Supplemental figures and tables for
Ke et al., Saturation mutagenesis reveals manifold determinants of exon definition

Table of Contents (Tables or Figures listed in order of citation in the text)

(Large Supplemental Tables 2, 6, 7, and 8 are available via separate links to .xlsx files at the Genome Research Web site.)

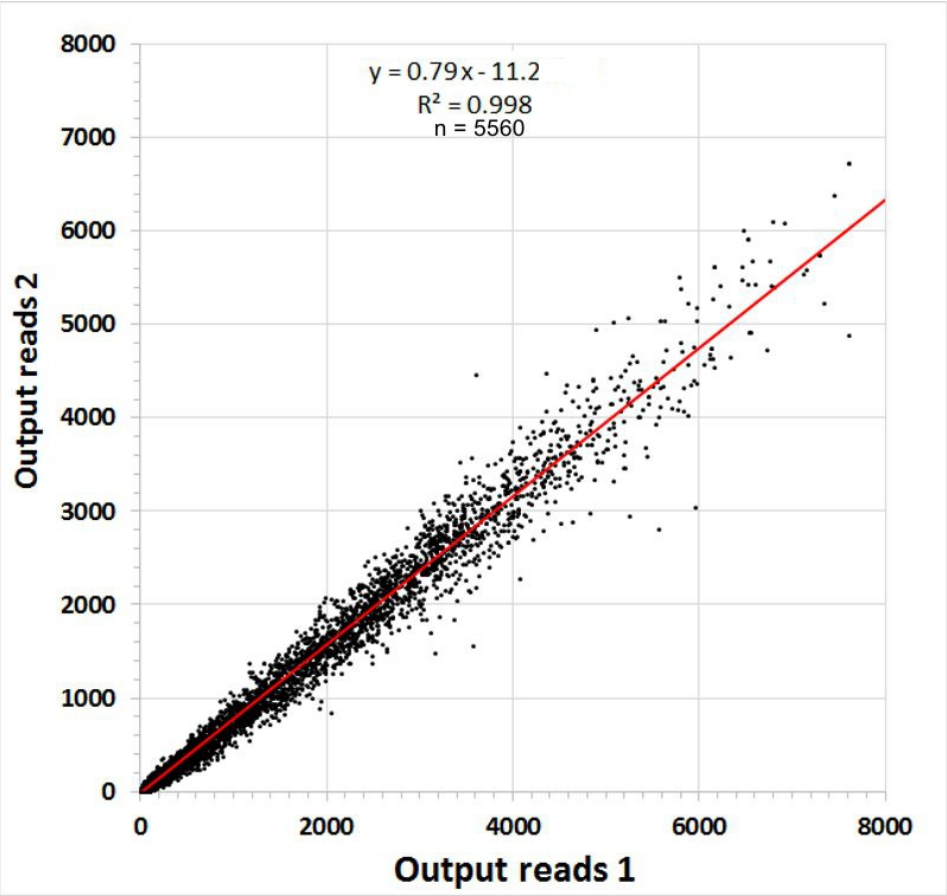


Figure S1. Reproducibility of successfully spliced mRNA levels between biological replicates. Duplicate cultures of human HEK293 cells were transfected with the single library of PCR products comprised of all 10 mutant minigenes described here (Hexmuts A to J). The cDNA region from mRNA molecules bearing the central exon was then prepared from each replicate, amplified and subjected to deep sequencing. The number of reads of each mutant molecule (passing the same accuracy filter) are shown compared on a scatter plot. The number of points is 5560; the reads numbered 7,814,806 in experiment 1 and 6,141,948 in experiments 2.

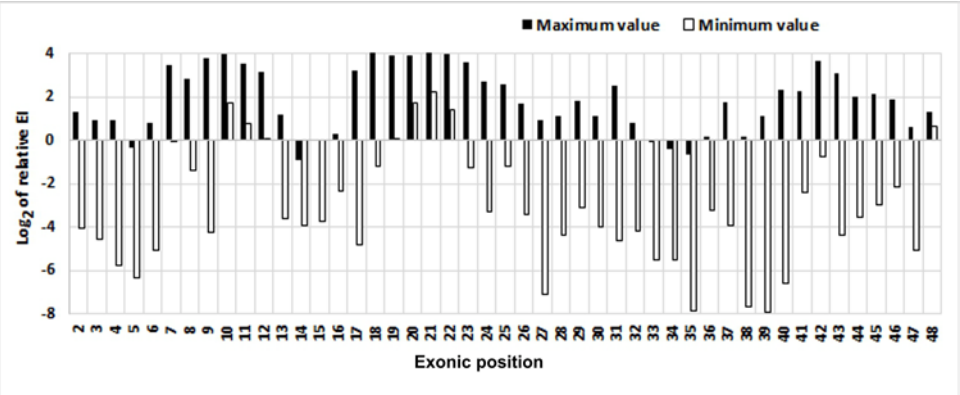


Figure S2. Distribution of extreme phenotypes (highest and lowest splicing) across the HMA exon. Relative LEI is the $\log_2$ of the relative EI (EI/EI_{WT}). Note there is no particular sensitivity close to the splice sites.

Table_S1. SBS and DBS mutations can affect multiple splicing motifs

HMA	39 A → T					
Start	WT	Mut	WT ESRs	ESR type	Mut ESR	ESR type
39	**A**AGGGC →	**T**AGGGC	-0.146	ESS	-0.635	ESS
38	G**A**AGGG →	G**T**AGGG	*0.002*		-0.733	ESS
37	AG**A**AGG →	AG**T**AGG	*-0.015*		-0.703	ESS
36	CAG**A**AG →	CAG**T**AG	0.390	ESE	-0.530	ESS
35	ACAG**A**A →	ACAG**T**A	*0.291*		*-0.274*	
34	GACAG**A** →	GACAG**T**	0.414	ESE	*0.055*	

ESRseq score changes for HMA mutant A>T at positon 39 help explain its splicing phenotype. This SBS reduces splicing to essentially zero. All 6-mers overlapping the position of the SBS are shown along with their ESRseq scores, where a positive value corresponds to an ESEseq and a negative value corresponds to an ESSseq (Ke et al., 2011). ESRseq scores in italics are not significant at the $p<0.05$ level. The SBS eliminates one ESE (34), converts another ESE to an ESS (36), strengthens a weak ESS over 4-fold (39) and creates 2 additional ESSs (37, 38).

