Supplementary Information for:

U-rich elements drive pervasive cryptic splicing in 3' UTR massively parallel reporter assays

Khoa Dao[1], Courtney F. Jungers[3], Sergej Djuranovic[3,4], Anthony M. Mustoe[1,2,*]

[1]Therapeutic Innovation Center (THINC), Verna and Marrs McLean Department of Biochemistry and Molecular Pharmacology, Baylor College of Medicine, Houston TX
[2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston TX
[3]Department of Cell Biology and Physiology, Washington University School of Medicine in St. Louis, St. Louis MO
[4]Department of Molecular Biology, Cellular Biology and Biochemistry, Brown University, Providence, RI, USA

*correspondence: anthony.mustoe@bcm.edu

Figure S1

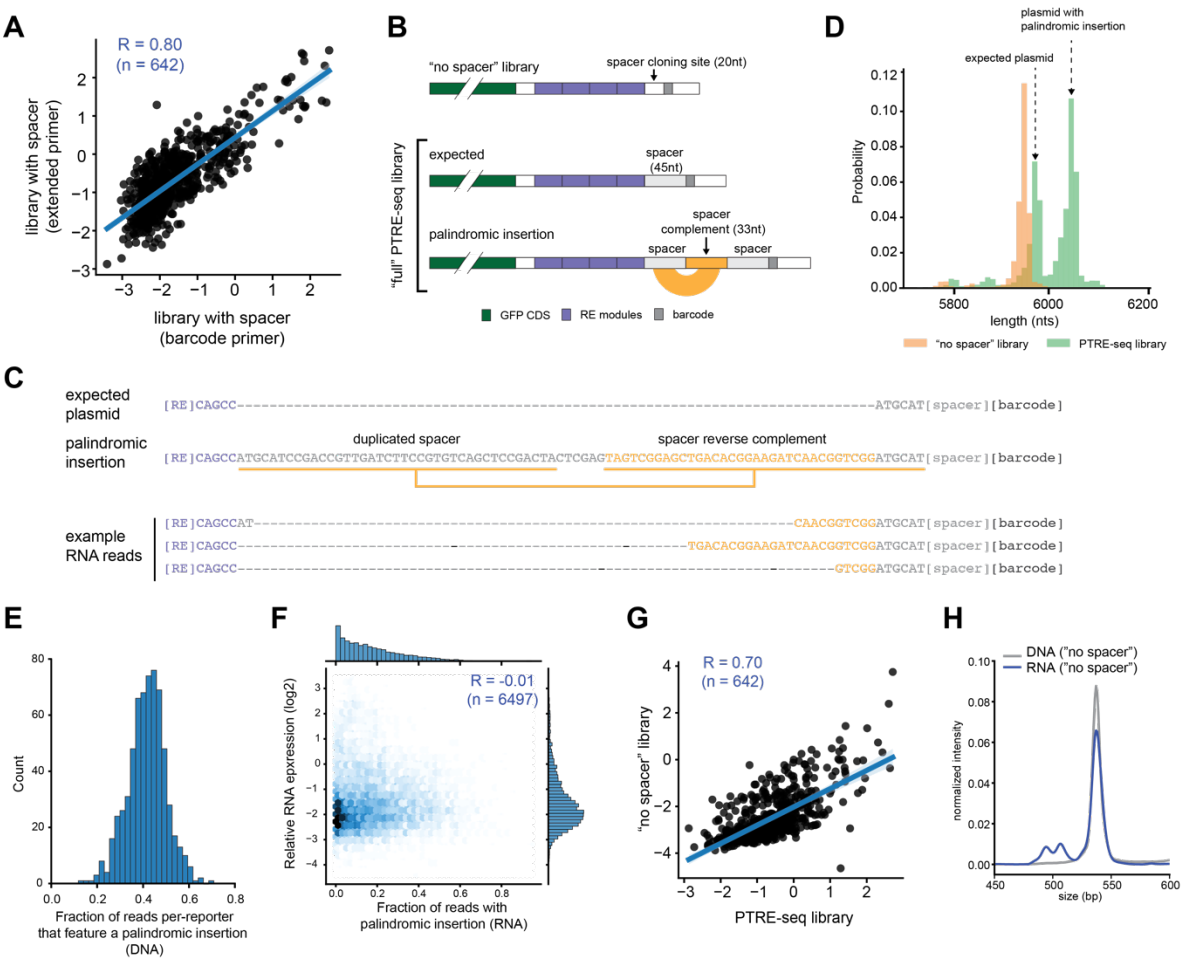

Figure S1. Detection of palindromic insertion artifact in PTRE-seq plasmids. (A) Comparison of RNA expression measurements obtained from extended primers (this study) compared to barcode primers (ref. 24). (B) Schematic of different PTRE-seq plasmids used in the study. The full PTRE-seq library was generated from the "no spacer" library by cloning the spacer sequence into the "spacer cloning site". This cloning step evidently produced two species in the full PTRE-seq library: the expected and palindromic insertion species. (C) Example DNA and RNA sequencing reads illustrating the palindromic insertion artifact. DNA reads are from full-length plasmid sequencing and RNA reads are from Illumina amplicon sequencing. (D) Distribution of read lengths from nanopore sequencing of the linearized "no spacer" and full PTRE-seq plasmid libraries. (E) The fraction of DNA sequencing reads for each reporter identity (regulatory element combination) that feature a palindromic insertion in the full PTRE-seq library. Fractions were computed from nanopore sequencing of the linearized plasmid library. (F) Relationship between relative RNA expression and frequency of palindromic insertion, computed at the barcode level from amplicon RNA sequencing data. (G) Comparison of RNA expression measurements from the full PTRE-seq library and the "no spacer" library. (H) Size heterogeneity is observed in the RNA "no spacer" library. Electropherogram was measured by TapeStation 4200. In (A, G), line of best fit and Pearson correlation coefficient is shown in blue.

