to be pathogenic in *BRCA1* and *BRCA2* (pathogenic SNPs), with unclassified or uncertain significance in *BRCA1* and *BRCA2* (unknown SNPs), 5′ss mutations across a broad range of genes and diseases (DBASS5), and for 5′ss SNPs with >10% frequency found in the human population (ExAC SNPs).

(C) RT-PCR validation of the usage of the 5′ss with the sequence ACG/GUAUCG, which showed high inclusion ratio in *BRCA2* and *SMN1* library results but does not occur naturally as a 5′ss in the human transcriptome. Gel image is representative of triplicates. Percent spliced in (PSI) is indicated below each lane.

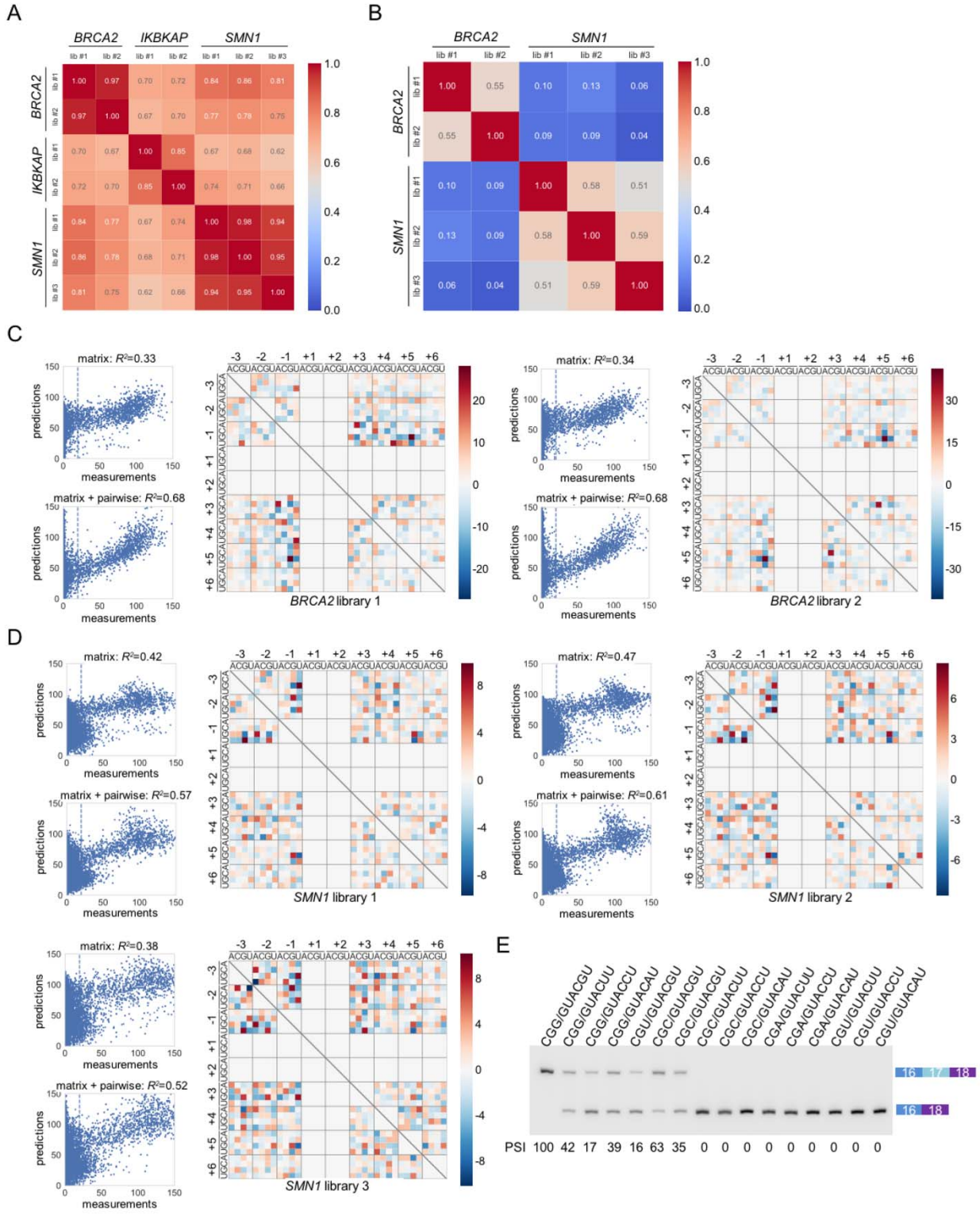

A
BRCA2
IKBKAP
SMN1
lib #1
lib #2
lib #3
1.00 0.97 0.70 0.72 0.84 0.86 0.81
0.97 1.00 0.67 0.70 0.77 0.78 0.75
0.70 0.67 1.00 0.85 0.67 0.68 0.62
0.72 0.70 0.85 1.00 0.74 0.71 0.66
0.84 0.77 0.67 0.74 1.00 0.98 0.94
0.86 0.78 0.68 0.71 0.98 1.00 0.95
0.81 0.75 0.62 0.66 0.94 0.95 1.00
1.0 0.8 0.6 0.4 0.2 0.0
B
BRCA2
SMN1
1.00 0.55 0.10 0.13 0.06
0.55 1.00 0.09 0.09 0.04
0.10 0.09 1.00 0.58 0.51
0.13 0.09 0.58 1.00 0.59
0.06 0.04 0.51 0.59 1.00
C
matrix: R²=0.33
matrix + pairwise: R²=0.68
predictions
measurements
BRCA2 library 1
matrix: R²=0.34
matrix + pairwise: R²=0.68
BRCA2 library 2
D
matrix: R²=0.42
matrix + pairwise: R²=0.57
SMN1 library 1
matrix: R²=0.47
matrix + pairwise: R²=0.61
SMN1 library 2
matrix: R²=0.38
matrix + pairwise: R²=0.52
SMN1 library 3
E
CGG/GUACGU
CGG/GUACUU
CGG/GUACCU
CGG/GUACAU
CGU/GUACGU
CGC/GUACGU
CGC/GUACUU
CGC/GUACCU
CGC/GUACAU
CGA/GUACUU