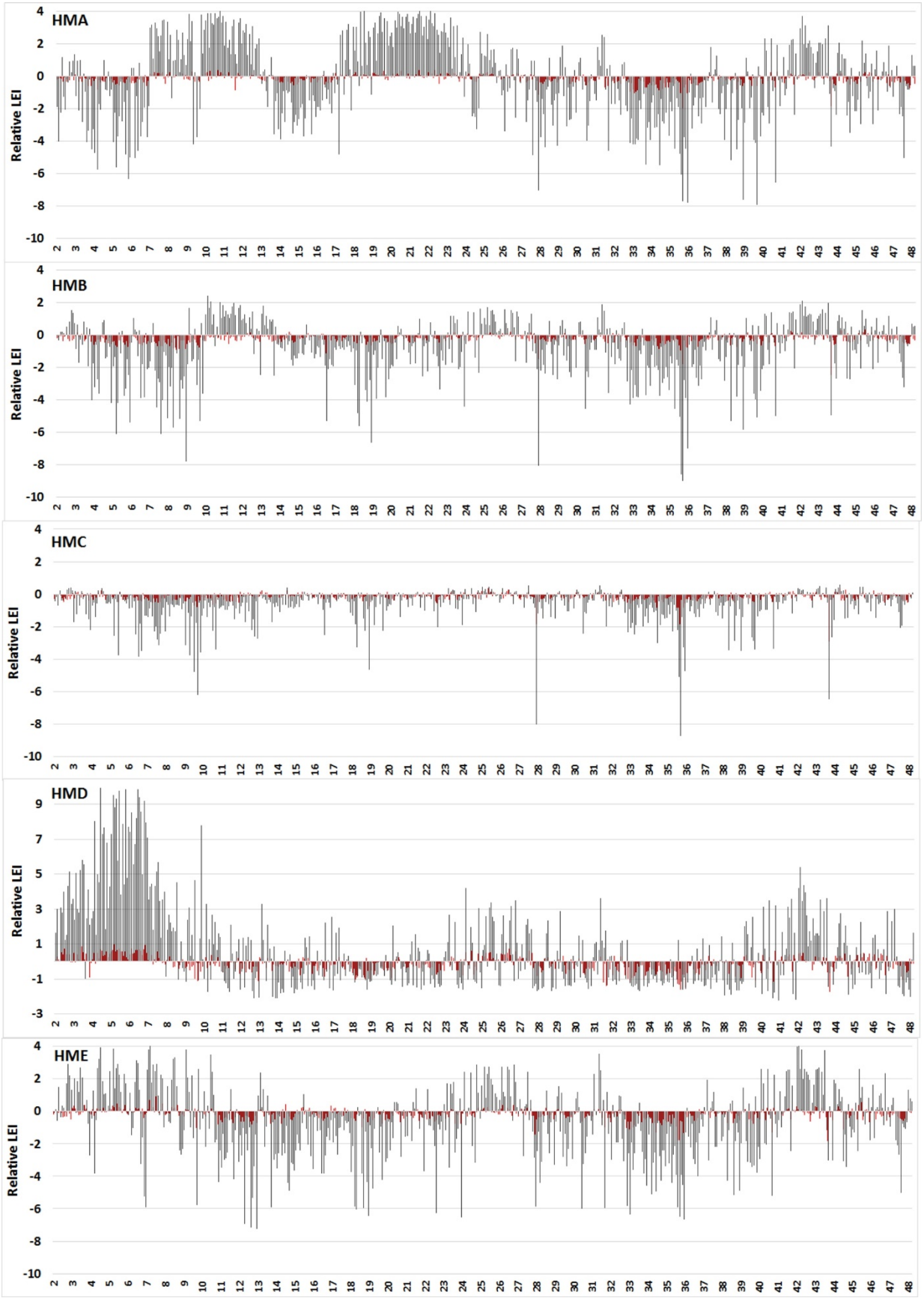


Figure S3. Maps of splicing phenotypes of all 10 Hexmuts. Panels are identified by the Hexmut letter in the upper left corner. Gray columns are the relative LEI values ($\log_2$ of the relative EI values, EI/EIwt) for the standard minigene library containing 2 introns; red columns are from the intronless minigene control library, which yields minor effects compared to the intron-containing counterparts. All scales are the same except for HexmutD, which has an hnRNPA1 binding sequence created at positions 4 to 10 that practically eliminates splicing. The psi values of the relative WT molecules are also indicated; note that the upper limits on fold increases are 1/WTpsi, or $-\log_2$(WTpsi) for the log transformed values used here on the y-axis.

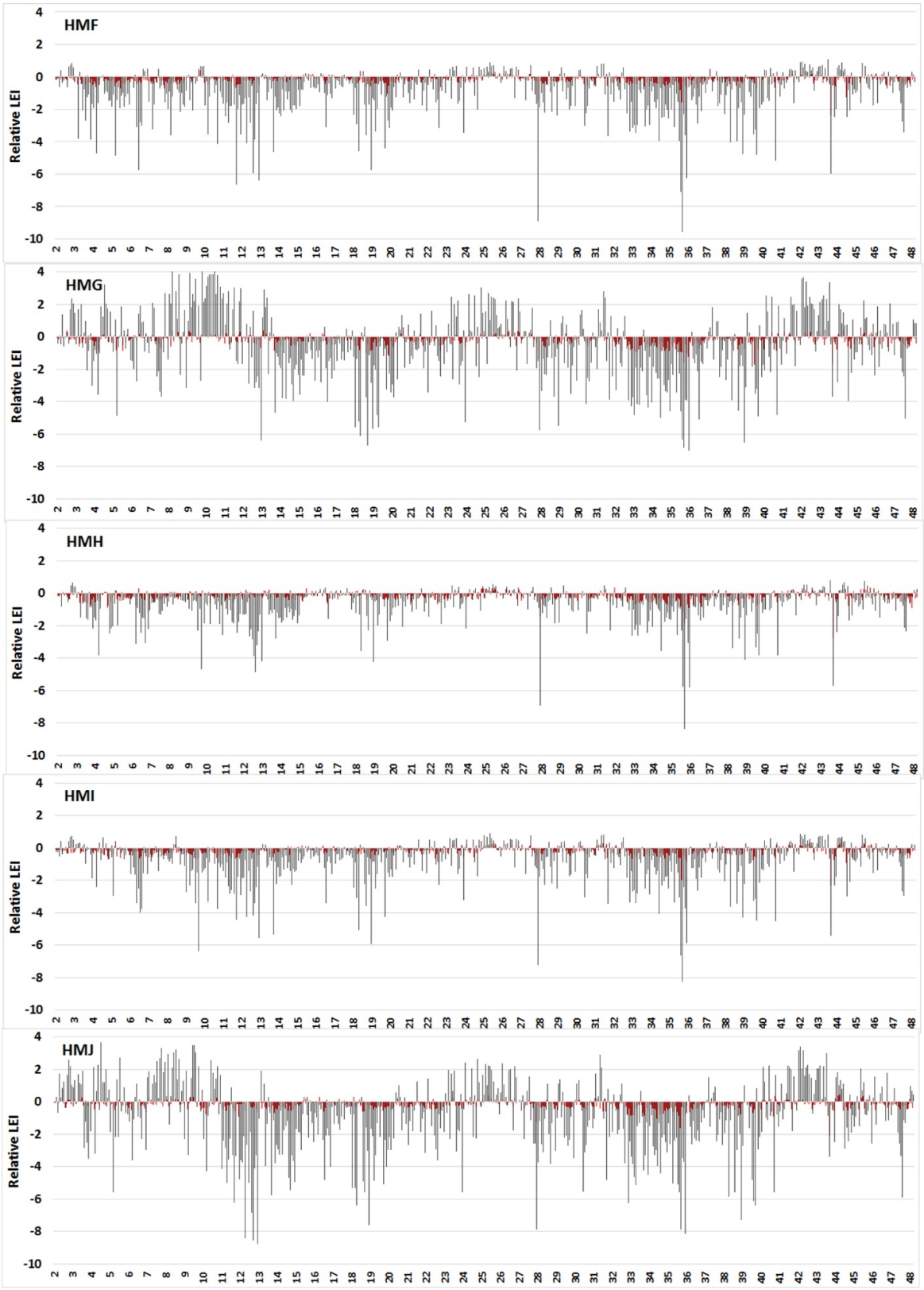


Figure S3 cont.. Legend repeated: Maps of splicing phenotypes of all 10 Hexmuts. Gray columns are the relative LEI values (log_2 of the relative EI values, EI/EIwt) for the standard minigene library containing 2 introns; red columns are from the intronless minigene control library, which yields minor effects compared to the intron-containing counterparts. All scales are the same except for HexmutD, which has an hnRNPA1 binding sequence created at positions 4 to 10 that practically eliminates splicing. The psi values of the relative WT molecules are also indicated; note that the upper limits on fold increases are 1/WTpsi, or $–log_2$(WTpsi).

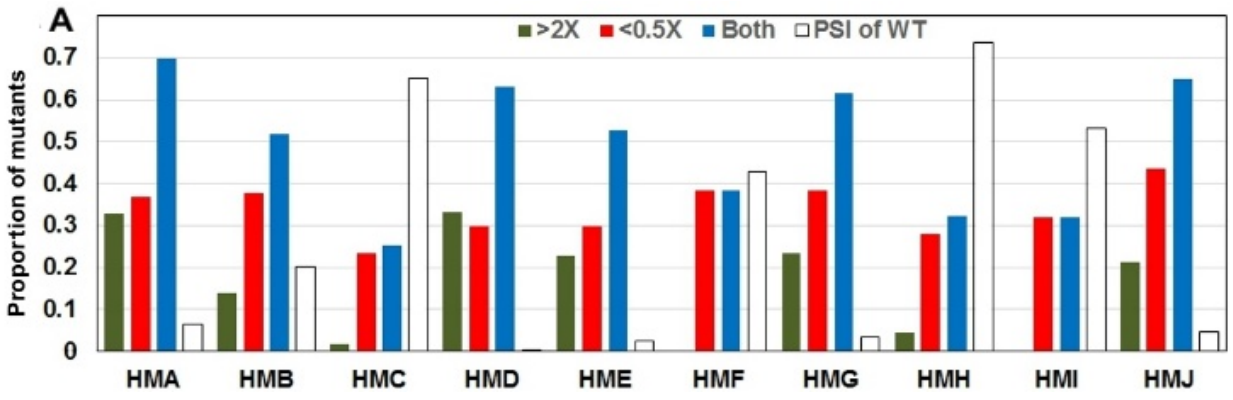


Fig. S4. Summary of mutations that substantially increased and decreased splicing in the 10 Hexmuts. Red and green columns show, respectively, the proportion of mutants that decreased splicing to less than half the relative WT or increased it more than 2-fold or at least to a psi of 0.9. Blue columns show the combined proportion of these changes. White columns indicate the starting psi values of the relative WT molecules, using the same values on the y-axis.

R^2	HMB	HMC	HMD	HME	HMF	HMG	HMH	HMI	HMJ
HMA	0.37	0.33	0.14	0.33	0.36	0.35	0.31	0.38	0.35
HMB		0.76	0.37	0.92	0.92	0.91	0.75	0.92	0.95
HMC			0.25	0.63	0.85	0.65	0.86	0.86	0.69
HMD				0.38	0.34	0.35	0.26	0.32	0.37
HME					0.84	0.89	0.64	0.83	0.92
HMF						0.86	0.85	0.95	0.90
HMG							0.66	0.85	0.92
HMH								0.82	0.70
HMI									0.88

0.1 0.5 0.9

Figure S5. Splicing phenotypes of mutations downstream of the mutated hexamer are highly correlated between Hexmuts. The 6-mer substitution that differentiates Hexmuts spans positions 5 to 10. These correlations are for LEIsc values from positions 16 to 48. Exceptions were those expected: HMA, which has a strong secondary effect and HMD, which includes many estimated very low splicing scores. Without these 2 Hexmuts, the range of R^2 values was 0.63 to 0.95.