Figure S2

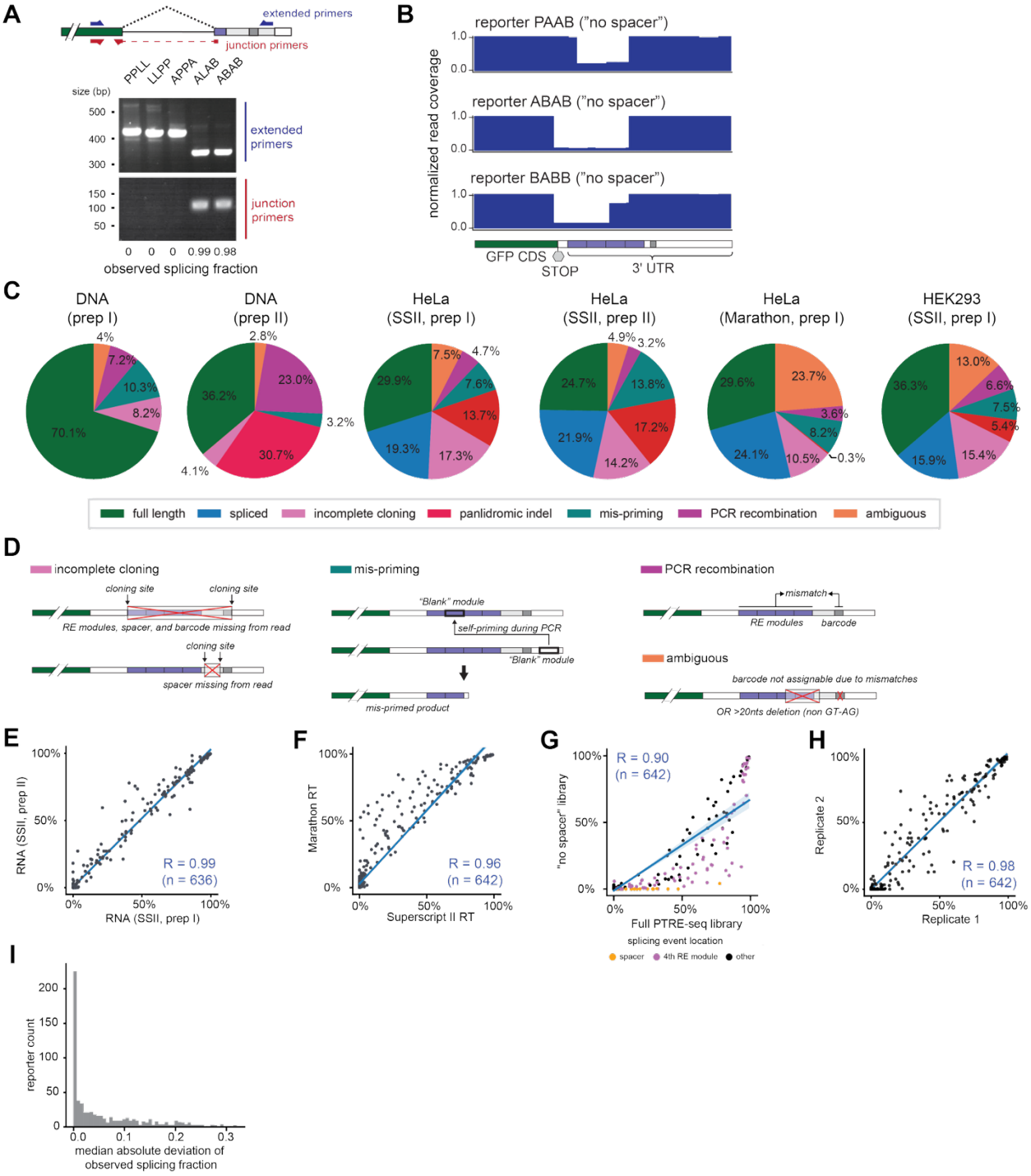

Figure S2. Quality control and quantification of splicing and other anomalies in the PTRE-seq library. (A) RT-PCR validation of splicing on individually transfected reporters. Reporters were transfected into HeLa cells and RT-PCR products were resolved on agarose gels using extended and junction primers. Splicing efficiency measured from sequencing of the PTRE-seq MPRA is shown at bottom. (B) Representative read coverage tracts of reporters with internal deletions in the "no spacer" library. (C) Quantification of read categories from sequencing of the

PTRE-seq libraries prepared from DNA plasmids and RNA reverse-transcribed with SuperScript II (SSII) or MarathonRT. “Prep I” and “prep II” refer to library preparation strategies based on emulsion PCR and standard PCR respectively (Methods). Labels show the mean percentage for each category over 2-3 biological replicates. HeLa (SSII, prep I) is replicated from Fig. 1E to facilitate comparison. HEK293 (SSII, prep I) is representative of quantifications from HEK, SH-SY5Y, and U87 cell lines. (D) Illustration of read anomalies detected in PTRE-seq sequencing libraries. (E) Comparison of splicing fraction measurements obtained from RNA libraries prepared with “prep I” and “prep II” protocol. (F) Comparison of splicing fraction measurements obtained from libraries prepared using Superscript II versus MarathonRT. (G) Comparison of observed splicing fractions obtained from the PTRE-seq library and the “no spacer” library. The majority of outliers correspond to splicing events occurring in the spacer and the 4^{th} RE module, which is immediately upstream of the spacer (which is missing in the “no spacer” library). (H) Comparison of splicing fractions measured across biological replicates in HeLa cells (SSII). (I) Median absolute deviation of observed splicing fractions computed across alternative barcodes for each 3’ UTR. In (E-H), line of best fits and Pearson correlation coefficients is shown in blue. Translucent bands show 95% confidence intervals. Uncropped gels are provided in the Source Data file.

