

Supplementary Information

Promoter identity shapes splicing outcomes and fidelity

Shaked Shanas, Anton M. Schwartz, Noa Neshet, Avital Melnik and Martin Mikl

Supplementary Figures

Supplementary Figure 1: Library overview.

Supplementary Figure 2: Characterization of experimental conditions

Supplementary Figure 3: Effect size comparison of experimental conditions.

Supplementary Figure 4: Genomically integrated reporter libraries maintain promoter-specific differences.

Supplementary Figure 5: Determining experimental reproducibility.

Supplementary Figure 6: Comparison of binary vs. non-binary cassette exons.

Supplementary Figure 7: The effect of pharmacological treatments and promoter switching on overall splicing behavior.

Supplementary Figure 8: The effect of designed sequence changes on PSI and expression levels.

Supplementary Figure 9: Exons in the reporter library cluster based on their sensitivity to promoter identity.

Supplementary Figure 10. Motif enrichment analysis between exon clusters.

Supplementary Figure 11. Motif enrichment is associated with changes in exon inclusion rates.

Supplementary Figure 12. Multiple isoforms are created through cryptic splice site usage.

Supplementary Figure 13. Cryptic splice sites resemble donor and acceptor sequences.

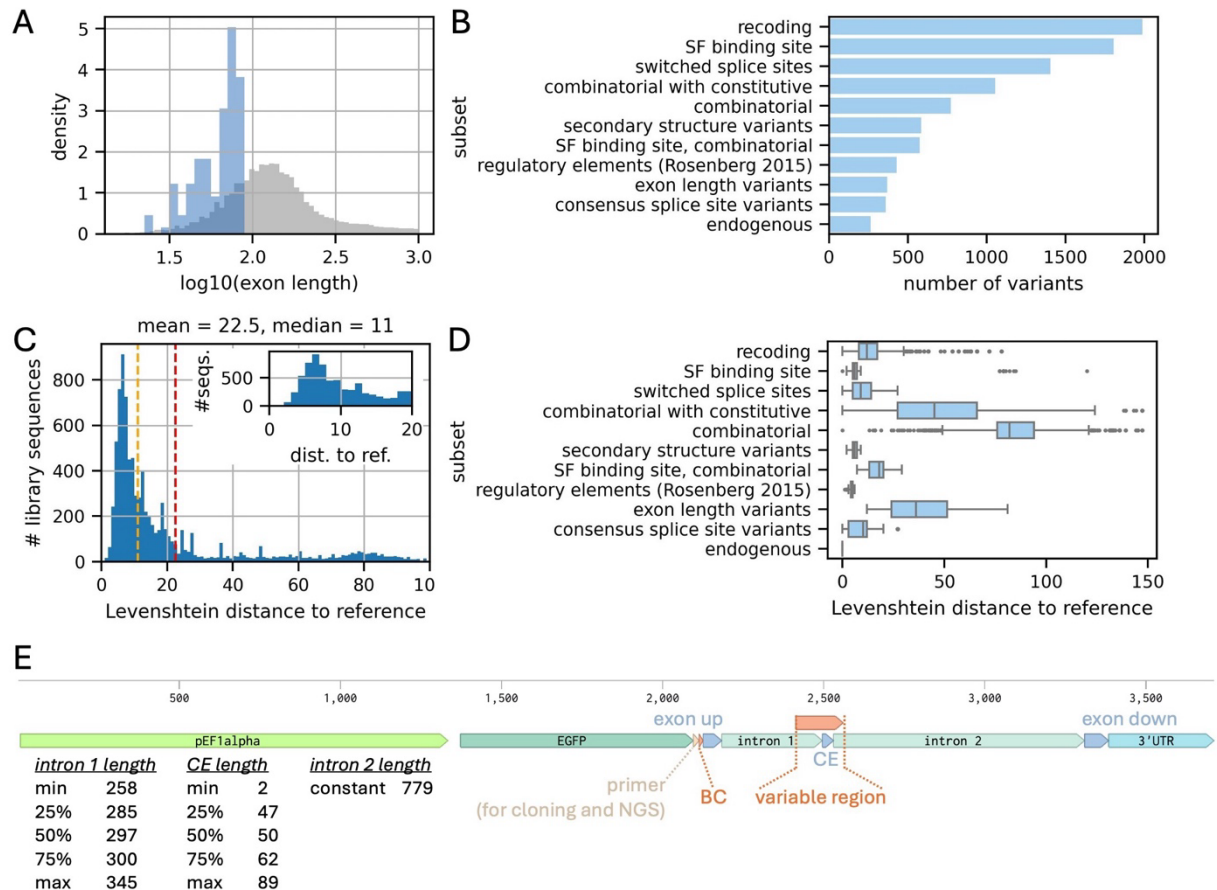
Supplementary Figure 14. Promoter identity affects cryptic splice site usage.

Supplementary Figure 15. Characterization of endogenous promoter constructs.

Supplementary Figure 16. The influence of endogenous promoters on splicing factor motif effects and splice site choice.

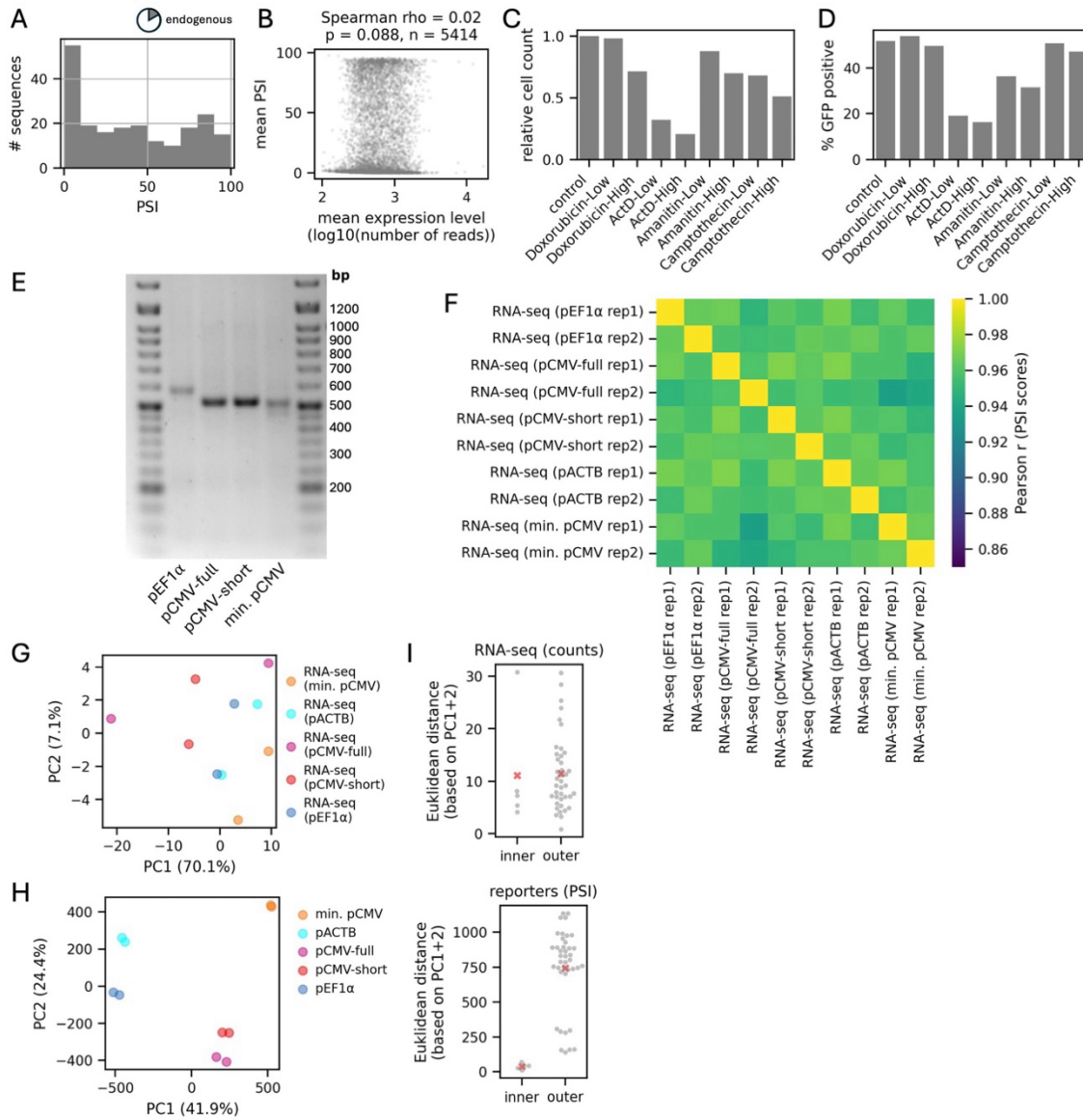
Supplementary Figure 17. Outline of the gating strategy used in the flow cytometry experiments based on three representative examples.