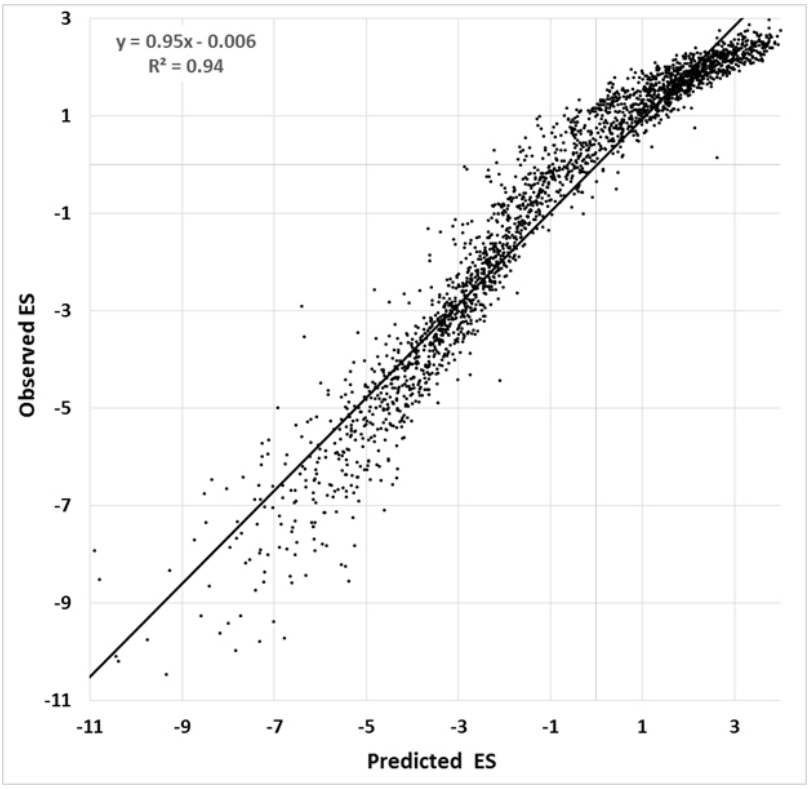


Fig. S6. Lack of epistasis between Hexmut mutations and most SBSs and DBSs. The method of Julien et al. (2016) was adapted to predict the result of combining the effects of two separated mutations, A and B. Here A was a Hexmut 6-base substitutions and B was one of the 555 S/DBSs. AB molecules contain both the Hexmut and the S/DBS. HMA and HMD were omitted from consideration because HMA is sensitive to a secondary structure and D contains many mutants that exhibit no detectable splicing, confounding log transformed quantification. Of the 8 remaining Hexmuts, HMB was used as a reference ("WT") for normalization; i.e., for each mutant, ES = $\log_2(EI_{mutx}/EI_{HMBmutx})$, where ES is the Enrichment Score of Julien et al. On the x-axis is plotted the prediction for linear additivity (AB = A + B); the y-axis shows the actual ES of AB. Wanting to rule out nearby mutations that might create a new RBP binding motif, we considered only mutation combinations that were at least 10 bp apart. The high R^2 value implies that the great majority of combinations acted additively. The number of mutant molecules in this set was 2373, so there may be a significant and potentially interesting number of combinatioins that do exhibit epistasis,

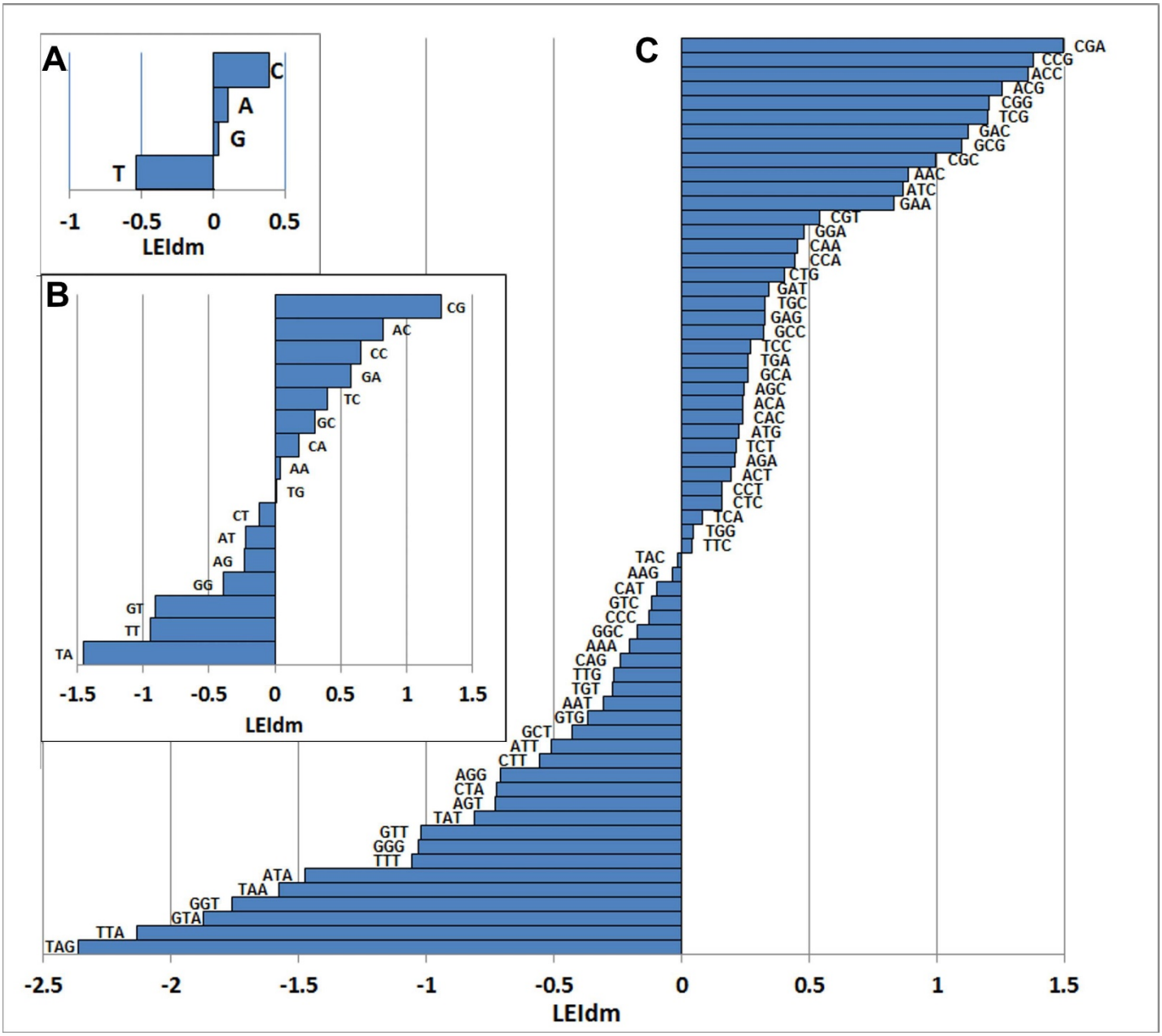


Figure S7. Motif splicing scores for short sequences. The effects of mono-, di- and trinucleotide changes on splicing were scored by their association with increased splicing when created. The changes in LEI (LEIdm) elicited by the indicated nucleotide changes at a given position were averaged for all positions and all 10 Hexmuts. (A) Mononucleotides. (B) Dinucleotides. (C) Trinucleotides.