CGA/GUACCU
CGA/GUACAU
CGU/GUACUU
CGU/GUACCU
CGU/GUACAU
16 17 18
16 18
PSI 100 42 17 39 16 63 35 0 0 0 0 0 0 0 0 0

Figure S7. Accounting for pairwise associations between nucleotides improves the predictability of the results. Related to Figure 4.

(A) Association matrix showing correlations of parameters in the linear matrix model.

(B) Same as A, but in pairwise associations. The number of active 5′ss was insufficient to analyze pairwise associations for *IKBKAP*.

(C) Scatter plot showing the predicted PSI versus the measured PSI when considering a linear matrix model only (top) and a matrix model together with pairwise interaction (bottom). The heat map shows pairwise associations between nucleotides at different positions in *BRCA2*, separated by libraries. Red indicates a positive interaction, and blue indicates a negative interaction.

(D) Same as C, but for *SMN1*.

(E) Validation of a comprehensive series of mutations at -1 and +5 positions of the 5'ss CGG/GUACGU, in the *BRCA2* context. Gel image is representative of triplicates. Percent spliced in (PSI) is indicated below each lane.

Table S1 - 5′ss sequences that have activity in library results, but do not occur naturally in the human transcriptome. Related to Figure 3E.

BRCA2	***SMN1***
ACGGUAUCG	ACGGUAUCG
CGCGUACGU	CGCGUACGU
AAAGCGCUG	AACGUACGG
CCAGUACCG	CACGUACGC
CCCGCCGUG	CCGGUAUAC
GCGGUAAAC	CGCGUACGA
GUGGCAUCG	GAAGCGGUA
	GACGCGGUA
	GACGUACGA
	GGCGCGUAU

Table S2 – Predicted strength of the upsteam and downstream 3′ss. Related to Figure 5.