Supplementary Figures



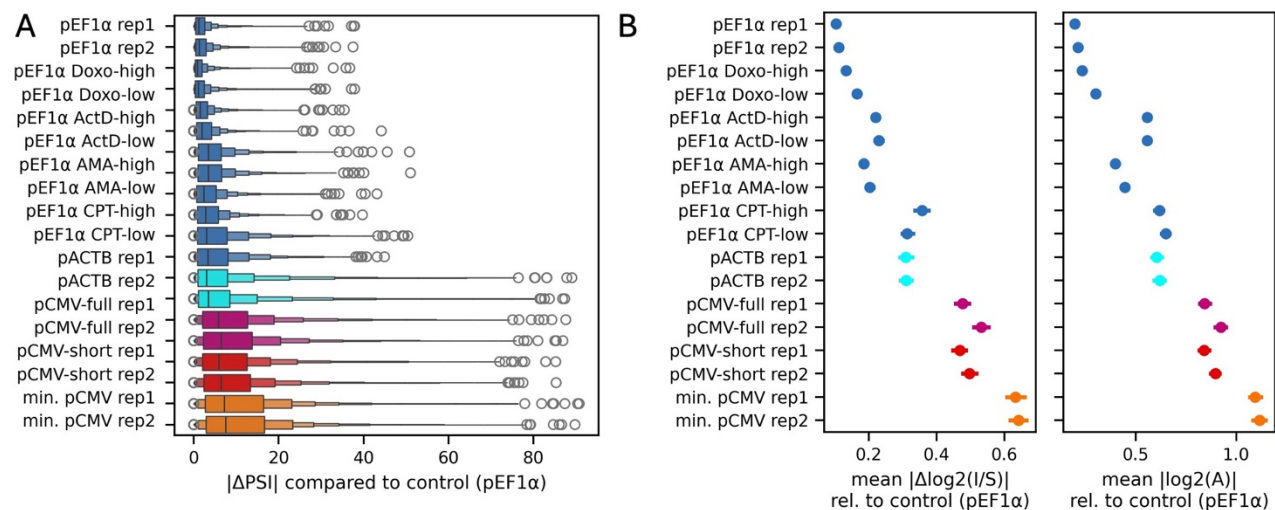
Supplementary Figure 1: Library overview.

A. Histogram showing the length distribution of all exons (grey, based on GRCh38) and exons underlying the sequence library (blue). B. Number of variants in each library subset (total $n=9608$). C. Histogram showing the distribution of the Levenshtein distance of each variant to its corresponding wild-type sequence, not including the “endogenous” subset; the inset shows the range of 0-20 ($n=9340$). D. Boxplot showing the distribution of Levenshtein distance of variants in each subset (ordered by the size of the subset, the number of variants in each subset are shown in panel B); box plots show the median (central line) and interquartile range (box, 25th-75th percentiles), with whiskers extending to the most extreme data points within 1.5x of the interquartile range; outliers are shown as points. E. Schematic of the mini-gene reporter, including overviews of intron and cassette exon (CE) length variability (minimum and maximum lengths, as well as 25th, 50th and 75th percentile across all library variants).



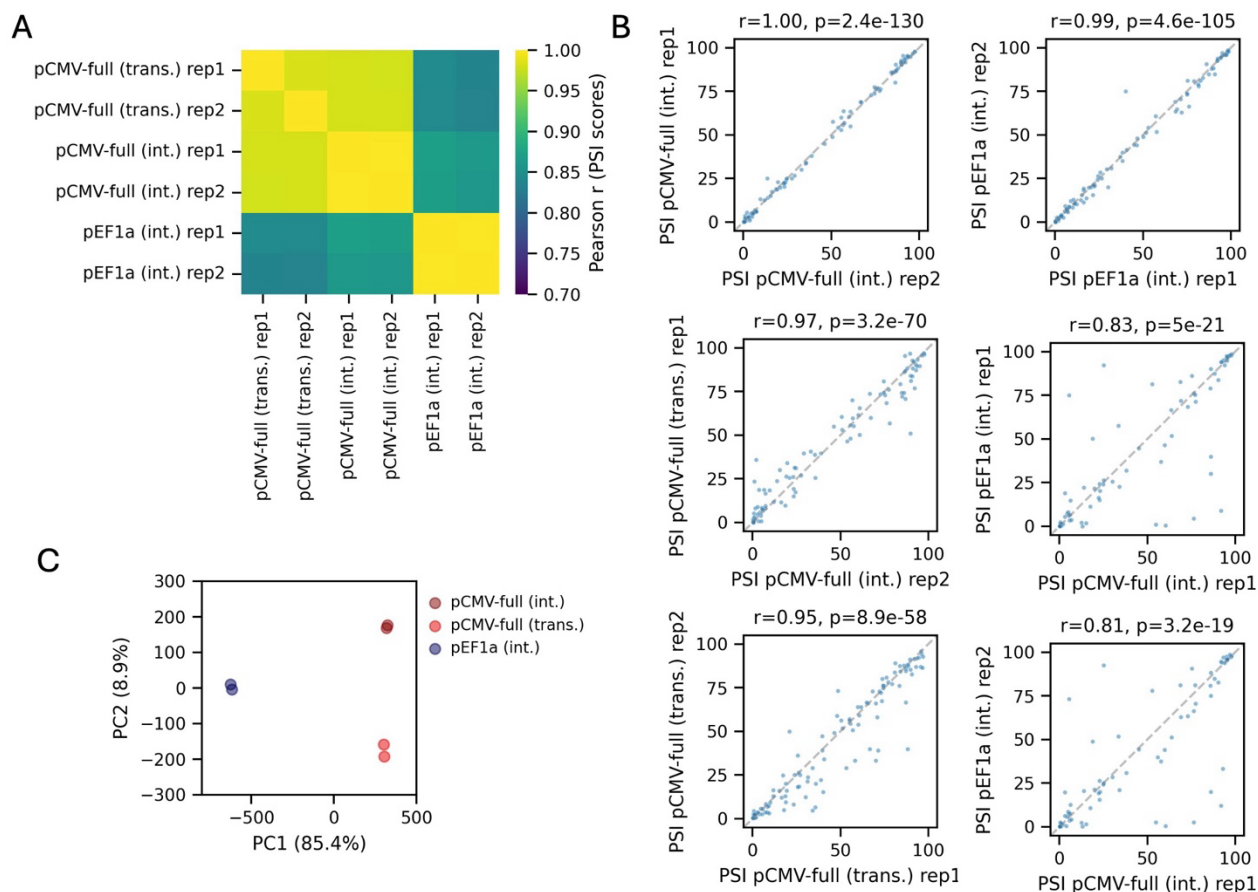
Supplementary Figure 2: Characterization of experimental conditions.

A. Distribution of PSI values as determined for reporter library variants containing native cassette exon sequences (covered by min. 100 reads) under the control of pEF1 α ; n=206. B. Scatterplot showing the relationship between mean PSI scores across samples and mean expression level (approximated by log10(number of reads) given that all variants in a sample are expressed from the same promoter); n=5414 (min. 100 reads in all samples). C. Cell counts (relative to transfection with the plasmid and mock treatment, “control”) as determined using a Novocyte flow cytometer, based on >4500 cells measured for each condition. D. Percent GFP positive cells (relative to transfection with the plasmid and mock treatment, “control”) as determined using a Novocyte flow cytometer, based on >4500 cells measured for each condition. E. Agarose gel showing the results of the 5'RACE on pEF1 α and the CMV promoter variants; the expected sizes (relative to the expected transcription start site) are ~590 bp for pEF1 α and ~510 for the CMV variants. F. Heatmap of Pearson correlation coefficients between log10(UMI) values measured in all conditions for 7449 genes with more than 10 UMIs in all samples. GH. The first two principle components are shown after PCA based on (G) log10(UMI) for 7449 human genes or (H) PSI values of 5653 constructs measured for different promoter constructs (>99 reads/variant/sample). I. Eukclidean distances calculated on G (top) or H (bottom), for pairs of samples within the same group (“inner”) or pairs of samples across groups (“outer”).



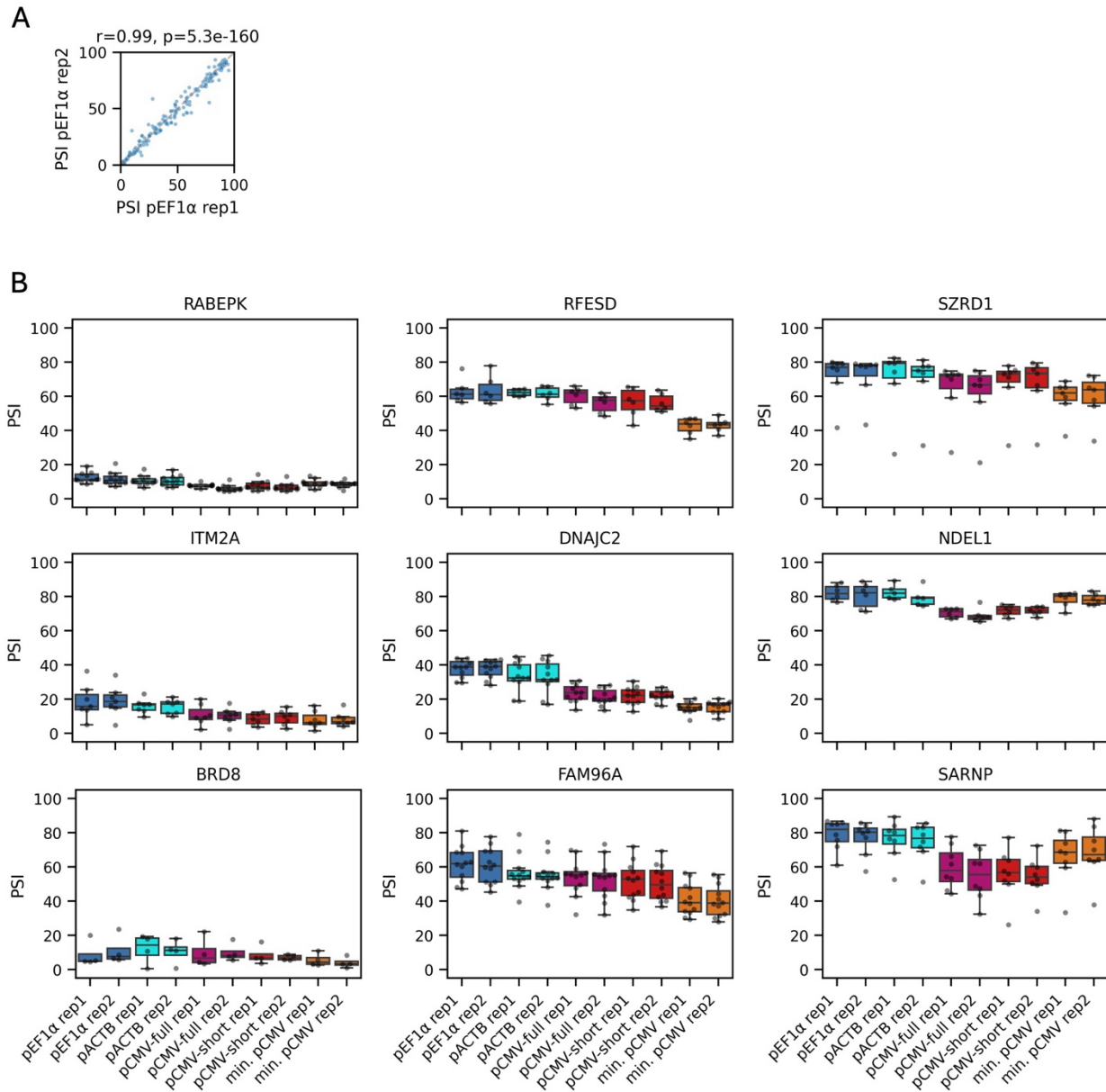
Supplementary Figure 3: Effect size comparison of experimental conditions.