Figure S3

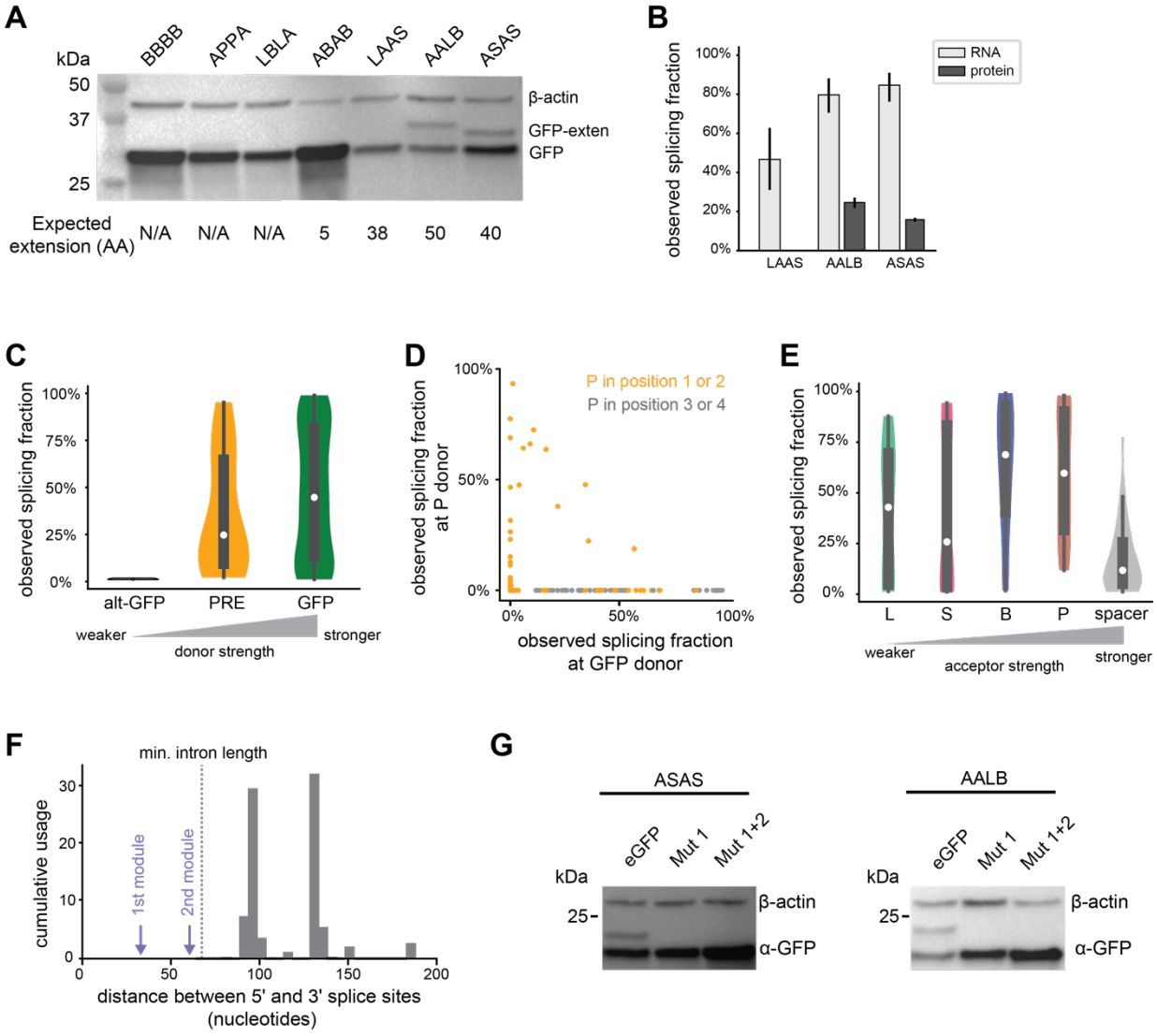

Figure S3. Validation and analysis of cryptic splicing donor and acceptor sites in PTRE-seq library. (A) Western blot analysis of GFP products from transfections of individual reporters. β-actin is shown as loading control. Predicted amino acid (AA) extension based on distance to the nearest in-frame stop codon is shown at bottom. (B) Comparison of observed splicing fractions measured by MPRA sequencing and western blot from spliced reporters in (A). The ABAB reporter was excluded from this analysis because its spliced versus full-length isoform cannot be distinguished via western blot. Error bars show standard deviation from 3 biological replicates. (C) Distribution of observed splicing fractions exhibited by different PTRE-seq splice donors. (D) Scatterplot of P donor versus GFP donor usage in PRE-containing reporters. (E) Distribution of observed splicing fractions observed at different splice acceptors in the PTRE-seq library. (F) Distributions of intron lengths observed in the PTRE-seq library, weighted by cumulative usage. Dashed line indicates minimum intron length identified from previous studies[29]. Purple arrows indicate intron lengths if 1st or 2nd module sites were used as splice acceptors. (G) Western blot analysis of ASAS and AALB reporters where GFP was synonymously recoded to abolish the 5'

cryptic splice site. Matching RT-PCR data are shown in Figure 2F. For (C) and (E), violins illustrate the probability density, boxplots denote quartiles, white dots represent the medians, and whiskers represent the min and max for each distribution. Uncropped blots are provided in the Source Data file.

