nature communications

Article

https://doi.org/10.1038/s41467-025-62000-9

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

Received: 29 July 2024

Accepted: 8 July 2025

Published online: 25 July 2025

Check for updates

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

Untranslated RNA sequences play essential roles in orchestrating gene expression. However, the sequence codes and mechanisms underpinning post-transcriptional regulation remain incompletely understood. Here, we revisit the finding from a prior massively parallel reporter assay (MPRA) that AU-rich elements in 3′ untranslated regions (3′ UTRs) can drive upregulation or downregulation of mRNA expression depending on 3′ UTR context. We unexpectedly discover that this variable regulation arises from widespread cryptic splicing, predominately from an unannotated splice donor in the coding sequence of GFP to diverse acceptor sites in reporter 3′ UTRs. Splicing is activated by U-rich sequences, which function as potent position-dependent regulators of 5′ and 3′ splice site choice and overall splicing efficiency. Splicing has diverse impacts on reporter expression, causing both increases and decreases in reporter expression via multiple mechanisms. We further provide evidence that cryptic splicing significantly impacts measurements made by other published 3′ UTR MPRAs. Overall, our work emphasizes U-rich sequences as principal drivers of splicing and provides strategies to minimize cryptic splicing artifacts in reporter assays.

Post-transcriptional regulation of RNA splicing, polyadenylation, translation, and degradation plays essential roles in shaping gene expression[1]. These post-transcriptional regulatory programs are predominantly encoded by non-coding RNA sequences located in intronic and untranslated regions (UTRs) that recruit diverse trans-acting factors, including small nuclear RNAs, RNA-binding proteins (RBPs) and microRNAs (miRNAs)[2,3]. Dysfunction of splicing and other post-transcriptional regulatory processes has been implicated in diverse human diseases[1]. However, the molecular codes that prescribe splicing and post-transcriptional regulation remain cryptic[4,5]. Understanding the functions and mechanisms of non-coding sequences continues to be a critical goal in biology.

Reporter assays have long been one of the most important strategies for studying post-transcriptional regulation. These assays place a reporter gene such as GFP under the control of non-coding regulatory sequences[6], enabling isolated measurement of non-coding sequence function. Recently, reporter assays have been extended to permit functional evaluation of thousands of non-coding sequences in parallel, termed massively parallel reporter assays (MPRAs)[6]. MPRAs leverage advances in gene-synthesis technology to clone large libraries of non-coding sequences into a common reporter vector in a pooled format. The pooled library is then transfected into cells and the activity of each non-coding sequence is measured via targeted next-generation sequencing. The flexibility and high throughput of MPRAs make them

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

useful for addressing diverse questions, including annotating the function of non-coding regulatory elements, evaluating the impact of genetic variants, and comparing homologous regulatory elements across species[7–13]. However, recent studies have identified multiple, frequently overlooked design choices that can convolute MPRA measurements[14]. For instance, the same candidate regulatory element can exhibit significantly different activity depending on whether the element is cloned with more or less of its endogenous surrounding sequence context[14]. Transfecting MPRA libraries as episomal vectors versus integrating them into chromatin using viral vectors, or use of alternative sequencing strategies to measure MPRA expression, can also generate different answers[15]. Fully defining the limitations and design caveats of MPRAs is important for ensuring that these experiments yield faithful measurements of non-coding sequence function.

Cryptic splicing has long been known to be an artifact that can plague reporter assays, causing alteration of reporter expression, false-positive identification of internal ribosome entry sites, and production of fusion proteins, among other examples[16–19]. Cryptic splicing arises when an inserted sequence element activates splicing from unexpected, usually weak 5′ donor or 3′ acceptor sequences. However, the requirements for activating cryptic splice sites remain poorly understood and are challenging to predict[20–22]. MPRAs may be particularly susceptible to cryptic splicing due to architectural features of reporter plasmids and the diversity of sequences that are assayed. Additionally, it is common for MPRA experiments to only sequence short segments of each reporter, meaning that cryptic splicing may go undetected. Better understanding the mechanisms underpinning cryptic splicing can help improve MPRA designs. Furthermore, studying cryptic splicing in the MPRA context may potentially offer insights into how mis-splicing arises in human disease contexts[23].