A. Absolute change in PSI compared to the reference (mean of pEF1 α replicates) for all variants covered in the reference (>9 reads); n=7627, 7587, 7605, 7736, 7723, 7863, 7505, 7572, 7723, 7771, 7550, 7709, 7183, 7141, 7405, 7422, 7803, 7697, 7189, 7048 in the order they appear in the graph, box plots show the median (central line) and interquartile range (box, 25th-75th percentiles), with whiskers extending to the most extreme data points within 1.5x of the interquartile range; outliers are shown as points. B. Mean absolute change in log2(exon included/exon skipped) (left) compared to the reference (mean of pEF1 α replicates) or mean absolute effect size (log2(A)) of the experimental condition compared to the reference (mean of pEF1 α replicates). Errorbars denote 99% CI. (n=4461 and 4336).



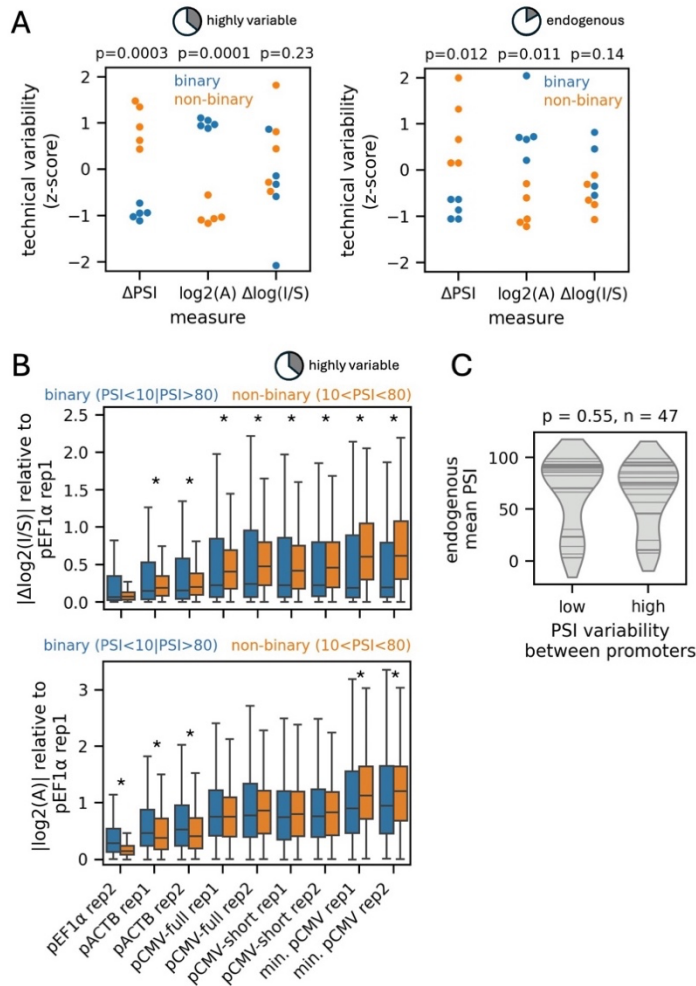
Supplementary Figure 4: Genomically integrated reporter libraries maintain promoter-specific differences.

A. Heatmap of Pearson correlation coefficients between PSI values measured in integrated and episomally expressed reporter libraries (n=1900). B. Scatterplots comparing PSI values for reporters containing native cassette exon sequences between integrated and episomally expressed reporter libraries; Pearson correlation coefficient and the associated p-values are shown above (n>66 for each comparison). C. The first two principle components are shown after PCA based on PSI values of 1900 constructs measured in integrated and episomally expressed reporter libraries.



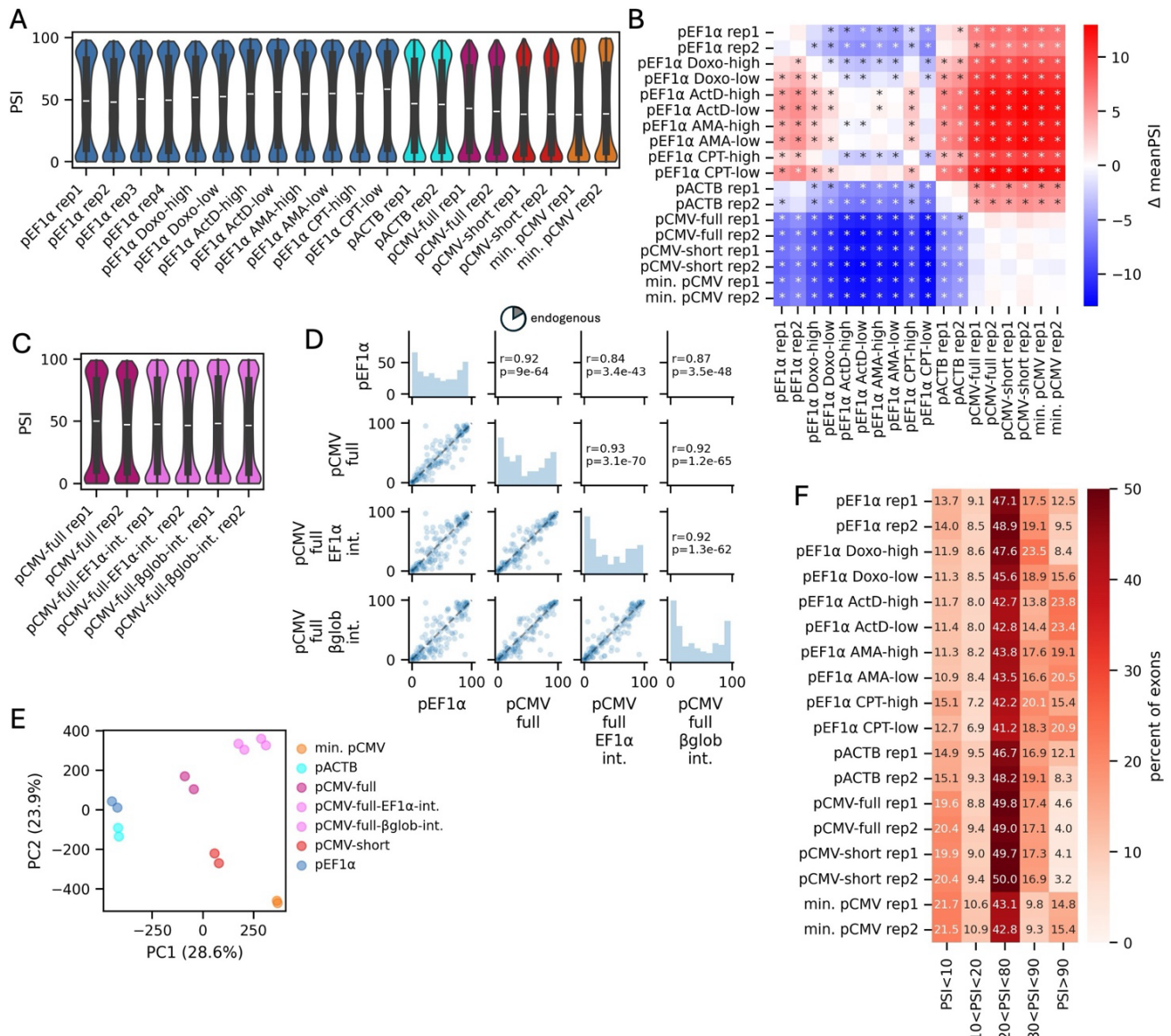
Supplementary Figure 5: Determining experimental reproducibility.