Figure S4

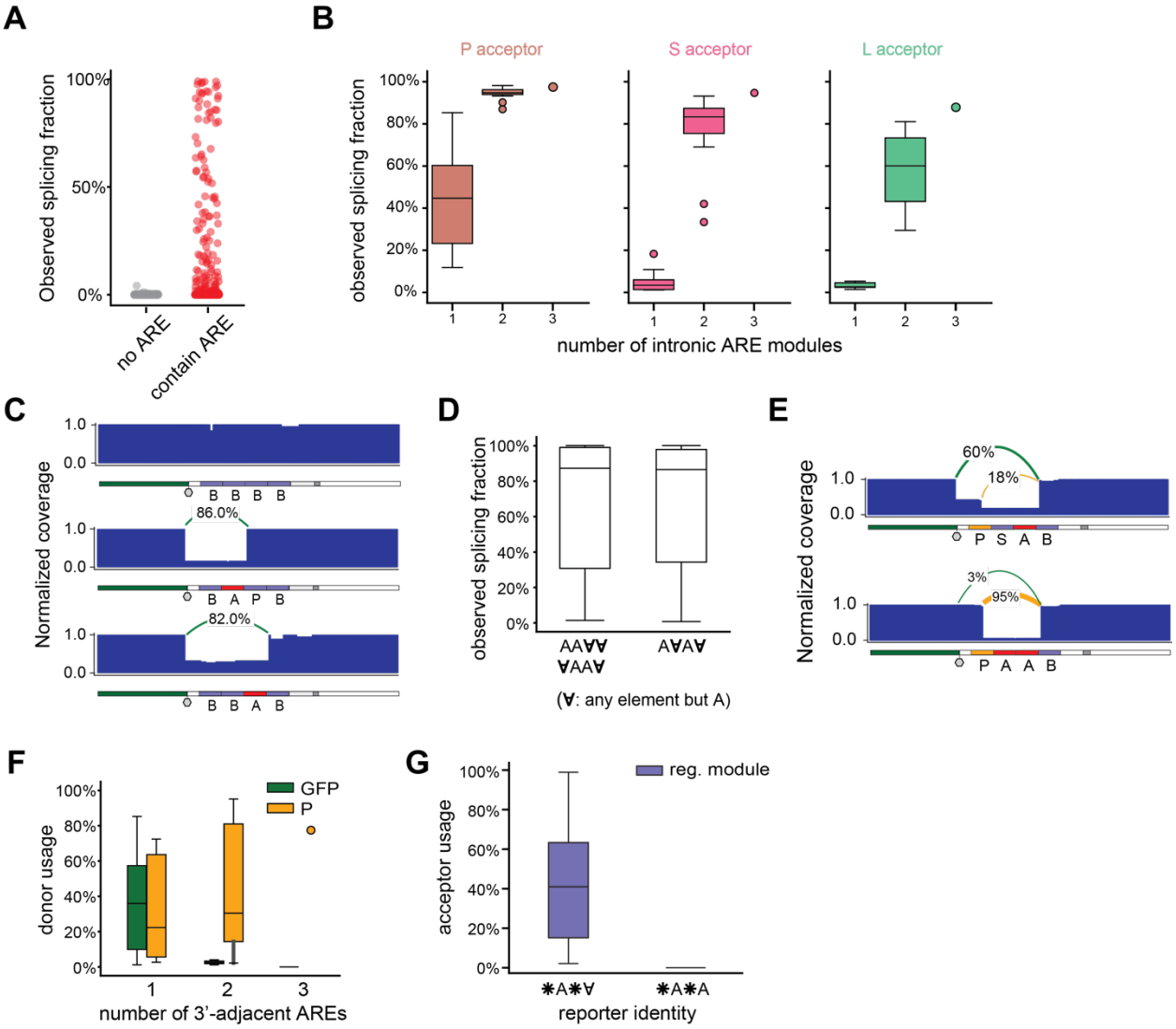

Figure S4. AREs are position-dependent regulators of cryptic splicing. (A) Observed splicing fractions of ARE-containing and non-ARE containing transcripts in the "no spacer" PTRE-seq library. (B) Relationship between the number of intronic AREs and observed splicing fractions for P (n = 45), S (n = 26), and L (n = 20) splice acceptors. (C) Example read coverage tracts showing a single ARE module is sufficient to activate highly efficient splicing at the B acceptor. (D) Comparison of observed splicing fractions between reporters containing two AREs that are contiguous versus non-contiguous. (E) Example read coverage tracts illustrating ability of adjacent AREs to bias selection of P versus GFP donors. (F) Comparison of splice donor usage in reporters with increasing numbers of ARE modules downstream of the P donor. (G) Comparison of acceptor usage in "no spacer" reporters with and without AREs in the 4th regulatory module. ✱ symbol represents all regulatory elements, ∀ symbol represents all regulatory elements except ARE. For (A, C, F, G), box plots span the 25th and 75th percentile and the centers indicate the median. Whiskers indicate the furthest datum 1.5x outside the interquartile range.

Figure S5

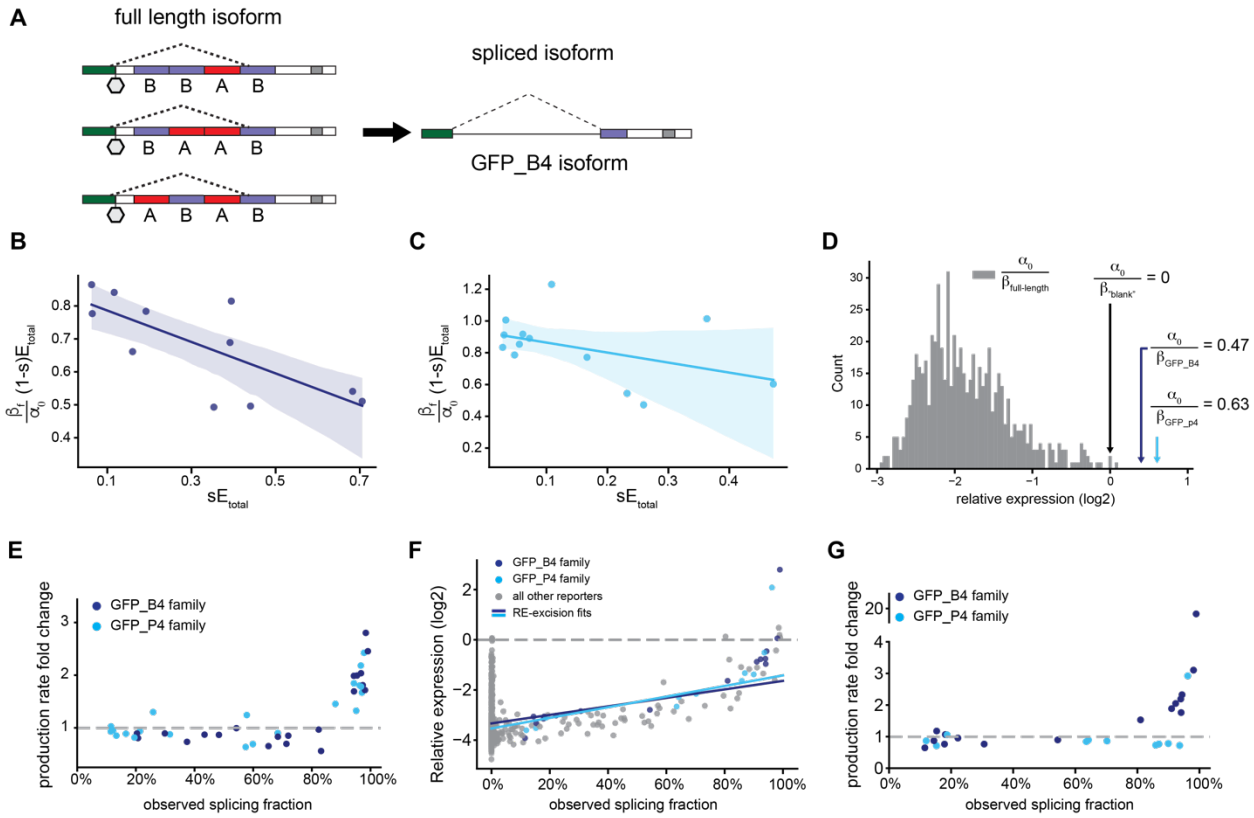

Figure S5. Modeling the impact of splicing on mRNA expression using RE-excision and production enhancement models. (A) Schematic showing the GFP_B4 family of reporters with different pre-mRNA sequences that are spliced at the same donor and acceptor sites, resulting in the same spliced isoform. (B) Fit of Equation 5 to the GFP_B4 family of reporters using reporters spliced from 10% to 80% efficiencies. The $\frac{\beta_S}{\alpha_0}$ of the GFP_B4 isoform is obtained from the slope of the linear regression model. Translucent band shows the 95% confidence interval for the regression line. (C) same as (B), but for the GFP_P4 family of reporters. (D) Estimated stability for the spliced isoforms GFP_B4 and GFP_P4 compared to stability coefficients of full-length reporters. $\frac{\beta_S}{\alpha_0}$ for GFP_B4 and GFP_P4 were computed using Equation 5 using reporters spliced from 10% to 80% efficiencies. Stabilities of full-length reporters were estimated using the regression model described in Equation 1. (E) The production enhancement effect is robust to alternative estimates of GFP_B4 and GFP_P4 spliced isoform stabilities using reporters spliced between 10% to 70% efficiency. (F) Relationship between observed splicing fraction and reporter expression in the no-spacer PTRE-seq library. Regression lines of respective colors show the fits of RE-excision model to GFP_B4 and GFP_P4 reporters. (G) Estimated production rate for GFP_B4 and GFP_P4 reporters in the no spacer library as a function of observed splicing fraction. For (F) and (G), $\frac{\beta_S}{\alpha_0}$ was computed using Equation 5 using reporters spliced from 10% to 80% efficiency.