	Gene (input sequence, 5´→3´)	MAXENT	MM	WMM
Upstream 3′ss	*BRCA2* (TTCTACTTTTATTTGTTCAGGGC)	8.33	9.46	10.78
	SMN1 (TTCCTTTATTTTCCTTACAGGGT)	10.92	13.08	15.51
	IKBKAP (ACTGCTTTAATTTATTTAAGATG)	6.36	6.51	4.57
Downstream 3′ss	BRCA2 (ATTTTTGTTTTCACTTTTAGATA)	11.50	12.16	12.62
	SMN1 (TTCTAATTTCTCATTTGCAGGAA)	10.77	10.86	10.52
	IKBKAP (CTTTCTCTGTCTTCTCACAGACT)	11.96	12.06	13.35

Table S3 – Primer sequences. Related to STAR methods.

Name	Sequence (5′→3′)
FRT F	CTGGCTAACTAGAGAACCCACTGC
BRCA2 18R	GCTGTGTCATCCCTTTCCATTATC
BRCA2 7R	GAGCACAGTAGAACTAAGGGTGG
SMN1 R	TAGTGGTGTCATTTAGTGCTGC
IKBKAP R	GATTGATTCTCAGCTTTCTCATGC
BRCA2 Bsu36I ss top	CATCATCCTAAGGAATTTGCTAATAGATGCCTAAGCCCAGAAAGGG TGCTTCTTCAACTAAAATANNNGNNNNNTTAAAGCAGCAGGTGGAT GCACATGATGACATAAT
BRCA2 NotI bc bot	TACTACCGCCGGCGNNNNNNNNNNNNNNNNNNNNTCTAGAATGCA GGTGATTATGTCATCATGTGCATC
BRCA2 PE1 top	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACG CTCTTCCGATCTNNNNNNNNNNT
BRCA2 PE1 bot	CATGTNNNNNNNNNNTCTAGCCTTCTCGCAGCACATCCCTTTCTCA CATCTAGAGCCACCAGCGGCATAGTAA
BRCA2 PE2 top	CGTNNNNNNNNNNAGATCGGAAGAGCGGTTCAGCAGGAATGCCGA GACCGATCTCGTATGCCGTCTTCTGCTT
BRCA2 PE2 bot	TTCGTCTTCTGCCGTATGCTCTAGCCAGAGCCGTAAGGACGACTTG GCGAGAAGGCTAGANNNNNNNNNNT
BRCA2 insert F	CATCATCACCTGCAGAGTTAAAGCATTACATTACG
BRCA2 insert R	TACTACCACCTGCACACTCTAGAATTACTACTTTAAC
BRCA2 17F	AGATGCCTAAGCCCAGAAAG
BRCA2 18F	GGCTCTCCTGATGCCTGTAC
BRCA2 18R	GCTGTGTCATCCCTTTCCATTATC
BRCA2 BC R	GGCAACTAGAAGGCACAGTCG
BRCA2 BC-LID F	CCCTACACGACGCTCTTCCGATCTNNNNNNNNNNGGCTCTCCTGAT GCCTGTAC
BRCA2 BC-LID R	CATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGGCAACTA GAAGGCACAGTCG
BRCA2 JNCT F	CCCTACACGACGCTCTTCCGATCTNNNNNNNNNNAGATGCCTAAGC CCAGAAAG
BRCA2 JNCT R	CATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGCTGTGTC ATCCCTTTCCATTATC
SMN1 BseRI ss top	CATCATGAGGAGCTTAAATTAANNNGYNNNNCTGCCAGCATGCAGG TGGATGCACATGATGACATAA
SMN1 NotI bc bot	ATGATGGCGGCCGCNNNNNNNNNNNNNNNNNNNNTCTAGAATGCA GGTGATTATGTCATCATGTGCATC
SMN1 PE1 top	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACG CTCTTCCGATCTNNNNNNNNNNGTGCT
SMN1 PE1 bot	CNNNNNNNNNNAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA GATCTCGGTGGTCGCCGTATCATT