A. Scatterplot comparing PSI values for reporters containing native cassette exon sequences between biological replicates; Pearson correlation coefficient and the associated p-values are shown above; $n=206$. B. Boxplots showing PSI values in reporters carrying native cassette exon sequences expressed under the control of the indicated promoter; each datapoint is one library variant with the identical variable region, but different barcode ($n=4-12$ for each group); box plots show the median (central line) and interquartile range (box, 25th-75th percentiles), with whiskers extending to the most extreme data points within $1.5\times$ of the interquartile range; outliers are shown as points.



Supplementary Figure 6: Comparison of binary vs. non-binary cassette exons.

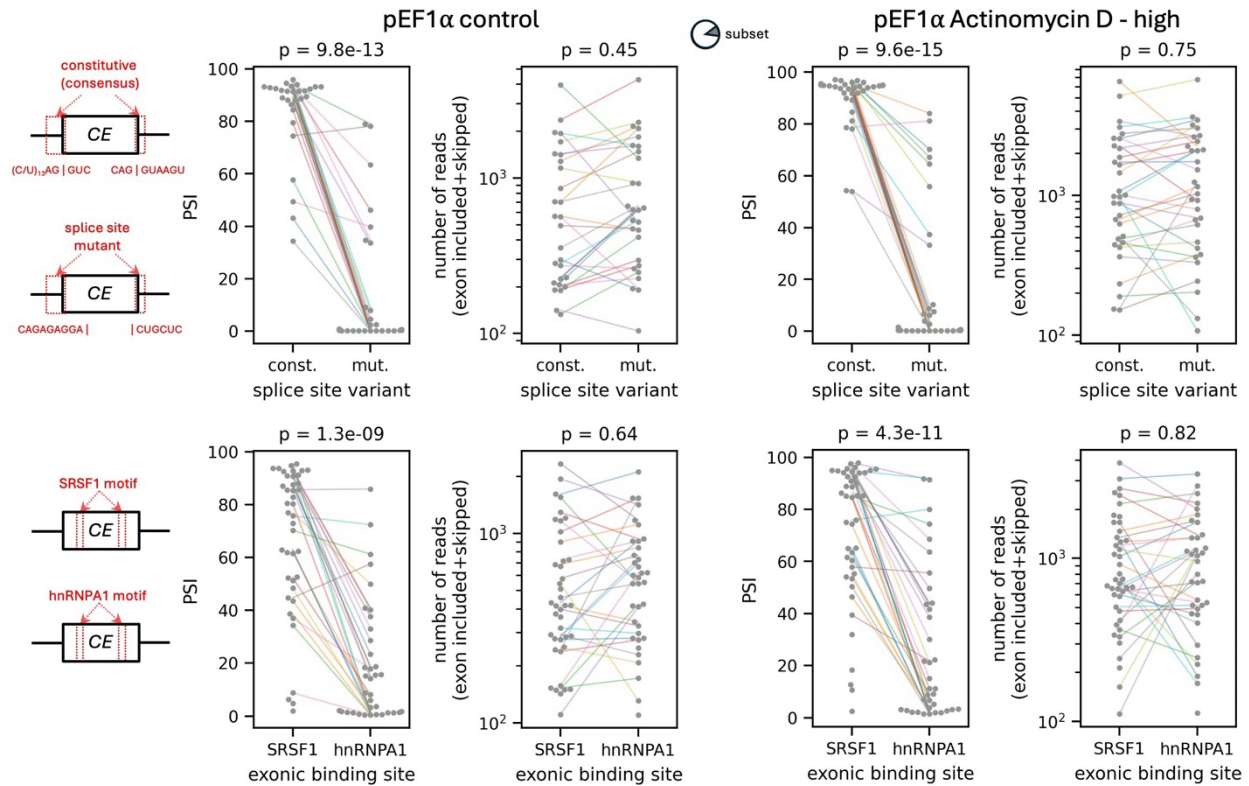
A. The technical variability, measured as the z-scored mean absolute difference between pairs of replicates for binary (blue) or non-binary (orange) cassette exons, is shown for three different measures of effect size, each data point represents one of the promoter constructs, p-values (paired t-test) for the comparison of binary and non-binary exons within each group is shown above; $n=5$. B. Boxplots showing the absolute difference in log ratio of exon included to exon skipped ($\log_2(I/S)$) (top) or absolute effect size ($\log_2(A)$) (bottom) for all pairwise comparisons to pEF1 α replicate 1; asterisks indicate significant differences between the two groups ($p < 0.01$, Mann-Whitney-U test); $n=794$ and 428 , box plots show the median (central line) and interquartile range (box, 25th-75th percentiles), with whiskers extending to the most extreme data points within $1.5\times$ of the interquartile range. C. Violin plot showing the mean endogenous PSI values across tissues for non-binary exons with low (< 50 th percentile) or high (> 50 th percentile) variability between promoters; $n=29$.



Supplementary Figure 7: The effect of pharmacological treatments and promoter switching on overall splicing behavior.

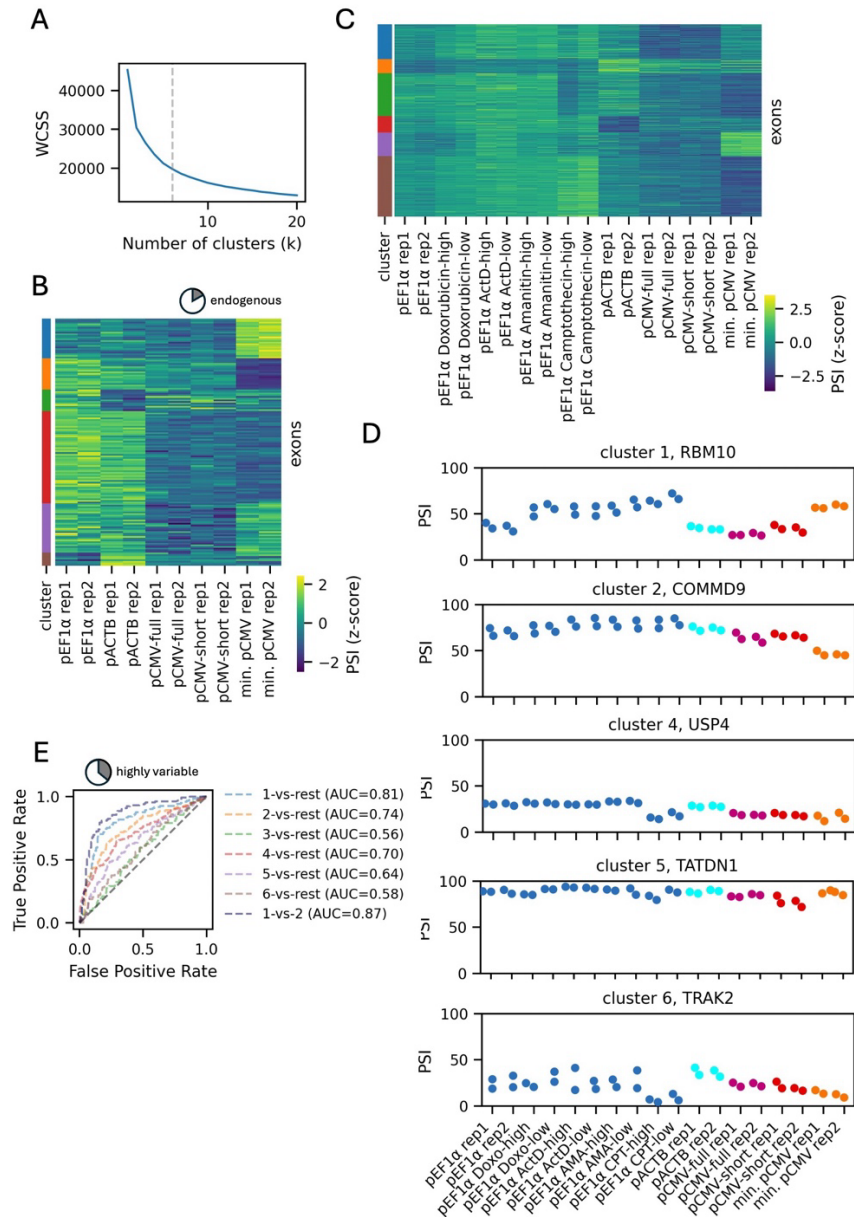
A. Violinplots showing the overall distribution of PSI values in the tested promoter variants and pharmacological treatments (n=7627, 7587, 7605, 7736, 7723, 7863, 7505, 7572, 7723, 7771, 7550, 7709, 7183, 7141, 7405, 7422, 7803, 7697, 7189, 7048 in the order they appear in the graph); inner box plots show the median (central line) and interquartile range (box, 25th-75th percentiles), with whiskers extending to the most extreme data points within 1.5x of the interquartile range. B. Heatmap showing the difference in mean PSI values for all pairwise combinations of samples; asterisks indicate statistically significant differences (Mann-Whitney-U test, corr. p<0.05, Bonferroni). C. Violinplots showing the overall distribution of PSI values in the CMV-full promoter variants with or without additional intron (n=7539, 7576, 7145, 7151, 6797, 6889 variants, in the order they appear in the graph); inner box plots show the median (central line) and interquartile range (box, 25th-75th percentiles), with whiskers extending to the most extreme data points within 1.5x of the interquartile range. D. Scatterplots comparing PSI values for reporters containing native cassette exon sequences between promoter variants with or without additional 5'UTR intron; Pearson correlation coefficient and the associated p-values are shown in the upper right half; the graphs along the diagonal show the distribution of PSI values in each sample; n=155 for each comparison. E. The

first two principal components are shown after PCA based on PSI values of 5575 constructs measured in different promoter constructs with or without 5'UTR intron. F. Heatmap showing the percent of library exons in the indicated PSI groups (n=5653 (=100%)).