Figure S6

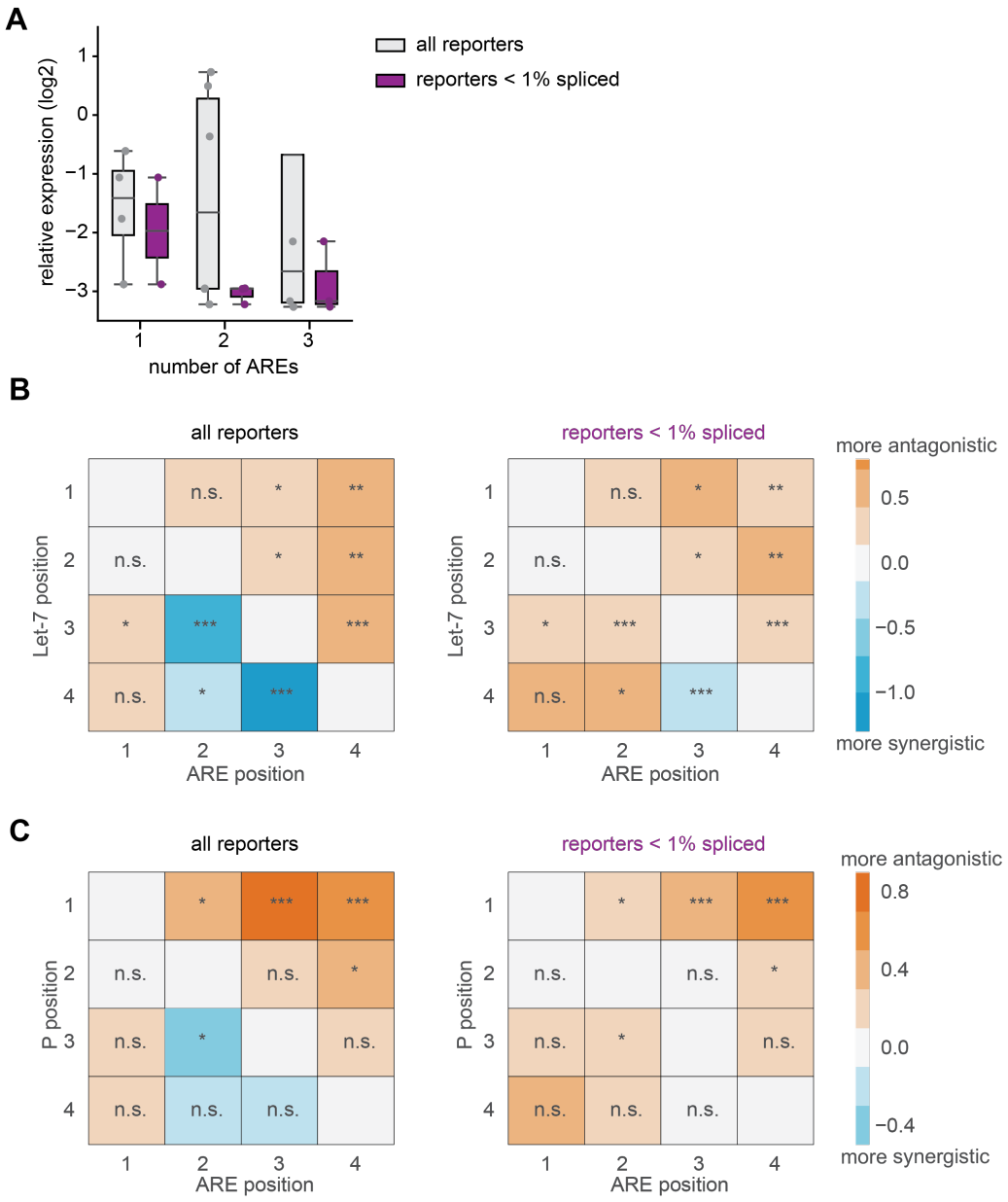

Figure S6. Impact of AREs on neighboring Let-7 and P sites. (A) . Effect of ARE module copy number on RNA steady state expression in the "no spacer" PTRE-seq library. Box plots span the 25th and 75th percentile and the centers indicate the median. Whiskers indicate the furthest datum 1.5x outside the interquartile range. Relative expression measurements are from the palindrome-free "no spacer" library. (B, C) Interaction coefficients between (B) ARE and Let-7 modules and (C) ARE and P modules obtained from fitting a linear regression model (Equation 1) to all reporters or only to unspliced reporters (<1% spliced). In the original PTRE-seq publication (ref. 24), AREs were concluded to have both synergistic and antagonistic relationship with other regulatory modules depending on their positions. However, when spliced reporters are excluded, the regression model indicates that AREs have minimal or moderately antagonistic impact on repression by neighboring Let-7 and P sites. Statistical significance was assessed using the Wald test, where *=P < 0.05, **=P < 0.01, ***=P < 0.001, and n.s.=not significant.