One of our groups recently used an MPRA-based strategy to evaluate combinatorial regulatory interactions between 3′ UTR sequence motifs[24]. This strategy, termed post-transcriptional regulatory element sequencing (PTRE-seq), employed a library of synthetic 3′ UTRs that encode arrays of four regulatory modules. Each module coded for either a "blank" control sequence, a let-7 miRNA site, Pumilio protein recognition element, Smaug protein recognition element, or an AU-rich element (ARE), which are bound by diverse ARE-binding RBPs. All 625 combinations of these regulatory modules and 17 additional natural let-7 sites were synthesized and cloned into the 3′ UTR of a common GFP reporter, resulting in a library of 642 unique reporters. PTRE-seq revealed that these regulatory elements generally function additively to repress mRNA expression and translation. However, as a surprising exception, AREs exhibited strong epistatic interactions with both adjacent AREs and other regulatory motifs, sometimes dramatically enhancing or repressing reporter expression depending on their relative position in the 3′ UTR. The mechanism for this variable, position-dependent impact of AREs on gene expression was unclear, but pointed to potential contextual control of 3′ UTR regulation[25].

In this work, we revisited the PTRE-seq MPRA experiment to better define the mechanism for position-dependent regulation by AREs. Unexpectedly, we discovered that PTRE-seq reporters undergo widespread cryptic splicing invisible to common MPRA measurement strategies that only sequence short reporter barcode segments. Cryptic splicing utilizes common GFP sequences and is activated by 3′ UTR ARE (U-rich) elements in a position-dependent manner, explaining the variable ARE phenotype. Repurposing this PTRE-seq dataset revealed new insights into splicing regulation, including how AU-rich elements modulate splice site selection and splicing efficiency, as well as the impact of splicing on mRNA expression. Additionally, we show that cryptic splicing impacts conclusions drawn by other MPRAs of 3′ UTR function. Overall, our work emphasizes the central role of U-rich elements in driving splicing and provides preventative solutions for minimizing splicing artifacts in MPRA designs.

# Results

## The PTRE-seq MPRA library features multiple unexpected size heterogeneities

As part of an effort to understand position-dependent ARE regulation, we repeated the PTRE-seq experiment[24] with slight modifications. HeLa cells were transfected with the same PTRE-seq plasmid library followed by total RNA isolation and amplicon sequencing to quantify reporter abundance. In the original experiment, reporter abundance was quantified via targeted sequencing of short co-transcribed barcode elements, which is a standard strategy used in MPRA experiments ("barcode primers", Fig. 1A)[6]. By contrast, we used a pair of "extended" primers to obtain sequencing coverage across most of the 3′ UTR and into the GFP CDS (Fig. 1A). Unexpectedly, the amplicon libraries obtained using these extended primers exhibited numerous shorter products (Fig. 1B). RNA abundance measurements obtained using extended primers were well-correlated with published barcode-based measurements (R = 0.80; Figure S1A), supporting that our experiments faithfully replicate the original PTRE-seq assay. Thus, these results indicated that mRNA transcripts in the PTRE-seq library have significant, unappreciated size heterogeneity.

Sequencing revealed multiple unexpected RNA products, including large numbers of spliced transcripts (detailed in subsequent sections), and a minority of transcripts featuring an insertion-deletion (indel) adjacent to the "spacer" region of each 3′ UTR (Fig. 1A, S1B). Long-read sequencing of the PTRE-seq plasmid library indicated that these indels derive from a cloning artifact (Figure S1B–D). Specifically, every PTRE-seq plasmid exists in two isoforms at a roughly 60:40 ratio (Figure S1E): (i) the expected reporter sequence, and (ii) a related copy containing a 78 bp palindromic insertion. This insertion artifact was missed by the original PTRE-seq study because it occurs in the variable region of the plasmid library and thus is invisible to Sanger and barcode-specific sequencing. Interestingly, the insertion product was also suppressed by our initial amplicon library preparation protocol, but was observed upon re-optimization of PCR (Fig. 1B). Despite the prevalence of the palindromic insertion at the DNA level, only 5–17% of RNA transcripts feature the insertion. This may reflect rapid turnover of RNA species containing the insertion, which is expected to fold into a perfect 39 base-pair hairpin (Figure S1B, S1C). Alternatively, the inserted hairpin may impede reverse transcription, resulting in sequencing dropout. We observed no correlation between RNA expression and the frequency of palindromic indels (R = −0.01, Figure S1F). To further evaluate whether the insertion impacts observed RNA expression, we repeated experiments using a variant of the PTRE-seq plasmid library that lacks a spacer sequence and is free of palindromic insertions (Figure S1B, S1D). RNA expression from this "no spacer" library was well correlated with the original PTRE-seq library (R = 0.7, Figure S1G). Together, these data support that the palindromic insertion does not significantly impact interpretation of PTRE-seq results. Notably, the "no spacer" library also exhibited significant numbers of shortened RNA products, confirming that RNA heterogeneity is unrelated to the palindromic insertion (Figure S1H). For the remainder of this paper, we focus on understanding this RNA heterogeneity, concentrating unless otherwise noted on the full PTRE-seq library for which we have more comprehensive data.