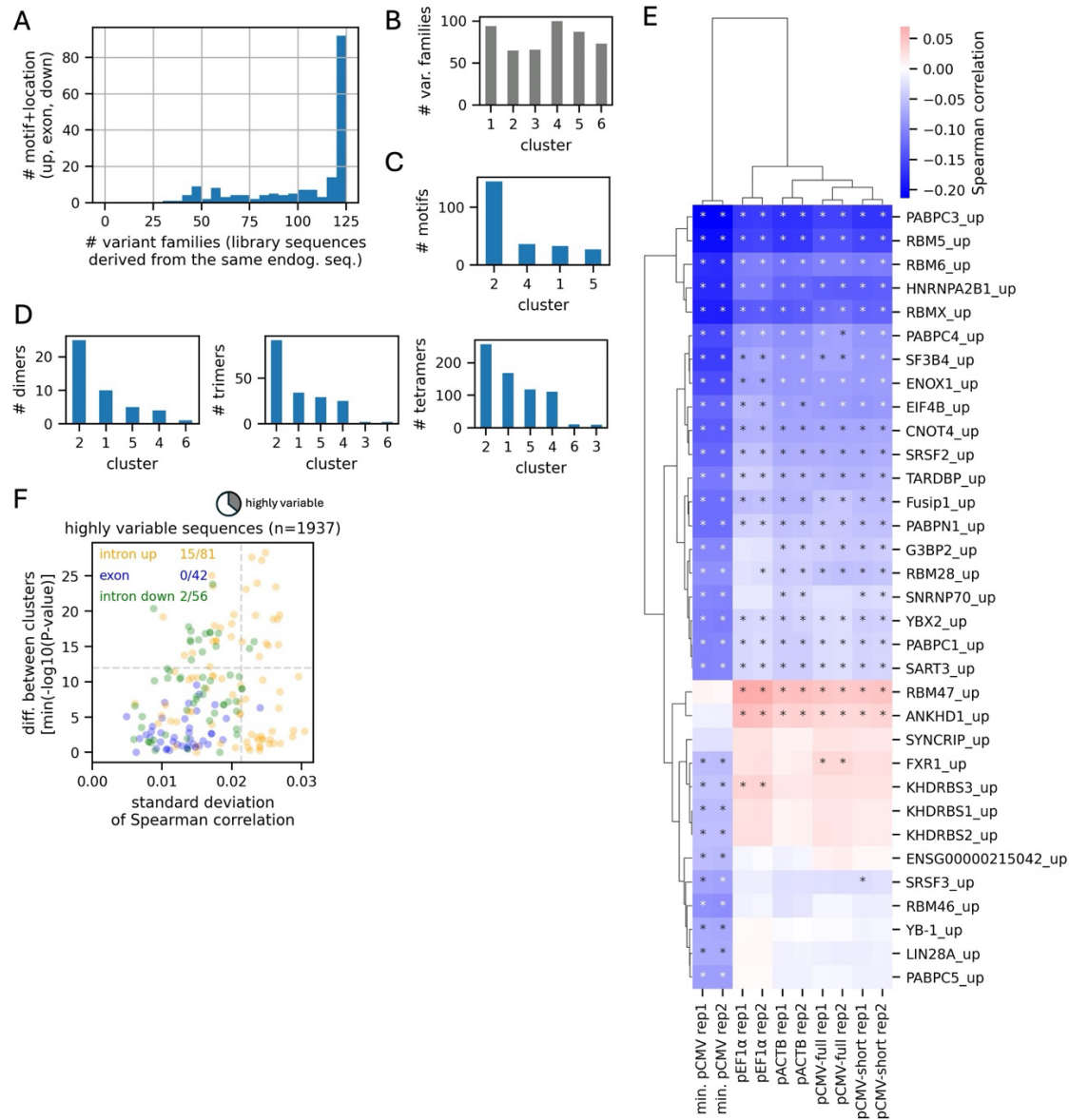
Supplementary Figure 8: The effect of designed sequence changes on PSI and expression levels.

The schemes on the left illustrate the systematic sequence changes introduced to bias cassette exon inclusion rates (top: replacing the inserts with consensus splice sites or non-spliceable sequences; bottom: introducing strong exonic splicing enhancer (SRSF1 motif - UCACACGAC) or silencer (hnRNPA1 motif - UUAGGGAAC) elements). The graphs to the right show measurements of PSI or total number of reads for pairs of sequences (connected by lines) derived from the same wild-type, but carrying either the sequence change promoting exon inclusion or the one promoting exon skipping (n=28 and 32 for the consensus or splice site mutant variants, n=36-41 for the splicing factor motif variants).



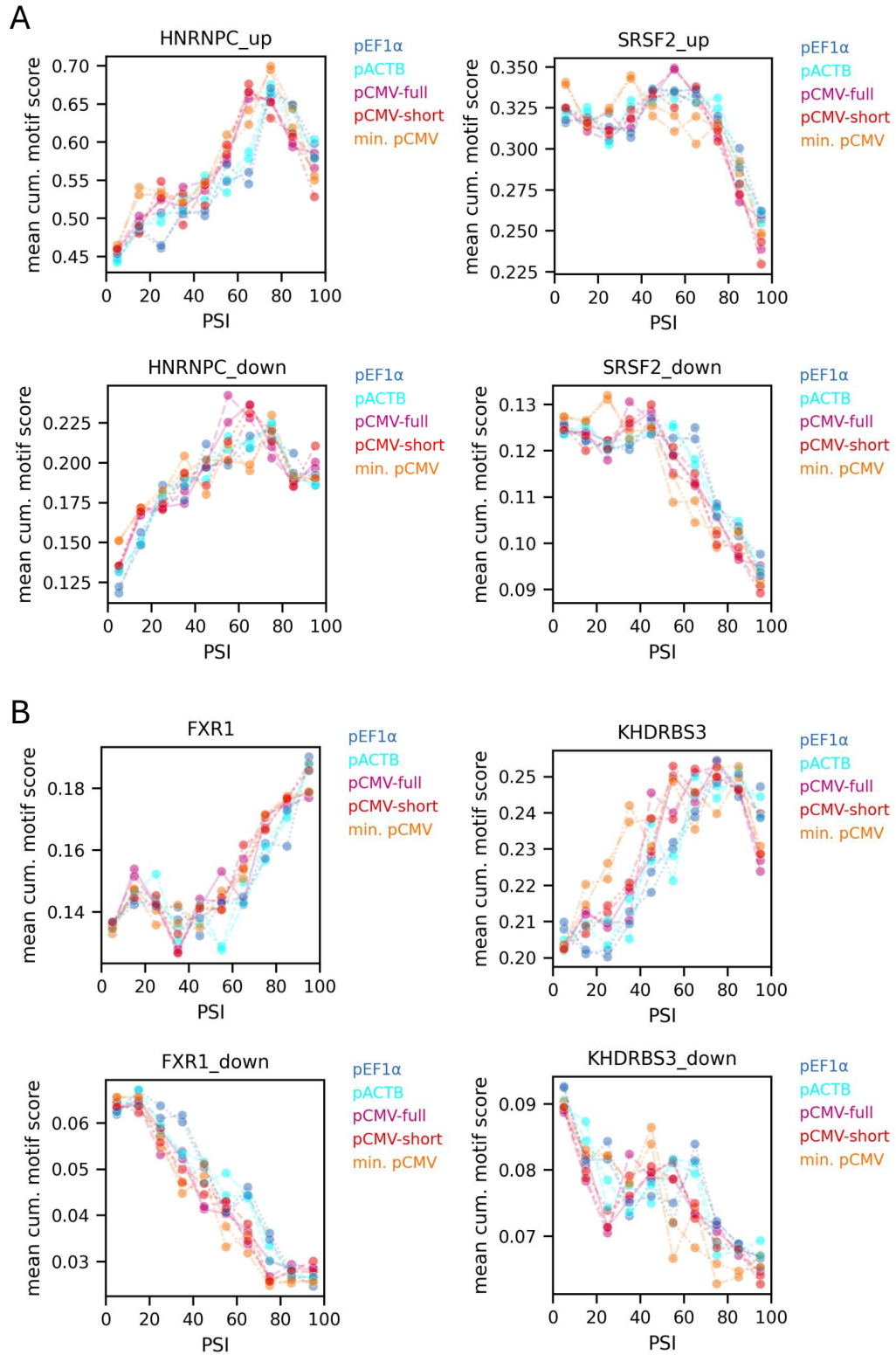
Supplementary Figure 9: Exons cluster based on their sensitivity to promoter identity.

A. Within-Cluster Sum of Squares (WCSS) for different numbers of clusters was used to choose a number of clusters for subsequent analyses (dashed vertical line). B. Clustered heatmap showing the PSI levels for all native cassette exon sequences (min. 10 reads per variant in all samples, min. 2 reads per isoform); clusters are indicated by colors on the left; z-scoring and clustering was done on the entire set of sequences (Fig. 4A); $n=150$. C. Clustered heatmap showing the PSI levels for all exon sequences based on pharmacological treatments and promoter variants (min. 10 reads per variant in all samples, min. 2 reads per isoform); clusters are indicated by colors on the left; $n=5653$. D. Representative examples for endogenous cassette exons from different clusters in Fig. 4A. Each data point corresponds to one measurement, each sequence was represented twice in the library (distinguishable by a unique barcode). E. Receiver Operating Characteristic (ROC) curves and the associated AUC values for the indicated classifications based on the subset of highly variable genes.