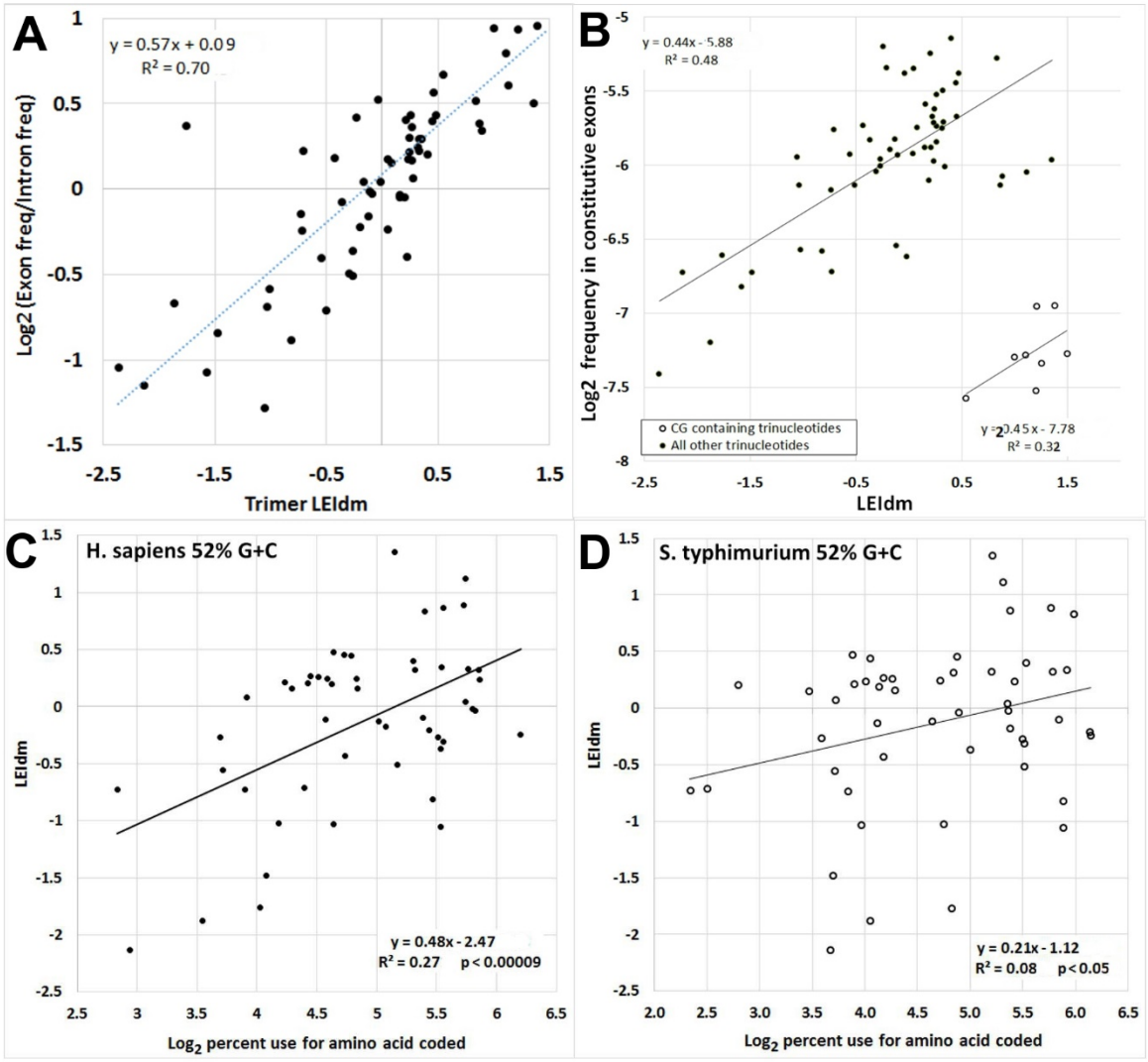


Figure S8. Splicing promotion by mutant trinucleotides correlates with genomic abundance in exons. (A) The ratio of the exonic frequency of each trinucleotide to flanking (100 nt) intronic frequency among 126,000 constitutive human exons is plotted against its average LEIdm value. LEIdm values are the averages from all mutated positions in all 10 Hexmuts. (B) The exonic frequency alone is plotted against LEIdm values. Unfilled symbols represent trinucleotides containing a CG dinucleotide; due to mutational instability CGs are underrepresented in the human genome and exome and have been omitted from the calculation of the regression line and its parameters. (C) LEIdm values for degenerate codons are plotted against their percent use. A moderate positive correlation (R=0.52) exists between splicing promotion and use. (D) As C but for *S. typhimurium* as a comparison, as pre-mRNA splicing does not take place in bacteria. This species was chosen because it has the same G+C content as human exons.

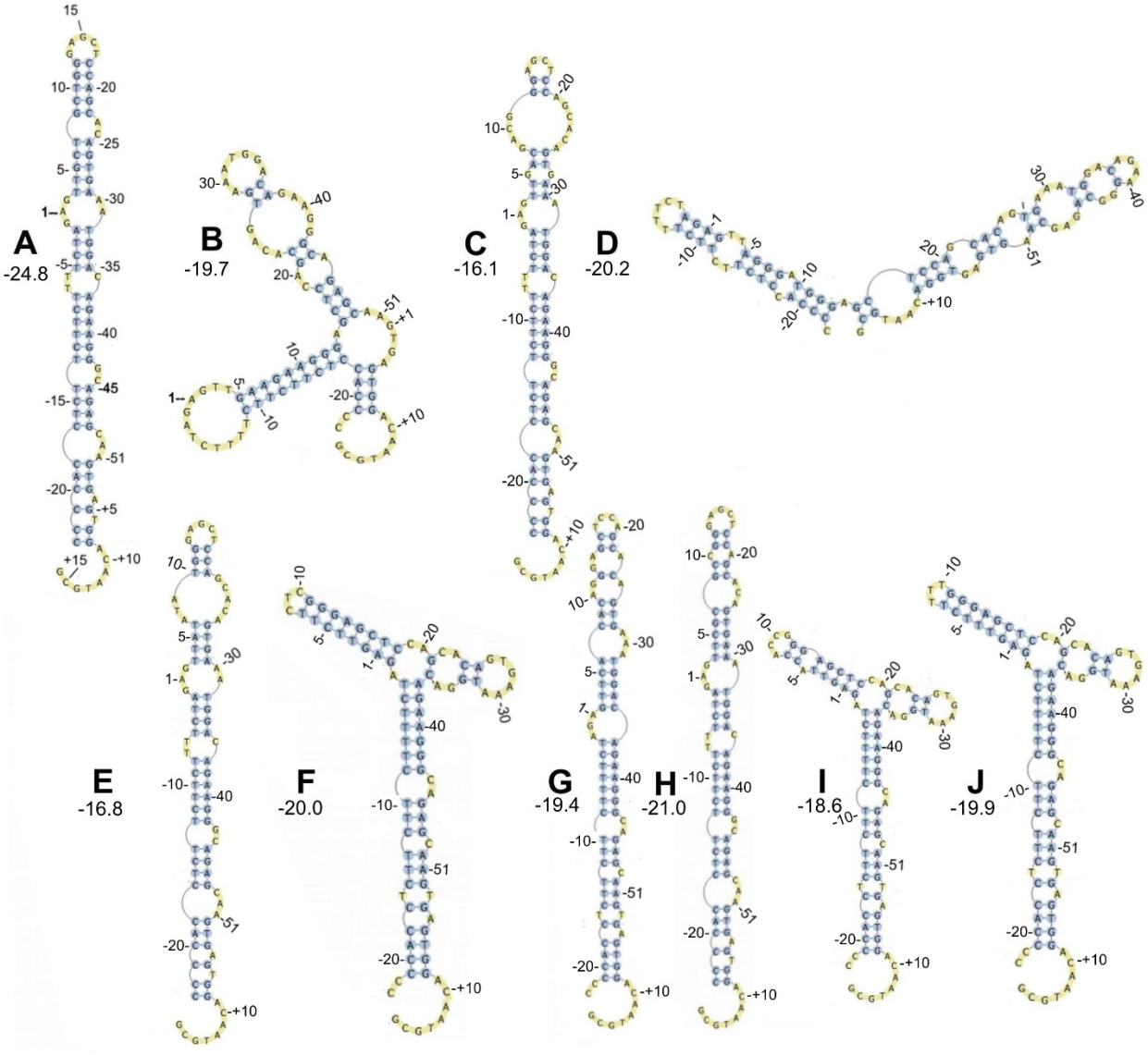


Fig. S9. The minimum free energy structures of the 10 Hexmuts. Regions of 90 nt from -23 to +16 relative to the exon were folded using RNAfold. Using longer flanks had little effect on the exon folding. The folding energies (ΔGº) of the MFE structures in kcal/mole are indicated. Structures were drawn using PseudoViewer3.