Figure S7

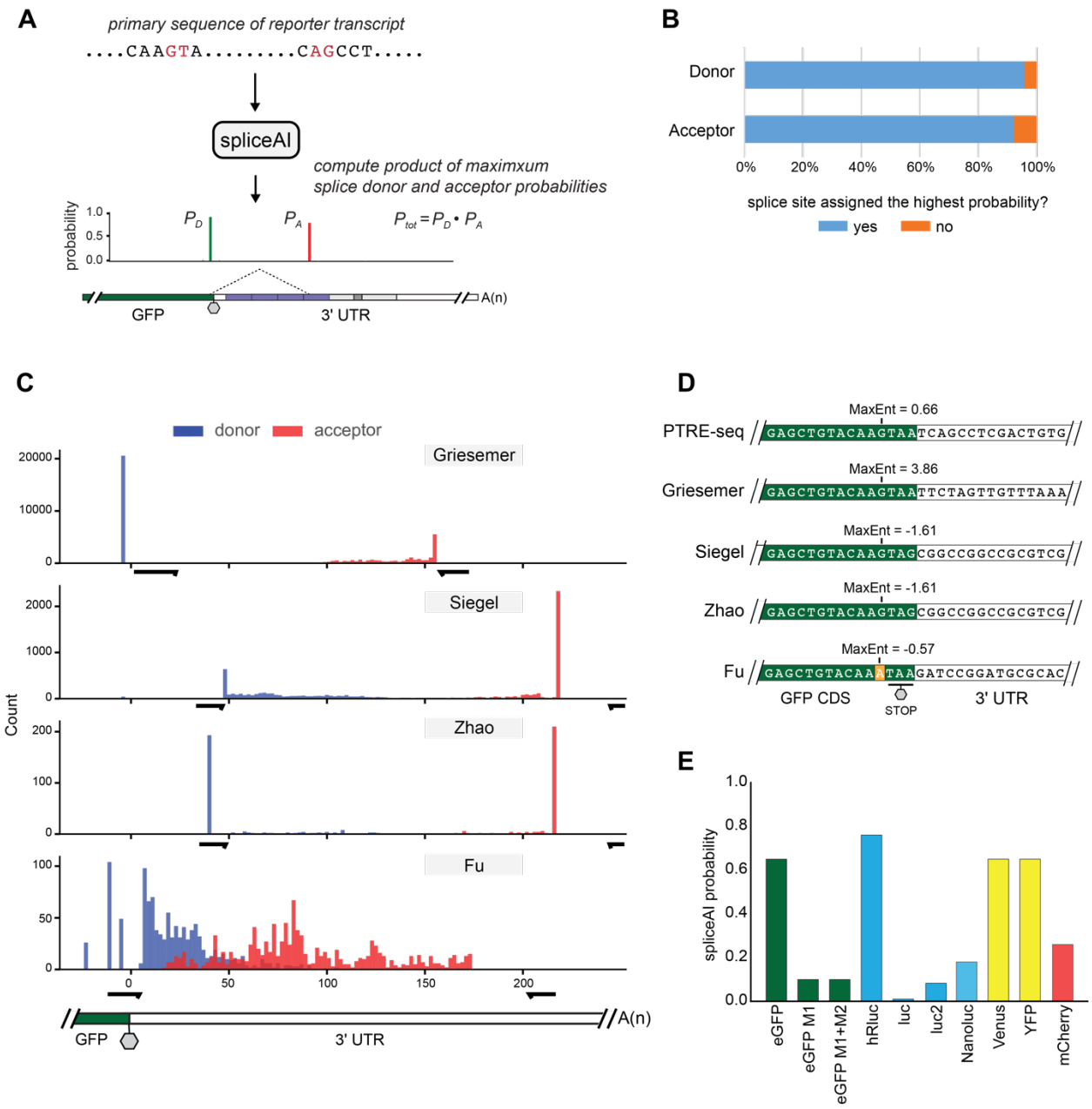

Figure S7. Prediction of cryptic splicing in MPRAs using SpliceAI. (A) Strategy for predicting cryptic splicing in reporter transcripts using spliceAI. The transcript sequence of each reporter is input into spliceAI[36], which computes the probability of every nucleotide being a splice donor or acceptor. To compute the overall probability of splicing, we compute the product of the highest probability donors and acceptors, respectively. (B) Percentage of splice sites in PTRE-seq with usage >1% correctly assigned as the highest probability sites by spliceAI. (C) Relative location of predicted 5' and 3' splice sites in *Griesemer*[7], *Siegel*[8], *Zhao*[9], and Fu[10] MPRAs. Only splice sites with probability above 0.01 are included. The forward and reverse primer binding sites used in each study are indicated by arrows. (D) Sequence context differences alter the predicted strength of the GFP cryptic donor site in the reporter vectors used for PTRE-seq, *Griesemer*, *Siegel*, and

Zhao. GFP donor strength is scored using Maximum Entropy[52]. (E) Predicted splicing probability for other commonly used reporter genes. The coding sequence for each gene was appended to the 3' UTR of reporter ABAB, which contains a strong splice acceptor.

Figure S8

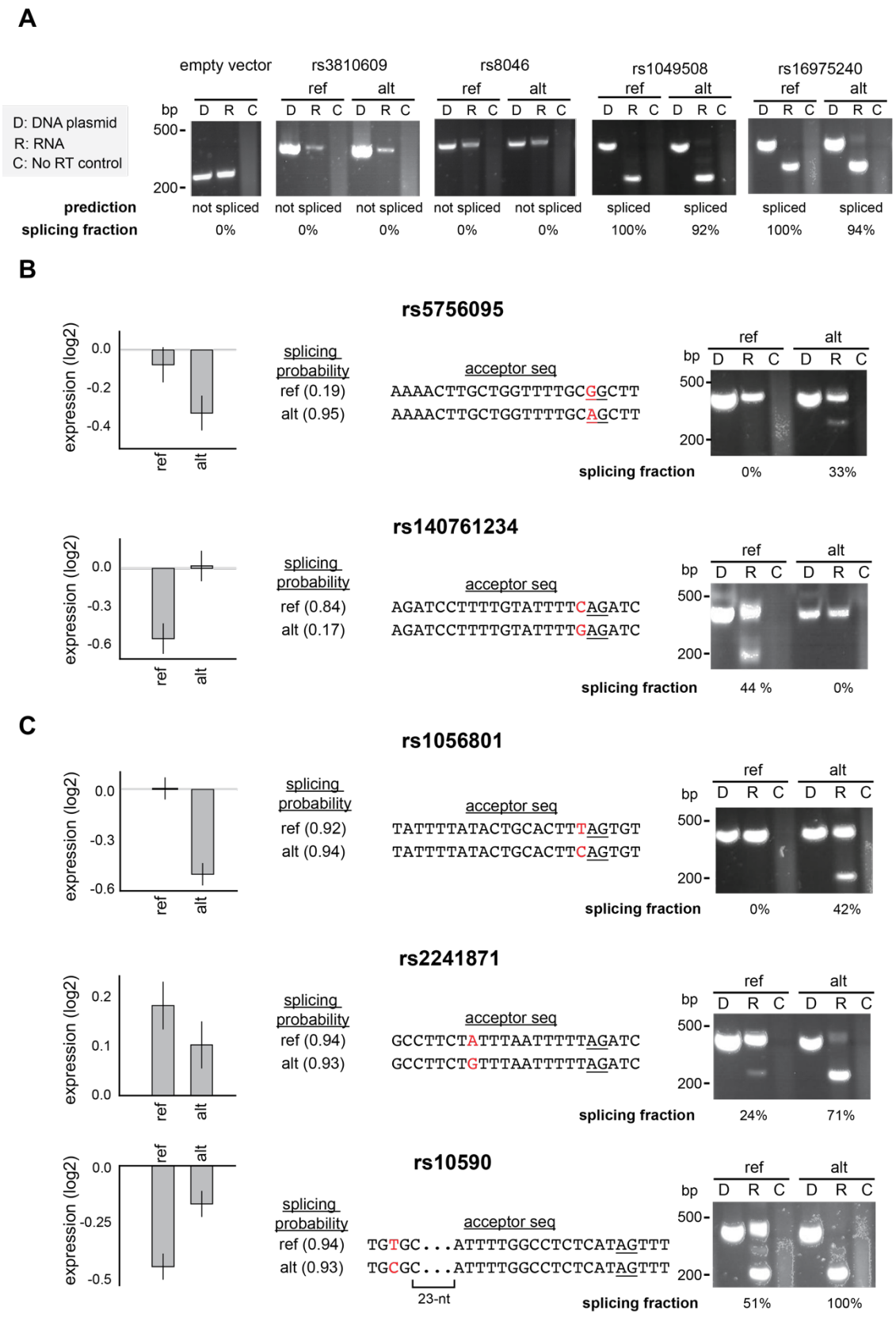
A
empty vector
rs3810609
rs8046
rs1049508
rs16975240
ref
alt
D R C
bp
500
200
D: DNA plasmid
R: RNA
C: No RT control
prediction
not spliced
spliced
splicing fraction
0%
100%
92%
94%
B
rs5756095
expression (log2)
splicing probability
ref (0.19)
alt (0.95)
acceptor seq
AAAACTTGCTGGTTTTGCGGCTT
AAAACTTGCTGGTTTTGCAGCTT
splicing fraction
0%
33%
rs140761234
ref (0.84)
alt (0.17)
AGATCCTTTTGTATTTTCAGATC
AGATCCTTTTGTATTTTGAGATC
44 %
C
rs1056801
ref (0.92)
alt (0.94)
TATTTTATACTGCACTTTAGTGT
TATTTTATACTGCACTTCAGTGT
42%
rs2241871
ref (0.94)
alt (0.93)
GCCTTCTATTTAATTTTTAGATC
GCCTTCTGTTTAATTTTTAGATC
24%
71%
rs10590
TGTGC...ATTTTGGCCTCTCATAGTTT
TGCGC...ATTTTGGCCTCTCATAGTTT
23-nt
51%
100%