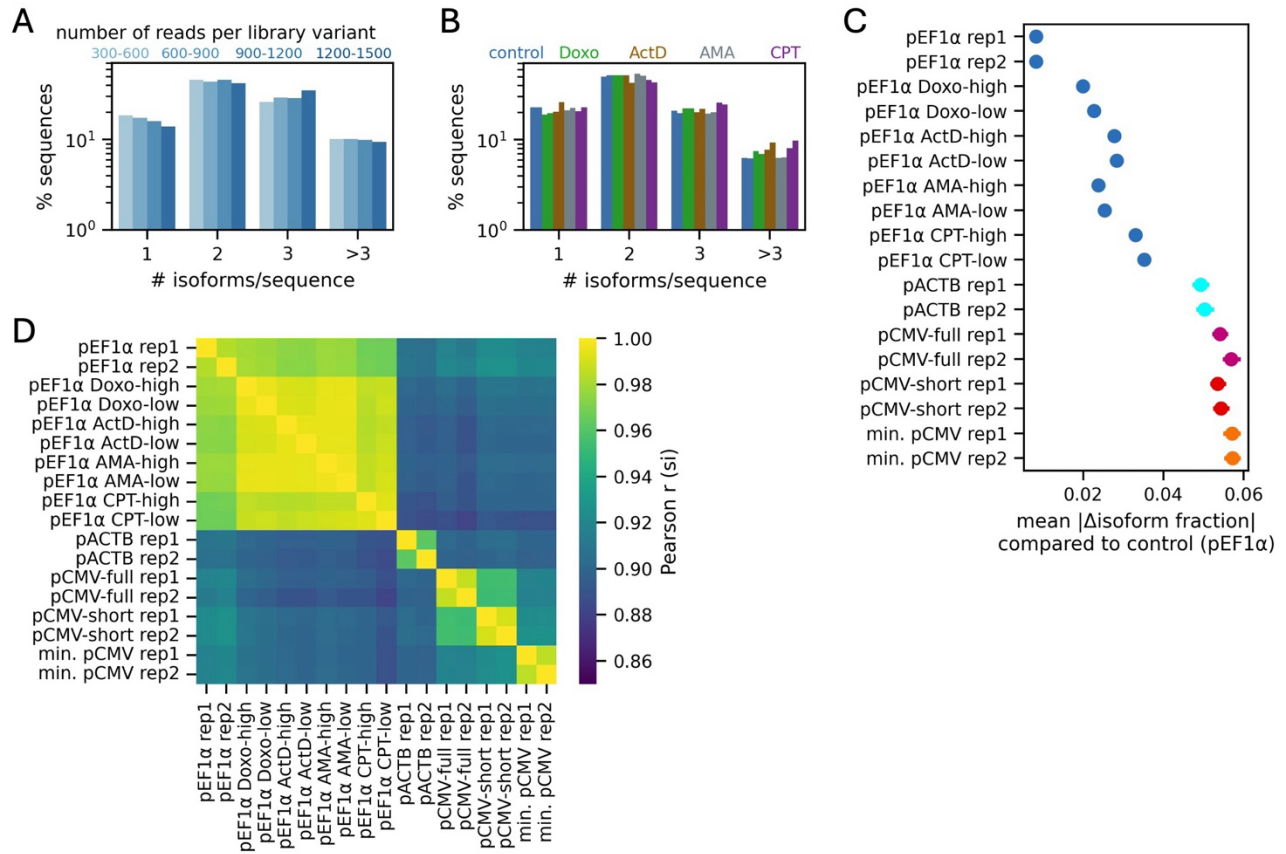
Supplementary Figure 10. Motif enrichment analysis between exon clusters.

A. Histogram showing the number of combinations of motif+location for which a similar sequence is present in the number of variant families indicated on the x-axis. B. Barplot showing the number of variant families (sequences derived from the same wild-type sequence) represented in each cluster. CD. Barplots showing for each cluster the number of motifs (C), dimers, trimers or tetramers (D) showing the lowest median p-value across the comparisons with all other clusters (Dunn's post test); absence of a cluster indicates that it did not have any motif fulfilling this criterion. E. Clustered heatmap showing the Spearman correlation coefficient between PSI measured for the indicated promoter context and all upstream motifs showing significant correlation (corr. $p < 0.05$, Bonferroni) in one of the conditions; asterisks indicate statistical significance (corr. $p < 0.05$, Bonferroni). F. The same analysis was performed as in Figure 5D, but restricting the library sequences used to highly variable (different from the corresponding wild-type sequence in at least 50/147 positions); the lowest p-value of pairwise comparisons of motif enrichment (F) is plotted against the standard deviation of Spearman correlation coefficients between promoter constructs; the number of motifs in the upper right quadrant (top 25% for each measure) out of all motifs included in the analysis for each location is shown.



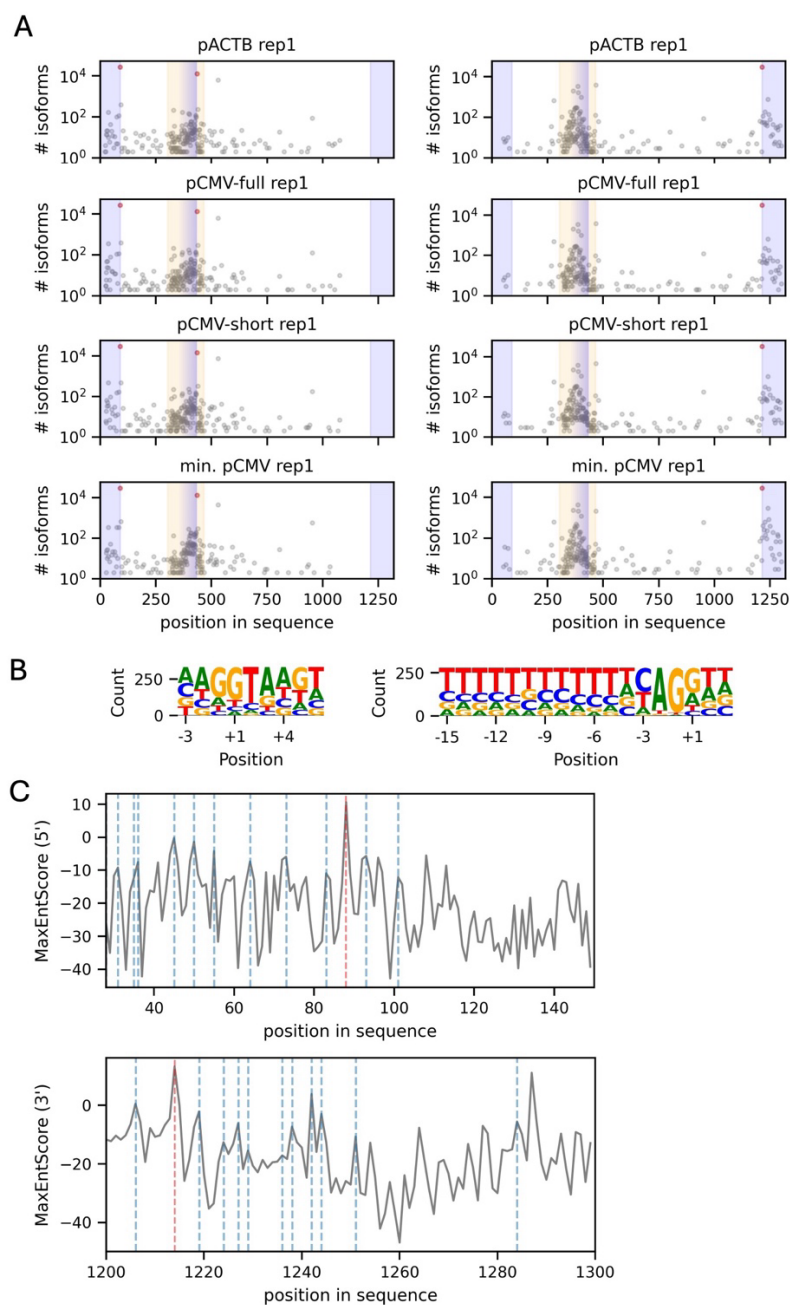
Supplementary Figure 11. Motif enrichment is associated with changes in exon inclusion rates.

AB. Mean cumulative motif score for the indicated RBP and location (up: upstream intron, down: downstream intron, only RBP name: exonic) in library sequences with a PSI in the indicated percentile range, each data points represents 1/10 of the span of PSI values (0-100); overall n=5734 for each sample.