SMN1 PE2 top	GGCCGCNNNNNNNNNNAGATCGGAAGAGCGGTTCAGCAGGAATG CCGAGACCGATCTCGTATGCCGTCTTCTGCTT
SMN1 PE2 bot	AAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAA CCGCTCTTCCGATCTNNNNNNNNNNGC
SMN1 insert F	CATCATCACCTGCTAGGGCCAGCATTATGAAACTGAATC
SMN1 insert R	ATGATGCACCTGCCCTATCTAGAATAACGCTTCACATTCCAGATC
SMN1 7F	GAAGGAAGGTGCTCACATTC
SMN1 8F	GACACCACTAAAGAAACGATCAG
SMN1 8R	CGCTTCACATTCCAGATCTG
SMN1 BC R	GGCAACTAGAAGGCACAGTCG
SMN1 BC-LID F	CCCTACACGACGCTCTTCCGATCTNNNNNNNNNNGACACCACTAAA GAAACGATCAG
SMN1 BC-LID R	CATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGGCAACTA GAAGGCACAGTCG
SMN1 JNCT F	CCCTACACGACGCTCTTCCGATCTNNNNNNNNNNGAAGGAAGGTG CTCACATTC
SMN1 JNCT R	CATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNCGCTTCAC ATTCCAGATCTG
IKBKAP BseRI ss top	CATCATGAGGAGAGTGGTTGGANNNGYNNNNGCCATTGTGCAGGT GGATGCACATGATGACATAAT
IKBKAP XhoI bc bot	ATGATGCTCGAGNNNNNNNNNNNNNNNNNNTCTAGAATGCAGG TGATTATGTCATCATGTGCATC
IKBKAP PE1 top	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACG CTCTTCCGATCTNNNNNNNNNNTTAAT
IKBKAP PE1 bot	TAANNNNNNNNNNAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTG TAGATCTCGGTGGTCGCCGTATCATT
IKBKAP PE2 top	TCGAGNNNNNNNNNNAGATCGGAAGAGCGGTTCAGCAGGAATG CCGAGACCGATCTCGTATGCCGTCTTCTGCTT
IKBKAP PE2 bot	AAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAA CCGCTCTTCCGATCTNNNNNNNNNNC
IKBKAP insert F	CATCGTCACCTGCAAGCGCCATTGTACTGTTTGCGACTAGTTAGC
IKBKAP insert R	ATGATGCACCTGCCATGTCTAGAACTACTTAGGGTTATGATCAT
IKBKAP 20F	GTTGTTCATCATCGAGCCCTGG
IKBKAP 21F	GCATGAGAAAGCTGAGAATC
IKBKAP 21R	GATTCTCAGCTTTCTCATGC
IKBKAP BC R	GGCAACTAGAAGGCACAGTCG
IKBKAP BC-LID F	CCCTACACGACGCTCTTCCGATCTNNNNNNNNNGCATGAGAAAGCT GAGAATC
IKBKAP BC-LID R	CATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGGCAACTA GAAGGCACAGTCG
IKBKAP JNCT F	CCCTACACGACGCTCTTCCGATCTNNNNNNNNNNGTTGTTCATCAT CGAGCCCTGG
IKBKAP JNCT R	CATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGATTCTCA GCTTTCTCATGC
PE1_v4	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACG CTCTTC

PE2_v4	AAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCT
ACTB-F	AGAGCTACGAGCTGCCTGAC
ACTB-R	AGCACTGTGTTGGCGTACAG