## The PTRE-seq MPRA library features widespread cryptic splicing

Although PTRE-seq reporters were not designed to be spliced and lack any known splice sites, sequencing of the PTRE-seq RNA indicated that the shorter species are products of cryptic splicing within reporter 3′ UTRs. A large subset of reporters featured internal deletions with precisely defined boundaries (Fig. 1C). Sequence analysis of the 5′ and 3′ deletion sites revealed strong enrichment for sequences that resemble splice sites, including characteristic GT and AG dinucleotides at position +1 of the 5′ site and −2 of the 3′ site respectively (Fig. 1D)[4]. U-rich regions were also strongly enriched upstream of the 3′ site,

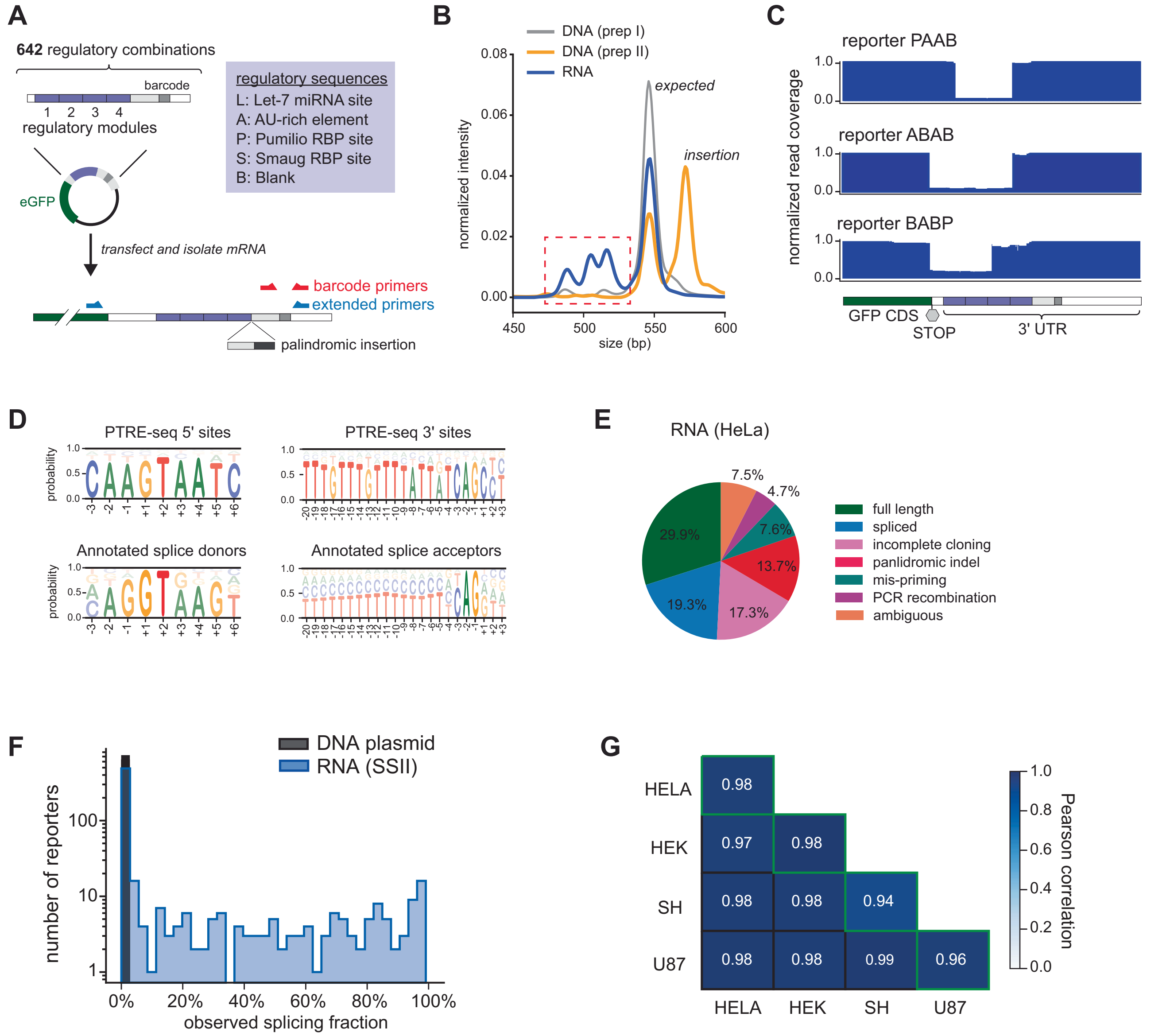

**Fig. 1 | The PTRE-seq MPRA library undergoes pervasive cryptic splicing.** **A** Schematic of the PTRE-seq library. Original "barcode" primers and "extended" primers used in this work are shown at bottom. eGFP, enhanced green fluorescent protein. The position of the palindromic insertion is indicated. **B** Size heterogeneity was observed in sequencing libraries prepared from the PTRE-seq DNA plasmid (grey and orange, denoting two different PCR protocols) and RNA (blue). Electropherogram was measured by Tapestation 4200. **C** Representative read coverage tracts of reporters with internal deletions in full PTRE-seq library. **D** Position weight matrices of sequences observed at 5′ and 3′ deletion boundaries observed in PTRE-seq reporters and annotated 5′ and 3′ splice sites retrieved from the human GRCh38 genome assembly. **E** Quantification of read categories from sequencing of the PTRE-seq libraries prepared from RNA reverse-transcribed with SuperScript II (SSII). Values denote the mean percentage for each category over two biological replicates. **F** Distribution of observed splicing fractions of PTRE-seq reporters. **G** Pearson correlations of PTRE-seq observed splicing fractions measured across biological replicates of the same cell line (diagonal, green outline) and between different cell lines.