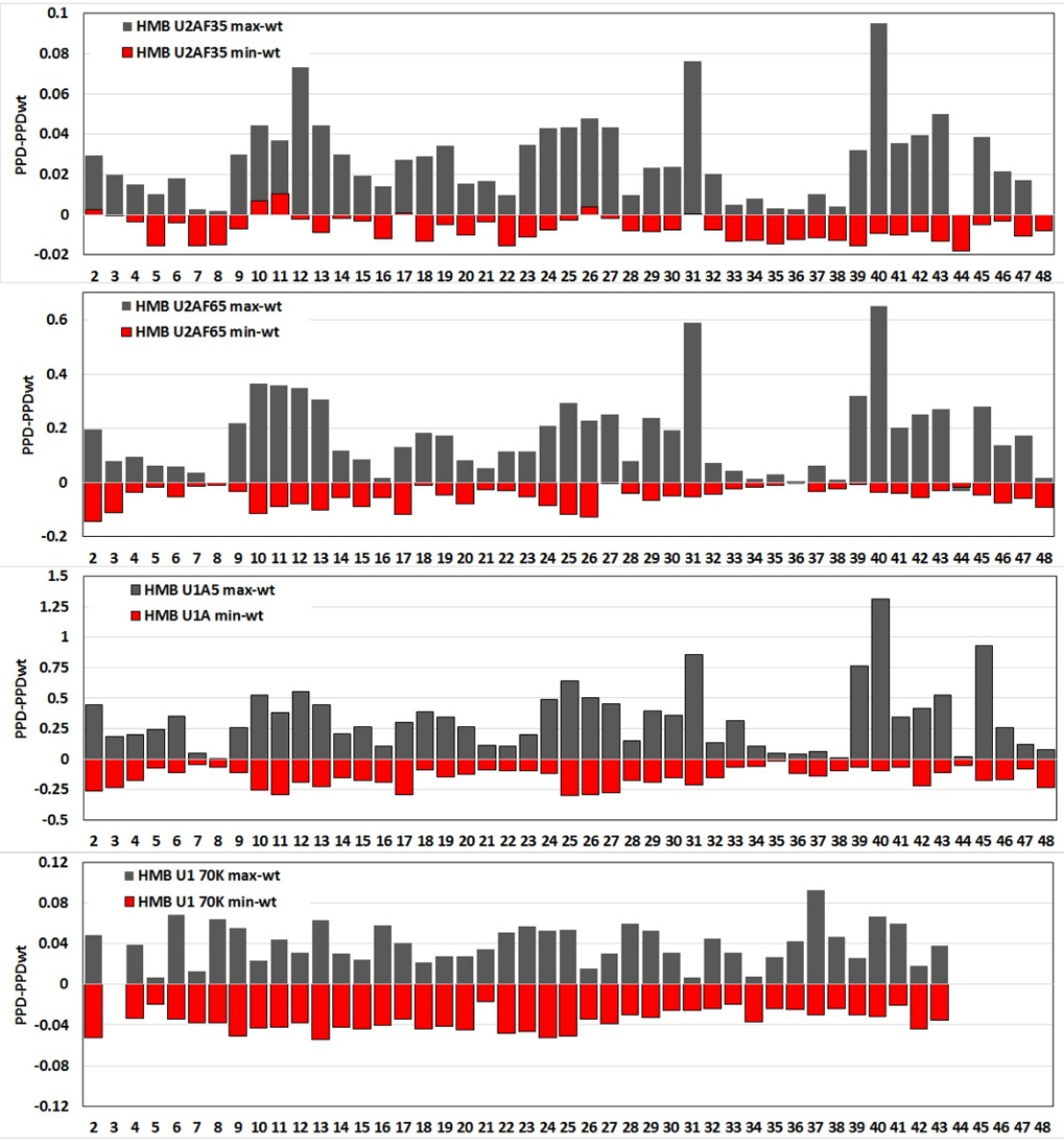


Fig. S10. Mutations affecting the *in vitro* binding of 4 spliceosomal proteins are distributed throughout the exon. Among the 12 mutations at each exonic position in HMB the gray bar indicates the maximum increase in binding and the red bar the maximum decrease.

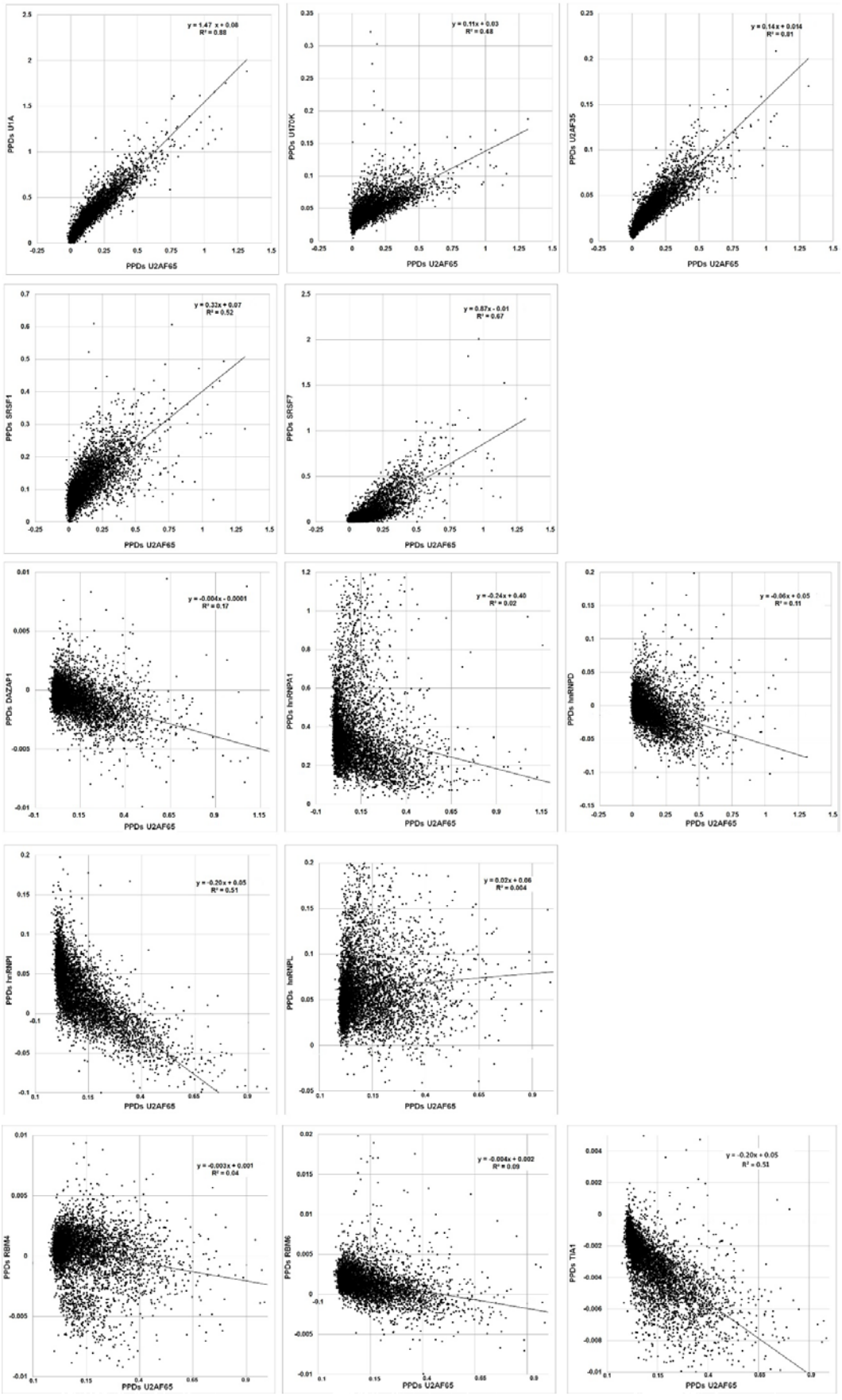


Fig. S11. Correlations of *in vitro* binding to mutant exon molecules comparing U2AF65 and 13 other RBPs in pairwise combinations. Scatter plots are shown with R^2 values indicated. The RBPs examined are indicated on the y axes. N=5217 to 5449; all p values are < 10-14.