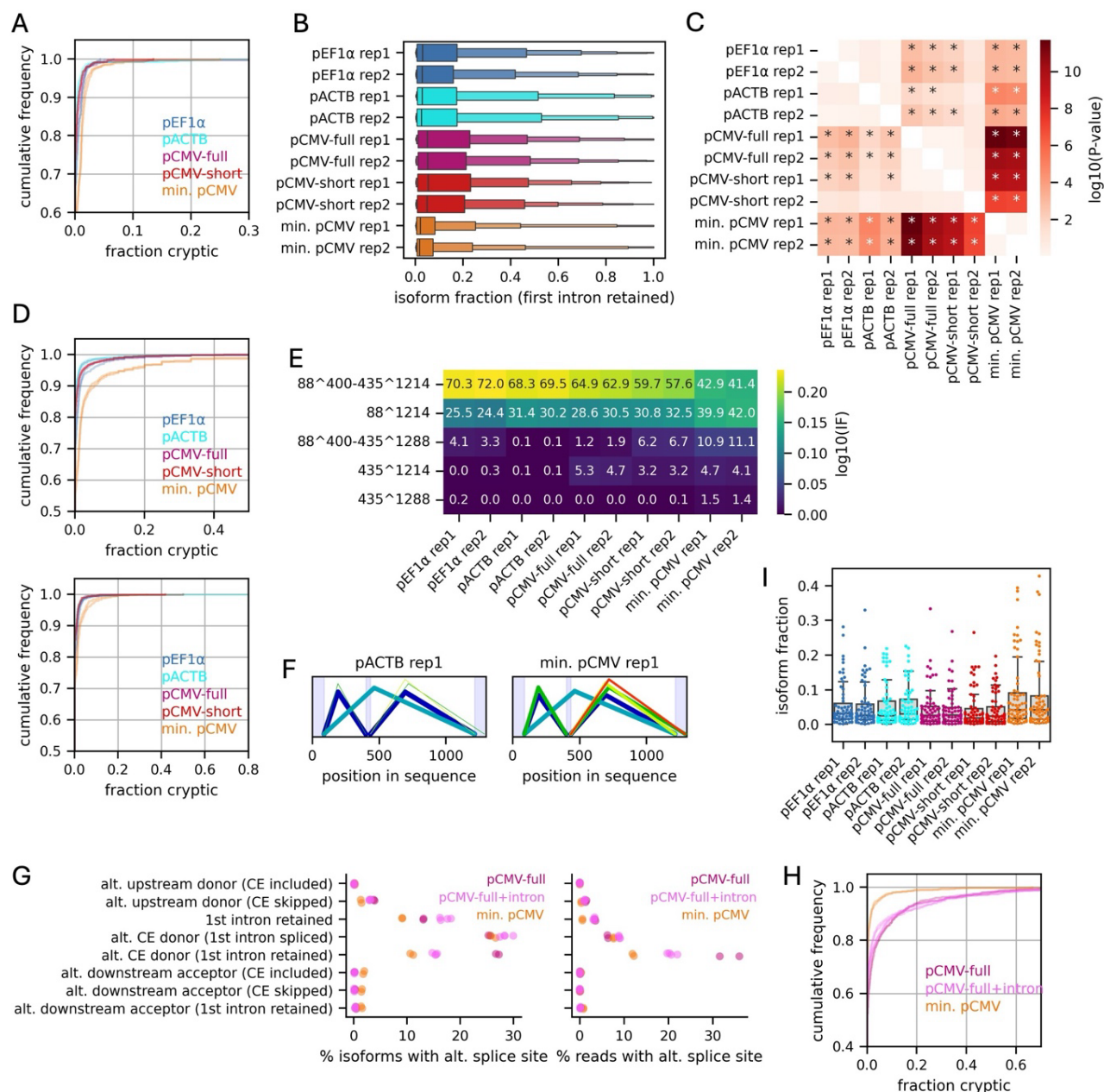
Supplementary Figure 12. Multiple isoforms are created through cryptic splice site usage.

A. Barplot showing the percent of library variants resulting in the indicated number of isoforms for the indicated numbers of reads per library variant (min. 2 reads per isoform, n=33-1133 per group). B. Barplot showing the percent of library sequences resulting in the indicated number of isoforms (min. 2 reads per isoform, 300-900 reads per library sequence); overall n=10261-12438 isoforms for each sample. C. Mean absolute changes in isoform fraction compared to the reference (mean of pEF1α replicates); errorbars denote 99% CI (n=41119). D. Heatmap of Pearson correlation coefficients between PSI values measured in all conditions (min. 2 reads per isoform per sample; n=41119 for each sample).



Supplementary Figure 13. Cryptic splice sites resemble donor and acceptor sequences.

A. Each data point indicates the number of library sequences with evidence for donor (left) or acceptor (right) splice sites (min. 2 reads per isoform per library sequence) at the position indicated on the x axis (relative to the end of the GFP coding sequence) in the sample indicated above the graph; red data points indicate the expected splice sites with fixed position according to the library design (see Fig. 1A). B. Sequence logos for cryptic donors (-3 to +6 relative to intron start, left) and acceptors (-15 to +3 relative to intron end, right). C. MaxEntScores at each position around the constant upstream donor (top) and downstream acceptor (bottom); red lines indicate the expected “canonical” splice sites (donor at 88 and acceptor at 1214 relative to the end of the GFP coding region), blue lines indicate the position of cryptic splice sites observed in the library.

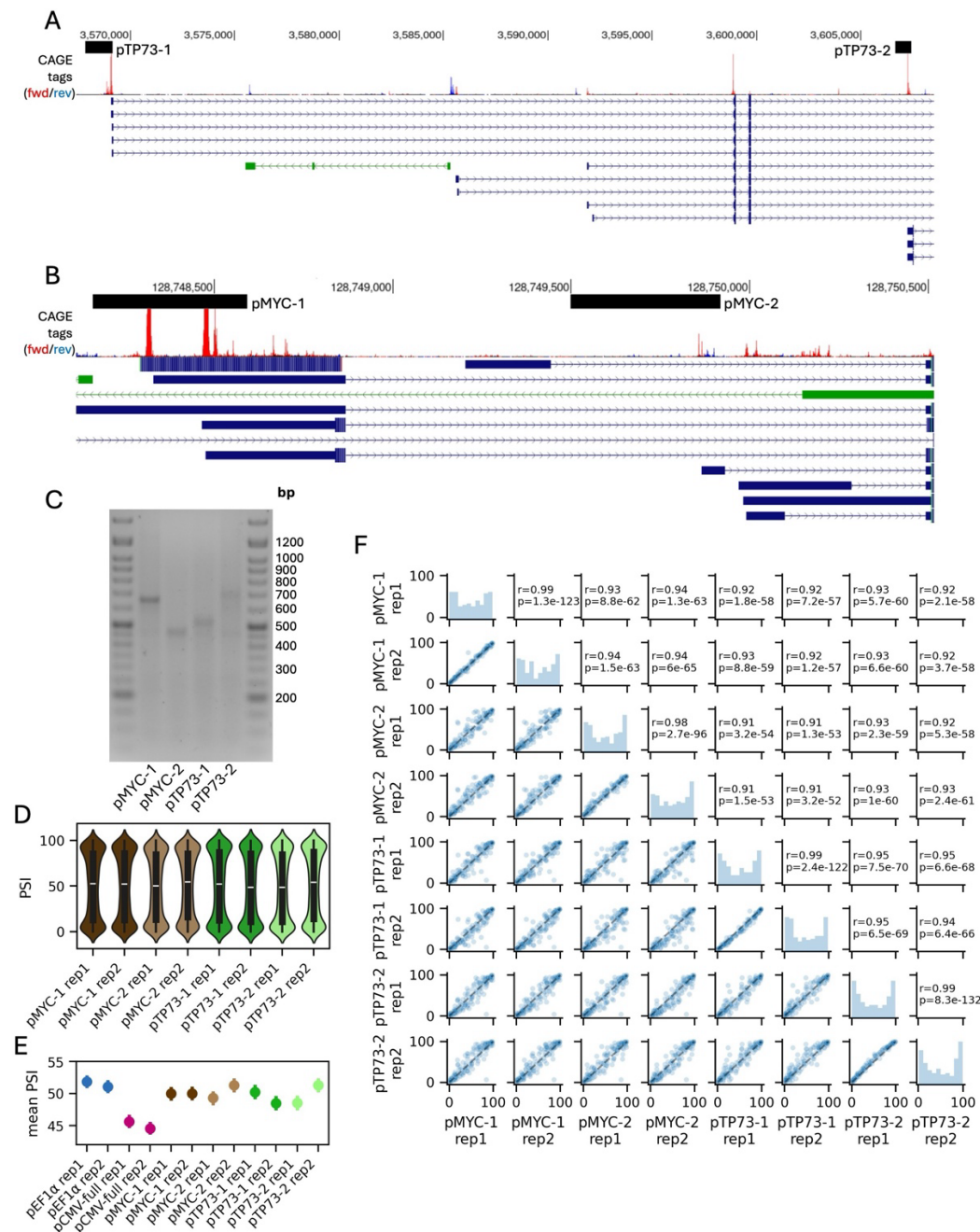


Supplementary Figure 14. Promoter identity affects cryptic splice site usage.

A. Cumulative distribution function showing the frequency of cryptic donor splice site use around the constant intron start site (position 88 in Fig. 1A) across 5653 variants if the cassette exon is included. B. Distribution of the fraction of isoforms with the upstream intron retained (n=5653 for each sample). C. Heatmap showing the corrected p-value (Mann-Whitney-U, Bonferroni) for all pairwise comparisons of the groups in panel B. D. Cumulative distribution function showing the frequency of cryptic acceptor splice site use around the constant downstream exon start site (position 1214 in Fig. 1A) across 5653 variants if the cassette exon is skipped (top) or included (bottom). E. Table showing relative isoform abundances in percent across samples for a representative library sequence with cryptic downstream acceptor usage (1288); colors represent $\log_{10}(\text{isoform fraction (IF)})$. F. Examples for cassette exon library variants showing different cryptic downstream acceptor usage (only isoforms covered by min. 2 reads, line width and hue indicate the isoform fraction, with thicker and darker lines corresponding to more abundant isoforms). G. Percent isoforms (left) and reads (right) with the properties indicated on the y-axis; each data point corresponds to one sample. H. Cumulative distribution function showing the frequency of cryptic donor splice site use around the constant intron start site (position 88 in Fig. 1A) across 5653 variants if the cassette exon is included. I. Isoform fraction across samples for a representative library sequence with cryptic downstream acceptor usage (1288).