Figure S8. Validation of cryptic splicing in *Griesemer* MPRA. Selected reporters from the *Griesemer* MPRA[7] were synthesized and validated with RT-PCR analysis. (A) Agarose gels of plasmid backbone, 2 tamVar pairs predicted to both be unspliced, and two tamVar pairs predicted to both be highly spliced. (B) Analysis of tamVar pairs predicted to be differentially spliced. See also Figure 5I. (C) Analysis of tamVar pairs in which both alleles were predicted to be strongly spliced, but RT-PCR indicated differential splicing. For (B) and (C), left bar plots show mean relative expression measured by *Griesemer*. Error bars denote reported standard error. SpliceAI probability for each reporter and sequence of predicted 3' splice sites are shown in the middle. Variant is highlighted red and the splice acceptor is underlined. At right is shown RT-PCR analysis of individually transfected plasmids resolved by agarose gel. Splicing fraction was quantified using ImageJ. D: DNA plasmid, R: RT-PCR of RNA, C: no RT control. Uncropped gels are provided in the Source Data file.

Table S1: Primers used in this study

Name	Sequence
RE_Amp_RT	AAAGGACAGTGGGAGTGGC
GFP_Amp_F_seq	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNGCTGCCCGACAACCAC
RE_Amp_R_seq	CCCTACACGACGCTCTTCCGATCTNNNNNAAAGGACAGTGGGAGTGGC
Junction_R	CACAGTCGAGGTTGTACAGCTC
M1_For	GAGCTATACAAATAATCAGCCTCGACTGTGGC
M1_Rev	GCCACAGTCGAGGCTGATTATTTGTATAGCTC
M2_For	GAGCTATACCAAATAATCAGCCTCGACTGTGGC
M2_Rev	GCCACAGTCGCCGGCCGCTACTTGTATAGCTC
Griesemer_Amp_F	CTAGTTGTTTAAAGCCCAACGCTAG
Griesemer_Amp_R	CGACGACGCTCTTCCGATC
Griesemer_F	GTTTAAAGCCCAACGCTAGTC
Griesemer_RT	CAGCGGTGGCAGCAG
Griesemer_F_seq	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNGTTTAAAGCCCAACGCTAGTC
Griesemer_R_seq	CCCTACACGACGCTCTTCCGATCTNNNNNCAGCGGTGGCAGCAG

Table S2: Evaluated sequences from the *Griesemer*[7] MPRA

Sequences were synthesized as IDT eBlocks, which have a minimum length requirement of 300 nts. To meet this minimum length requirement, the following sequence was appended 3' to all sequences during synthesis:

TCGGATCCCGAAGAACTGCGCAGTCCGGTTATAACGTCTGACATGTCCTTTTTTGATTTGGAATCCAACC
ACTCAAGTGGATCCGTTCATTTACTTTGCGAAAAATATACCCACAAATTGCCCATCGAAAGTGA

This flanking sequence was then removed by the initial PCR amplification step (Methods).

Name	Sequence
rs8046_ref	CTAGTTGTTTAAAGCCCAACGCTAGTCACTTTCCGAGCTCGCTAGCCTCCAACAAGAATG CATTCCCTGAATCTGTGCCTGCACTGAGAGGGCAAGGAAGTGGGGTGTTCTTCTTGGGAC CCCCACTAAGACCCTGGTCTGAGGATGTAAGATCGGAAGAGCGTCG
rs8046_alt	CTAGTTGTTTAAAGCCCAACGCTAGTCGAAAGTCGAGCTCGCTAGCCTCCAACAAGAATG CATTCCCTGAATCTGTGCCTGCACTGAGAGGGCAAGGAGGTGGGGTGTTCTTCTTGGGAC CCCCACTAAGACCCTGGTCTGAGGATGTAAGATCGGAAGAGCGTCG
rs140761234_ref	CTAGTTGTTTAAAGCCCAACGCTAGTCATGCCGCGAGCTCGCTAGCCTCTCCAGCCTGGG AAAAGTTCTCCTTATTTGTTTAGATCCTTTTGTATTTTCAGATCTCCTTGGAGCAGTAGA GTACCTGGTAGACCATAATAGTGGAAAAGAGATCGGAAGAGCGTCG
rs140761234_alt	CTAGTTGTTTAAAGCCCAACGCTAGTCCGGCATCGAGCTCGCTAGCCTCTCCAGCCTGGG AAAAGTTCTCCTTATTTGTTTAGATCCTTTTGTATTTTGAGATCTCCTTGGAGCAGTAGA GTACCTGGTAGACCATAATAGTGGAAAAGAGATCGGAAGAGCGTCG
rs5756095_ref	CTAGTTGTTTAAAGCCCAACGCTAGTCCGGCTACGAGCTCGCTAGCCTTACACTCCCTCC CCTTTTGAAAGTCCCTAATAAAAACTTGCTGGTTTTGCGGCTTGTGAGGCATCACGGAAC CTACCGATGTGTGATGTCTCCCCTGGACAAGATCGGAAGAGCGTCG