consistent with a polypyrimidine tract[26]. We validated these MPRA-based observations with RT-PCR analysis on selected individually transfected reporters, which confirmed the presence of novel splice junctions in the 3′ UTR of PTRE-seq transcripts (Figure S2A). Identical splicing events were also observed in the "no spacer" PTRE-seq library (Figure S2B).

We developed a bioinformatics pipeline to rigorously quantity these splicing events. We filtered out all reads deriving from plasmids featuring palindromic insertions, and then categorized reads as "full length", "spliced" if having extended deletions flanked by +1 GT and −1 AG dinucleotides, or other anomalous (Methods). In the DNA library, 30% and 36% of reads contain "palindromic indels" or are "full length", respectively (Figure S2C). As expected, <0.1% of reads in DNA libraries were classified as "spliced", supporting that our pipeline has high specificity for identifying splicing (Figure S2C). Incomplete cloning artifacts (4%), mis-priming to homologous UTR regions during PCR (3%), PCR recombination (23%), and ambiguous deletions (3%) accounted for the remainder of DNA reads (Figure S2C, D). These anomalies persisted across multiple alternative PCR conditions, indicating that they are intrinsic to the repetitive PTRE-seq library (Figure S2C). By comparison, 19% of RNA reads are spliced, 30% are full length, 14% contain palindromic insertions, and 37% feature anomalies similar to those observed in DNA. Of note, the number of incompletely cloned reads doubles in RNA samples, consistent with reporters with incomplete 3′ UTRs being expressed at disproportionately higher levels. Usage of MarathonRT[27] versus