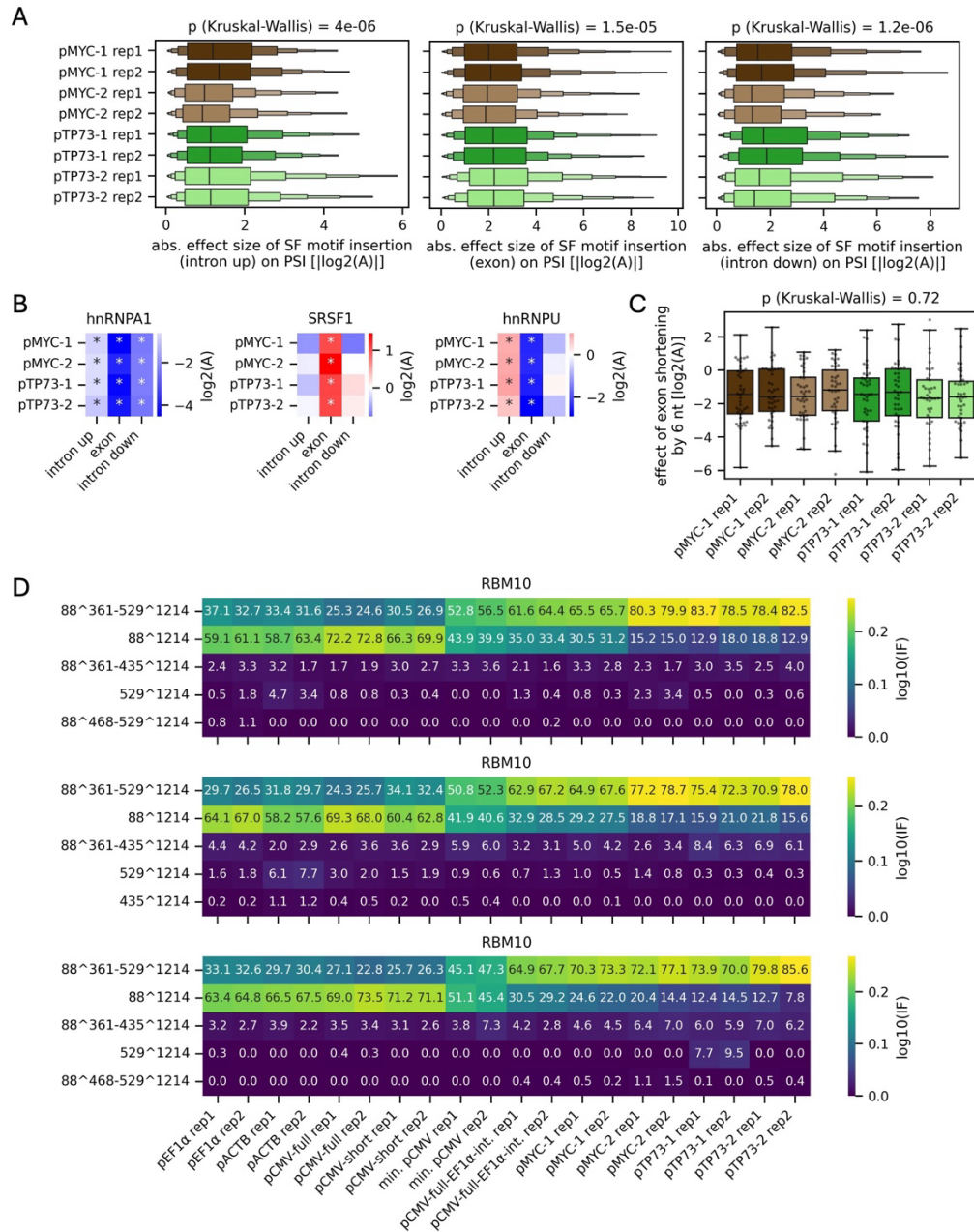
H. Cumulative distribution function showing the frequency of cryptic donor splice site use around the constant intron start site (position 88 in Fig. 1A) across 8263 variants if the cassette exon is skipped.

I. Fraction of isoforms with cryptic acceptor site usage at position 277 across samples; $n=63$, box plots show the median (central line) and interquartile range (box, 25th-75th percentiles), with whiskers extending to the most extreme data points within 1.5x of the interquartile range; outliers are shown as points.



Supplementary Figure 15. Characterization of endogenous promoter constructs.

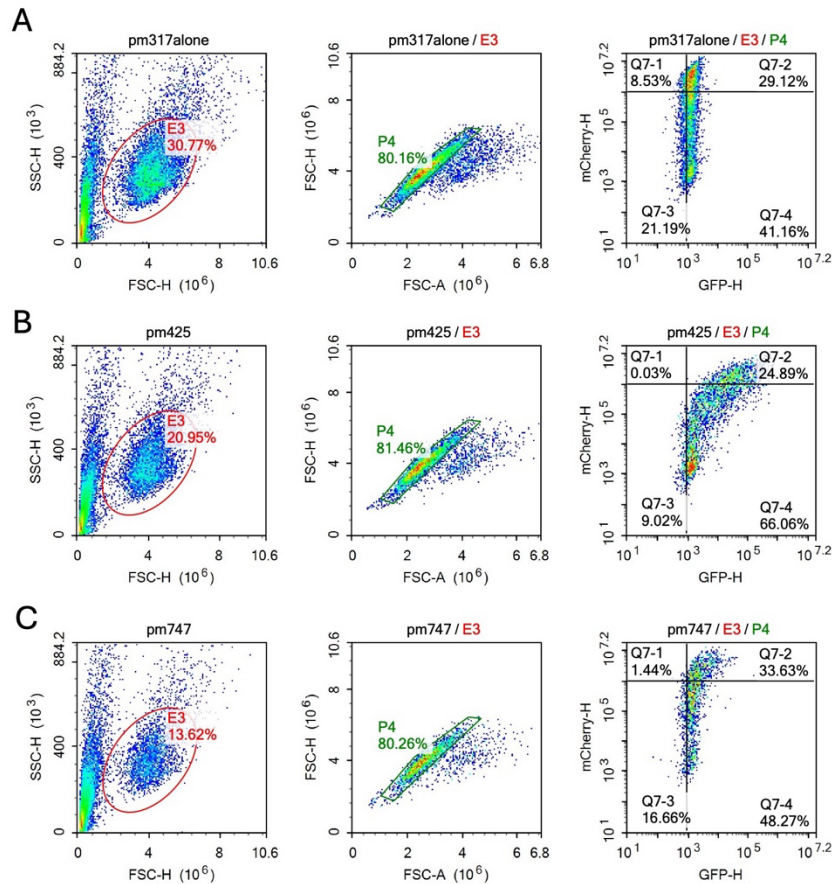
AB. UCSC genome browser visualization of parts of the TP73 (A) and MYC (B) gene, including the selected promoter regions in black as well as a track showing the density of CAGE tags and isoform models based on RefSeq (GRCh38; <http://genome.ucsc.edu>). C. Agarose gel showing the results of the 5'RACE on MYC and TP73 promoter variants. D. Violinplots showing the overall distribution of PSI values in the tested endogenous promoter constructs (n=6562-7388); inner box plots show the median (central line) and interquartile range (box, 25th-75th percentiles), with whiskers extending to the most extreme data points within 1.5x of the interquartile range. E. Mean PSI across library variants measured in the endogenous promoter constructs; errorbars denote 99% CI (n=6562-7388). F. Scatterplots comparing PSI values for reporters containing native cassette exon sequences between different experimental conditions; Pearson correlation coefficient and the associated p-values are shown in the upper right half; the graphs along the diagonal show the distribution of PSI values in each sample (n=137).



Supplementary Figure 16. The influence of endogenous promoters on splicing factor motif effects and splice site choice.

A. Boxenplots showing the distribution of effect sizes of motif insertions in the exon in different endogenous promoter constructs (n=667, 667, 619, 617, 650, 643, 660, 657 in the order they appear in the graphs), based on insertions of 16 motifs in the upstream intron (left), the exon (middle) or the downstream intron (right) in 1-4 different locations in 13 endogenous sequence contexts). B. Heatmap showing the effect ($\log_2(A)$) of the indicated motif in or around the cassette exon; asterisks indicate a statistically significant effect ($p < 0.05$, Wilcoxon signed-rank test; $n = 19-48$ variants). C. Effect size ($\log_2(A)$) of exon shortening by 6 nt (minimal resulting exon length: 20; $n = 50$); box plots show the median (central line) and interquartile range (box, 25th-75th percentiles), with whiskers extending to the most extreme data points within 1.5x of the interquartile range; outliers are shown

as points. D. Tables showing relative isoform abundances in percent across samples for three barcode variants of the same library sequence containing the RBM10 wild-type cassette exon sequence; colors represent $\log_{10}(\text{isoform fraction (IF)})$ of each variant.



Supplementary Figure 17. Outline of the gating strategy used in the flow cytometry experiments based on three representative examples.

Flow cytometry gating for a negative control (A; corresponding to “no GFP” in Figure 1D), a strong promoter (B; corresponding to “pEF1 α ” in Figure 1D) and a weak promoter (C; corresponding to “pTP73-2” in Figure 7B) driving GFP and co-transfected with an mCherry expressing plasmid based on forward and side scatter (left) to select cells, based on forward scatter height and area to select single cells (middle) and based on GFP and mCherry fluorescence intensity (right) to select a population for analysis (gate Q7-2). A high mCherry threshold (10^6) was used to ensure to provide equal conditions for all promoters, independent of the varying strength and correspondingly varying GFP expression levels.