Fig. S12

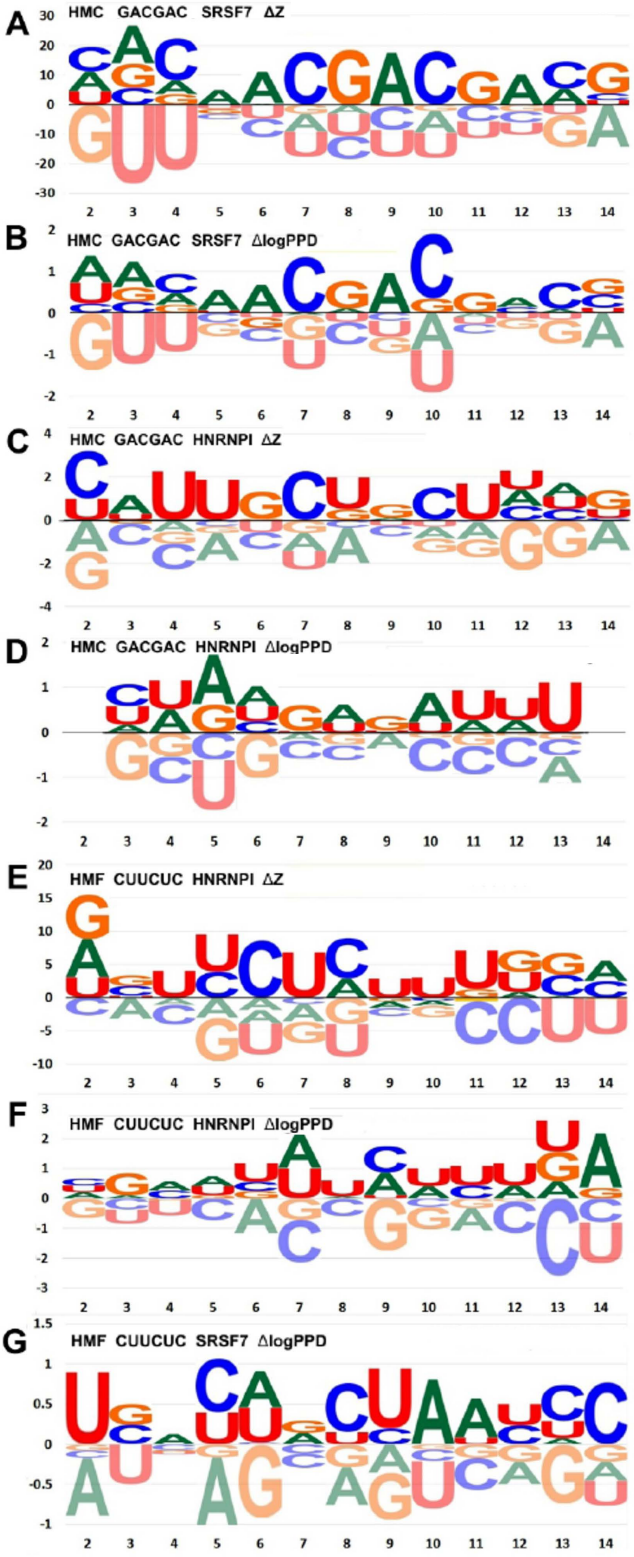


Fig. S12. Binding specificity in exon immunoprecipitation mirrors the specificity of the purified RNA-binding domain. Position-specific affinity matrices (PSAM, Foat et al., 2005) are shown for 13 mutated positions that span a Hexmut hexamer sequence at positions 5 to 10. HMC and HMF are compared. The hexmut regions in HMC and HMF were designed for binding SRSF7 (GACGAC) or HNRNPI (CUUCUC), respectively. PSAM logos designated ΔZ were generated using changes in mutants relative to the wild type for the summed CISBP-RNA Z-scores of the 156 mutated 7-mers for each Hexmut. PSAM logos designated ΔPPD were generated using the changes in IP pull down values. (A) *In vitro* binding of mutant HMC exons to the purified SRSF7 RNA binding domain (ΔZ) as reported by CISBP-RNA z-scores. (B) Immunoprecipitation of HMC exons targeting SRSF7 after exposure to a nuclear extract (ΔPPD). The logos in A and B are almost identical, indicating that the competitive environment of the nuclear extract had little effect on specificity in this PPD assay. (C) and (D) A control comparison for HMC, but focusing on HNRNPI as the RNA-binding protein. The expected CU-rich specificity is detected using ΔZ (panel C), despite the fact that the starting wild type sequence exhibits no resemblance to the HNRNPI consensus. However the IP (ΔPPD, pasnel D) was unable to detect this weak binding. (E) and (F) The same comparison as in C and D for HNRNPI but using HMF, which does carry a consensus HNRNPI binding sequence at position 5 to 10. Now both ΔZ and ΔPPD return an expected CU-rich logo, and once again the exon definition assay and the purified RBD yield similar results. (G) SRSF7 in HMF using IP (ΔPPD). Since there is little chance for the HMF sequence to be mutated to an SRSF7 binding sequence, the correct logo is not detected. The negative results of this control and panel D rule out the possibility that the background provided by the wild type sequence itself is capable of generating the correct logo.

Table S3. Single base mutations can have multiplex consequences on RBP binding (Z-scores) and splicing (EI): Two examples.

Pos. 39	Mut. pos.	7-mer start	EI	Z-scores	EI change	Z-score change	Conclusion
HNRNPA1	39	37					
WT 7mer		AG**A**AGGG	0.187	2.48	-0.186	+7.17	silencer
Mut. 454 7mer	A→T	AG**T**AGGG	0.001	9.65			
DAZAP1	39	37					
WT 7mer		AG**A**AGGG	0.187	5.69	-0.186	+5.08	silencer
Mut. 454 7mer	A→T	AG**T**AGGG	0.001	0.61			
MSI1	39	35					
WT 7mer		ACAG**A**AG	0.187	0.44	-0.186	+5.68	silencer
Mut. 454 7mer	A→T	ACAG**T**AG	0.001	6.12			
FXR2	39	33					
WT 7mer		GGACAG**A**	0.187	5.03	-0.186	-3.47	enhancer
Mut. 454 7mer	A→T	GGACAG**T**	0.001	1.56			
CNOT4	39	33					
WT 7mer		GGACAG**A**	0.187	7.32	-0.186	-5.88	enhancer
Mut. 454 7mer	A→T	GGACAG**T**	0.001	1.44			
Pos. 10							
HNRNPA1	10	8					
Mut. 95: TA at 9,10		GT**A**GGGA	0.045	9.76	-2.505	+9.21	silencer
Mut. 96: TC at 9,10	A→C	GT**C**GGGA	2.55	0.55			
RBM4	10	10					
Mut. 95: TA at 9,10		**A**GGGAGC	0.045	0.03	-2.505	-4.00	enhancer
Mut. 96: TC at 9,10	A→C	**C**GGGAGC	2.55	4.03			

Table S4. Significant splicing-RBP binding regressions in the 10 Hexmuts

HM	No. of significant regressions*	Proportion significant
HMA	813	0.18
HMB	1046	0.23
HMC	631	0.14
HMD	637	0.14
HME	824	0.18
HMF	694	0.16
HMG	679	0.15
HMH	690	0.15
HMI	710	0.16
HMJ	1003	0.22
average	773	0.17
median	702	0.16

* of 4459 total regressions attempted

Table S5. Empirical FDRs for significant LEI**:**z-score correlations for each Hexmut

HM	FDR
HMA	0.053
HMB	0.060
HMC	0.092
HMD	0.053
HME	0.055
HMF	0.077
HMG	0.054
HMH	0.084
HMI	0.079
HMJ	0.055
average	0.066
median	0.058

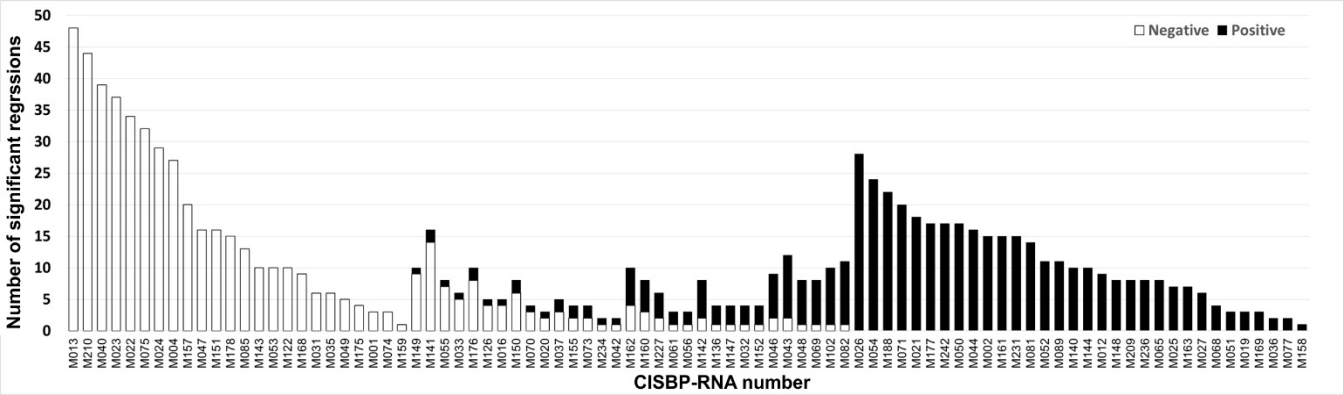


Fig. S13. Consistency of positive vs. negative correlations between splicing and RBP relative affinities. The sign of the correlation (black, positive; white, negative) is indicated for each of the 87 RBPs exhibiting significant correlations between splicing (LEI) and RBP relative affinities (z scores) for HexmutB. The RBPs are arranged from left to right in order of increasing positive to negative ratio and then by decreasing number of regressions. RBPs numbers are from the CISBP-RNA database.

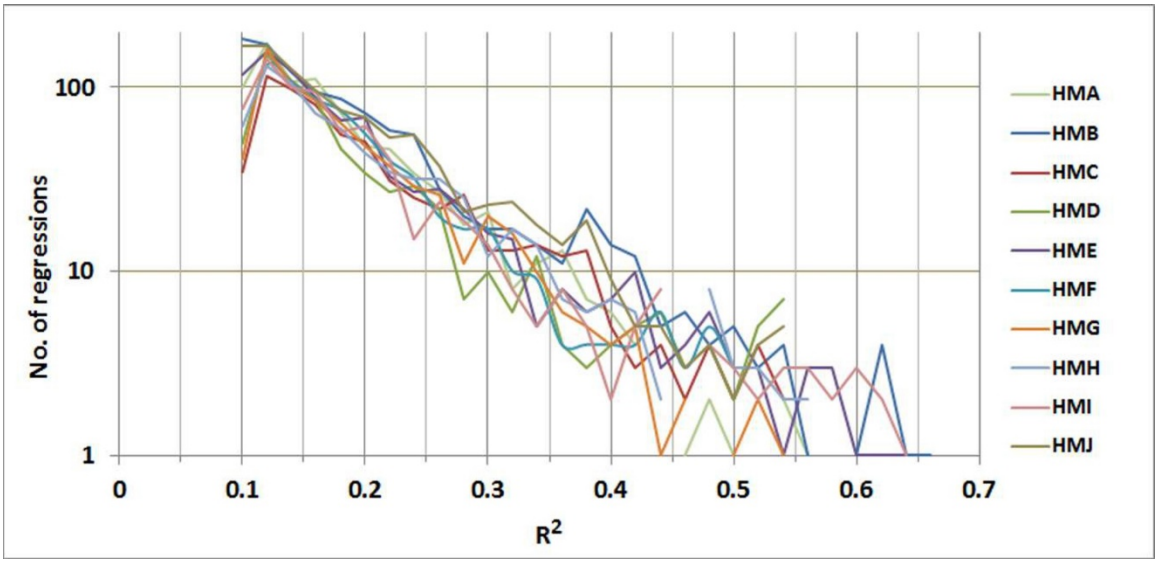


Fig. S14. Distribution of R^2 values for correlation between splicing (LEI) and RBP relative affinities (z-scores). The results for each of 10 Hexmuts are shown; 4459 regressions were performed for each, of which about 800 were significant at an FDR of 0.05.