rs5756095_alt	CTAGTTGTTTAAAGCCCAACGCTAGTCTAGCCGCGAGCTCGCTAGCCTTACACTCCCTCC CCTTTTGAAAGTCCCTAATAAAAACTTGCTGGTTTTGCAGCTTGTGAGGCATCACGGAAC CTACCGATGTGTGATGTCTCCCCTGGACAAGATCGGAAGAGCGTCG
rs1056801_ref	CTAGTTGTTTAAAGCCCAACGCTAGTCCGTACTCGAGCTCGCTAGCCTAGCAGTTGGTCT ATTCAGAATCAAACCTTTTTATATTTTATACTGCACTTTAGTGTATTTTCTGTCACTGTA GGTATAGAAGATCTGCCTCCCCTGTGGAAAGATCGGAAGAGCGTCG
rs1056801_alt	CTAGTTGTTTAAAGCCCAACGCTAGTCAGTACGCGAGCTCGCTAGCCTAGCAGTTGGTCT ATTCAGAATCAAACCTTTTTATATTTTATACTGCACTTCAGTGTATTTTCTGTCACTGTA GGTATAGAAGATCTGCCTCCCCTGTGGAAAGATCGGAAGAGCGTCG
rs1049508_ref	CTAGTTGTTTAAAGCCCAACGCTAGTCCGTGTCCGAGCTCGCTAGCCTTTTTCATATACA TCTTACCTCATTTCAAGTGAATTATTTTAATCTTTTTCTCTCTTTCCAAAAATTTACAGG AATGTTTAGTGTAATTGGATTTCGCTATCAGATCGGAAGAGCGTCG
rs1049508_alt	CTAGTTGTTTAAAGCCCAACGCTAGTCGACACGCGAGCTCGCTAGCCTTTTTCATATACA TCTTACCTCATTTCAAGTGAATTATTTTAATCTTTTCCTCTCTTTCCAAAAATTTACAGG AATGTTTAGTGTAATTGGATTTCGCTATCAGATCGGAAGAGCGTCG
rs2241871_ref	CTAGTTGTTTAAAGCCCAACGCTAGTCCTTCTTCGAGCTCGCTAGCCTTCTTGCCATAAA TTGCTGACTTCACGGTTAGGTAAAGATGGATAGATAAAGAGAAGCTTAAACAGTGGCATT TTACTACAAAGGCCTTCTATTTAATTTTTAGATCGGAAGAGCGTCG
rs2241871_alt	CTAGTTGTTTAAAGCCCAACGCTAGTCAAGAAGCGAGCTCGCTAGCCTTCTTGCCATAAA TTGCTGACTTCACGGTTAGGTAAAGATGGATAGATAAAGAGAAGCTTAAACAGTGGCATT TTACTACAAAGGCCTTCTGTTTAATTTTTAGATCGGAAGAGCGTCG
rs13004845_ref	CTAGTTGTTTAAAGCCCAACGCTAGTCATGAACCGAGCTCGCTAGCCTTTTTTTTTTTTC TTTACATTGGCTTTTTAAAAGACAGTTTTTATTTTTCAGATACTGGAATGTCAAGTATCT GGGTTGTCCCAAGACGTTGGAGATAATTCAGATCGGAAGAGCGTCG
rs13004845_alt	CTAGTTGTTTAAAGCCCAACGCTAGTCGTTCATCGAGCTCGCTAGCCTTTTTTTTTTTTC TTTACATTGGCTTTTTAAAAGACAGTTTTTATTTTTCAAATACTGGAATGTCAAGTATCT GGGTTGTCCCAAGACGTTGGAGATAATTCAGATCGGAAGAGCGTCG
rs3810609_ref	CTAGTTGTTTAAAGCCCAACGCTAGTCAATCACCGAGCTCGCTAGCCTCGTCTCAGCATC GCCAGTCTAGACTGTCTATGGAGAGCAGAAAGTTGTCTGGGGCTGCCTGGGGAACTGTGA GGCCAGCTATATCACCGTCGCTGATGGTGAGATCGGAAGAGCGTCG
rs3810609_alt	CTAGTTGTTTAAAGCCCAACGCTAGTCGTGATTCGAGCTCGCTAGCCTCGTCTCAGCATC GCCAGTCTAGACTGTCTATGGAGAGCAGAAAGTTGTCTAGGGCTGCCTGGGGAACTGTGA GGCCAGCTATATCACCGTCGCTGATGGTGAGATCGGAAGAGCGTCG
rs16975240_ref	CTAGTTGTTTAAAGCCCAACGCTAGTCGCAGGTCGAGCTCGCTAGCCTCGACGTAGGGTA AGTGCAATCCCAAGCCGTTTAAAATAATCCCAGACTGCCTGGAGGCTTTGTTCTTATTTT CTGATTCTTTTTTCTTTGTCTTTGTTGGAAGATCGGAAGAGCGTCG
rs16975240_alt	CTAGTTGTTTAAAGCCCAACGCTAGTCACCTGCCGAGCTCGCTAGCCTCGACGTAGGGTA AGCGCAATCCCAAGCCGTTTAAAATAATCCCAGACTGCCTGGAGGCTTTGTTCTTATTTT CTGATTCTTTTTTCTTTGTCTTTGTTGGAAGATCGGAAGAGCGTCG
rs10590_ref	CTAGTTGTTTAAAGCCCAACGCTAGTCCTCTACCGAGCTCGCTAGCCTTTGTGTATGCTT CATCACCTATATTAGGCAAATTCCATTTTTTTCCCTTGTGCTAAGGTAAAGATTTAATTA AATAATTTTGGCCTCTCATAGTTTTCTCTAGATCGGAAGAGCGTCG
rs10590_alt	CTAGTTGTTTAAAGCCCAACGCTAGTCGTAGAGCGAGCTCGCTAGCCTTTGTGTATGCTT CATCACCTATATTAGGCAAATTCCATTTTTTTCCCTTGCGCTAAGGTAAAGATTTAATTA AATAATTTTGGCCTCTCATAGTTTTCTCTAGATCGGAAGAGCGTCG