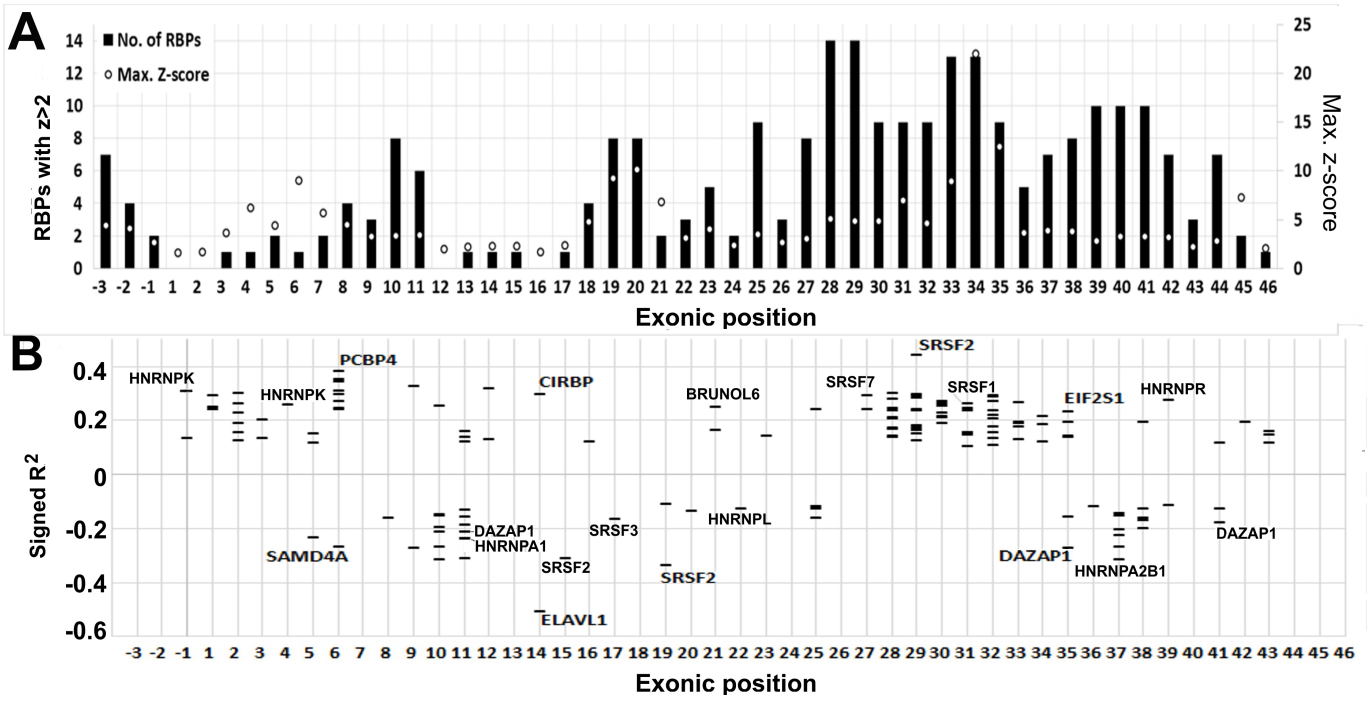
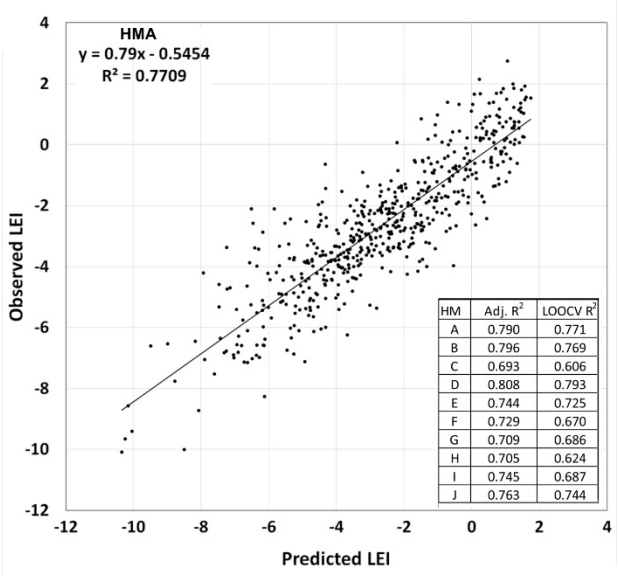


Fig. S15. RBP binding and splicing correlations for HMA, the wild type exon. (A) A map of the HMA exon showing the starting positions of RBP 7-mer binding sites with CIS-BP RNA z-scores >2. Open circles indicate the maximum z-score among RBPs bound. Taking into account the length of the 7-mers, the entire exon is covered with significsnt RBP binding sites. (B) A map of the HMA exon showing signed R^2 values for significant correlations between splicing (LEI) and mutant z-scores that decreased relative to the wild type z score (i.e., loss of function mutations, implying the wild type sequence mediated functional binding). A positive correlation infers an enhancer sequence and a negative correlation infers a silencer. Some of the RBPs that yielded high absolute R^2 values are labeled. Note that RBPs such as SRSF1 and SRSF7 exhibited positive correlations and HNRNPs such as HNRNPA1 and DAZAP1 exhibited negative correlations.



HM	Adj. R^2	LOOCV R^2
A	0.790	0.771
B	0.796	0.769
C	0.693	0.606
D	0.808	0.793
E	0.744	0.725
F	0.729	0.670
G	0.709	0.686
H	0.705	0.624
I	0.745	0.687
J	0.763	0.744

Fig. S16. Prediction of splicing efficiencies for each set of Hexmut mutants. An equation for multiple linear regression was derived using HMA data only; 40 significant protein-positions were found and used. The results shown are values derived by leave-one-out cross validation (LOOCV) for each of the 556 sequences. The inset table shows the R^2 values for the leave-one-out cross validation of all 10 Hexmuts analyzed separately.

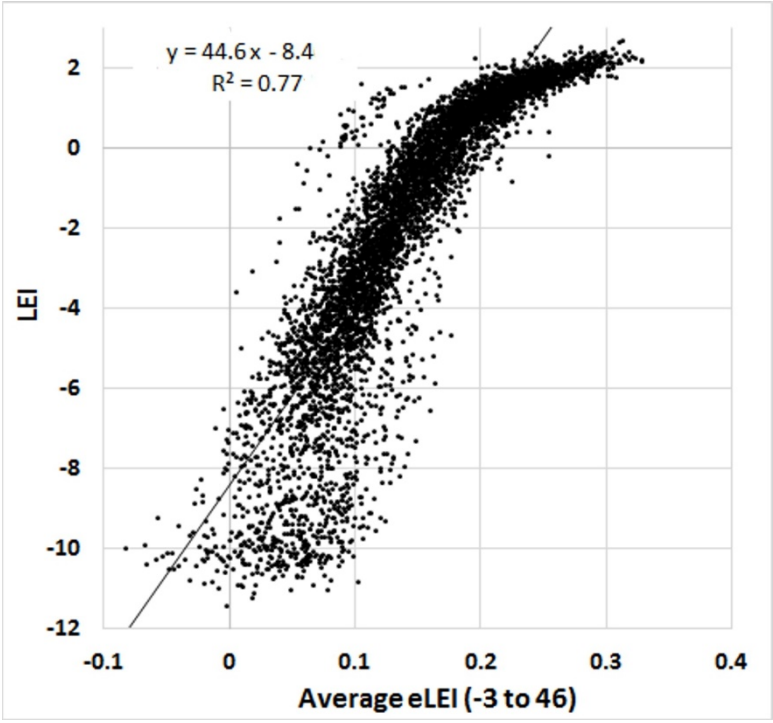


Fig. S17. Correlation between splicing efficiency and average eLEI values across the exons. Splicing is expresed as LEI (y-axis). The x-axis plolts the average of eLEI values for all 7-mer from positions -3 to 46. The eLEI values are derived from the 7-mers generaterd in the mutant population. N = 5560 exons.

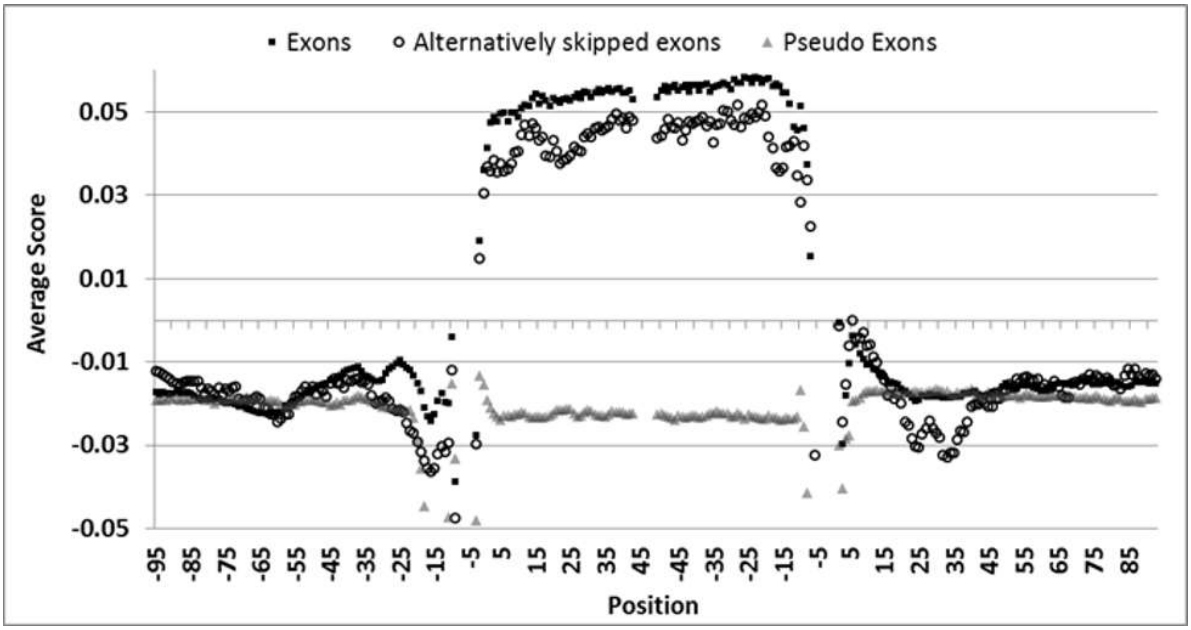


Fig. S18. Composite map of SMS scores for 7-mers at single nt resolution across real and pseudo exons. Data was collected for approximately 100,000 constitutive, 30,000 alternative and 100,000 pseudo exons. Only the 50 most upstream and downstream nts of the exons were scored. Some extremely low points as expected at the distinctive splice site sequences are not shown.

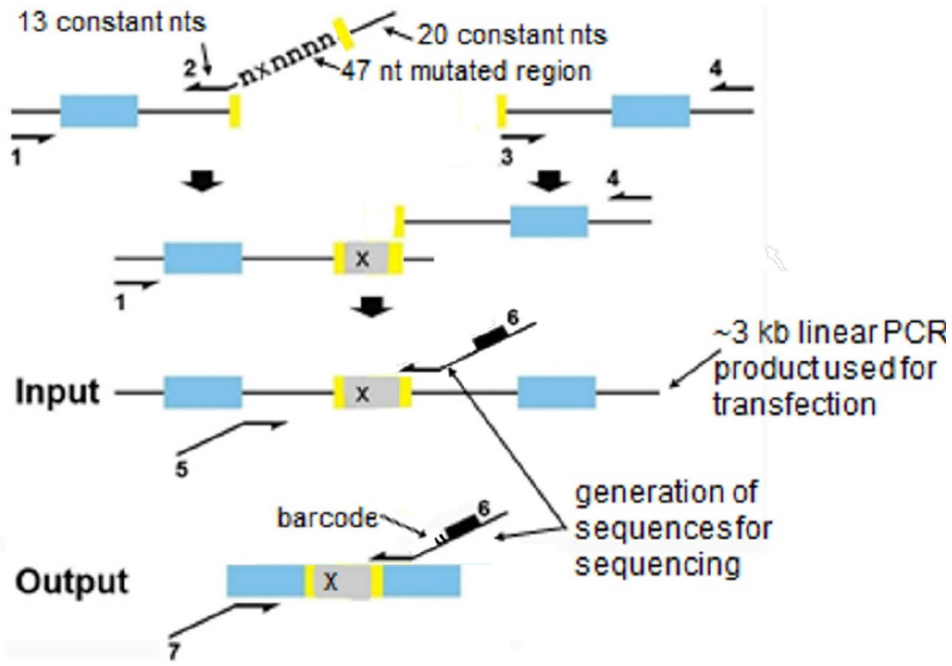


Fig.S19. Scheme for mutant library generation. Overlap extension PCR was used to construct 3-exon minigenes of about 3000 nt that were used directly for transfection. The central exon is gray with yellow edges indicating the few nts spared from mutagenesis. Primers are numbered for distinction; primer sequences are available on request. Primers 2 comprised the 80 nt products of the primer extension plus ligation using the custom DNA microarray. Primers 6 were barcoded with CG or AT (2 short bars) and used to distinguish cDNAs from independent transfections. The black bar on primers 6 denotes the template for the Illumina sequencing primer.

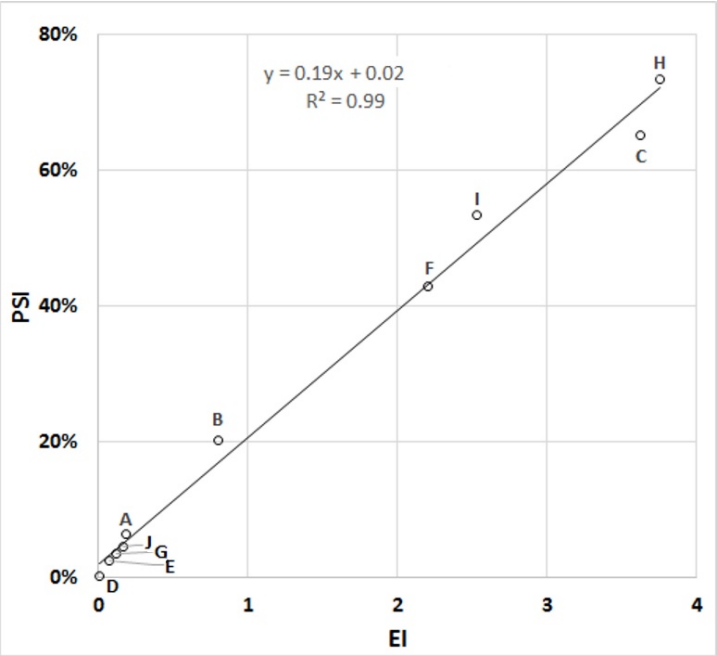


Fig. S20. Linear relationship between EI values and empirically measured psi values for the 10 WT Hexmuts A to J, as indicated. Psi values for all mutant molecules were calculated using this data as a calibration.

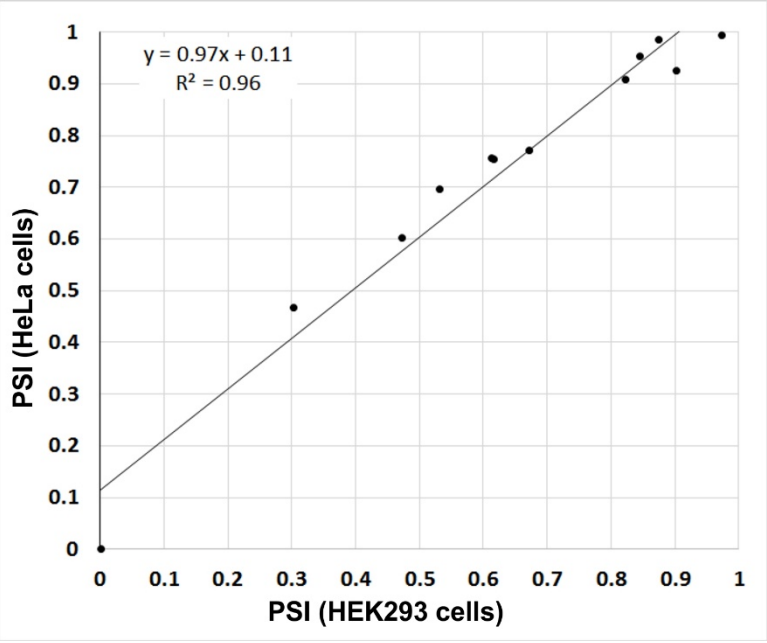


Fig. S21. Minigene mutants mature similarly in HeLa and HEK293 cells. The observed proportion spliced in (psi) resulting from testing 10 cloned minigenes transfected into HeLa cells (y-axis) were compared to HEK293 cells (x-axis). After transfection, RT–PCR products were visualized on ethidium bromide-stained gels and quantified using ImageJ. Psi is included/(included + skipped). The mutants used (in order of descending psi) were: 4704, 3907, 3265, 3967, 1637, 567, 3090, 240, 2707, 5489, 3456, and 2046 (Table S2).

Table S9. List of the first 25 mutant sequences to illustrate mutagenesis scheme

1. tctagAGTTGCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
2. tctagA**AA**TGCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
3. tctagA**AC**TGCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
4. tctagA**AG**TGCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
5. tctagA**A**TTGCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
6. tctagA**CA**TGCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
7. tctagA**CC**TGCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
8. tctagA**CG**TGCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
9. tctagA**C**TTGCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
10. tctagA**TA**TGCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
11. tctagA**TC**TGCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
12. tctagA**TG**TGCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
13. tctagA**T**TTGCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
14. tctagAG**AA**GCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
15. tctagAG**AC**GCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
16. tctagAG**AG**GCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
17. tctagAG**A**TGCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
18. tctagAG**CA**GCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
19. tctagAG**CC**GCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
20. tctagAG**CG**GCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
21. tctagAG**C**TGCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
22. tctagAG**GA**GCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
23. tctagAG**GC**GCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
24. tctagAG**GG**GCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg
25. tctagAG**G**TGCTGCTGGGAGCTCCAGCACAGTGAAATGGACAGAAGGGCAGAGCAAgtg

:

Bold red characters are those that differ from the wild type. Note that, for instance, SBSs at position 3 appear on lines 17, 21 and 25. Thus all 15 possible SBSs and DBSs stem from 12 entries per position.

Table S10. Antibodies used for immunoprecipitation

Target	Type	Source	Mfr. No.
U2AF65	mouse	Sigma	MC3
U2AF35	rabbit	Abcam	ab86305
U1A	mouse	Abcam	ab55751
U1-70K	goat	Santa Cruz	sc-9571
SRSF1	mouse	Santa Cruz	sc-33652
SRSF7	rabbit	Bethyl	A303-773A
DAZAP1	rabbit	Abcam	ab168820
hnRNPA1	mouse	Sigma	9H10
hnRNPD/AUF1	rabbit	Abcam	ab50692
hnRNPI	mouse	Santa Cruz	sc-56701
hnRNPL	rabbit	Bethyl	A303-896A
RBM4	rabbit	Proteintech	16614-1-AP
RBM6	rabbit	Bethyl	A301-013A
TIA1	rabbit	Abcam	ab140595