Superscript II reverse-transcriptase resulted in significantly fewer palindromic insertion containing reads (Figure S2C). Nonetheless, splicing fraction estimates were strongly correlated across diverse library preparation strategies (Figure S2E, S2F), and for the remainder of the paper we focus on sequencing libraries prepared using Superscript II. Per-reporter splicing fractions were also strongly correlated with fractions measured in the “no spacer” PTRE-seq plasmid library, with the exception of splicing events occurring in the spacer (R = 0.9, Figure S2G). This reproducibility across distinct plasmid libraries and sequencing preparation strategies supports that splicing is intrinsic to the PTRE-seq library and that our quantification is robust to palindromic insertion and other artifacts. Collectively, these analyses emphasize the need for thoughtful quality control during library preparation and bioinformatic quantification[14].

Quantification of the fraction of spliced versus full-length reads observed for each reporter revealed significant reporter-specific cryptic splicing (Fig. 1F). Out of 642 reporters, 119 (18%) were spliced at 20% or higher rates and 28 (5%) were nearly completely spliced ( > 90%). Observed splicing fraction of reporters were reproducible between biological replicates (Pearson correlation = 0.98) (Fig. 1G, S2H) and across internal barcode replicates (Figure S2I). Observed splicing fractions were also in agreement with orthogonal quantification by semi-quantitative RT-PCR of individually transfected reporters (Figure S2A). We also explored whether this cryptic splicing varied across different cell types. Observed splicing fractions were highly consistent across experiments performed in three additional cell lines: human embryonic kidney (HEK293), human neuroblastoma (SH-SY5Y), and human glioblastoma (U87) (Fig. 1G, S2C). Thus, cryptic splicing in the PTRE-seq library is constitutive and generalizable across diverse cell contexts.

### Splicing occurs at diverse weak donor and acceptor sites

Given that the PTRE-seq library was not designed to contain splice sites, we sought to better understand the sequence features driving efficient 3′ UTR splicing. Strikingly, 165 splicing events ( > 1% usage fraction) feature a cryptic 5′ splice site located at the stop codon of the eGFP coding sequence (Fig. 2A). Splicing also occurred from a donor located two codons upstream of the stop codon, albeit at significantly reduced observed fraction (alt-GFP; Fig. 2A). Splicing at these 5′ sites excises the stop codon, resulting in a C-terminal extension of GFP protein that was detectable by Western Blot (Figure S3A). These C-terminal-extended GFP proteins exhibit significantly lower expression than expected based on RNA abundance, consistent with the known destabilizing effect of C-terminal protein extensions (Figure S3B)[28]. While both the GFP and alt-GFP donor sites contain GT dinucleotides, they are otherwise weak sites compared to known human splice donors (1st percentile for both donors; Fig. 2B). Nevertheless, the GFP donor was able to support nearly 100% splicing fraction for multiple reporters (Figure S3C).

In addition to the GFP coding sequence donors, we also observed usage of a second set of cryptic 5′ splice sites in 31 spliced isoforms. These 5′ sites mapped to Pumilio recognition element (P) regulatory modules located in the 1st position of the 3′ UTR regulatory array (Fig. 2A, C). Like the GFP site, this P donor is weak (1st percentile relative to annotated human splice donors; Fig. 2B) but is capable of supporting >90% splicing fraction depending on neighboring sequence context (Figure S3C). Identical P modules located in the 2nd module position were also used as 5′ donors, although at significantly lower apparent frequency, potentially because usage of 2nd module sites may trigger non-sense mediated decay[29] (Fig. 2A). No splicing was observed from P donors in the 3rd or 4th positions, likely because there are no appropriately spaced downstream acceptor sites. Notably, because all PTRE-seq reporters contain potential GFP CDS donors, usage of 3′ UTR P donors occurs in competition with GFP, and complex patterns of alternative splicing were observed in some P-containing reporters (Fig. 2C, S3D).

Given that all PTRE-seq transcripts contain viable 5′ splice sites (the GFP CDS), we hypothesized that cryptic splicing might be determined by the availability and strength of downstream 3′ acceptor sites. However, 3′ splice acceptors exhibited minimal sequence dependence. 3′ splice sites predominantly mapped to the 3rd and 4th regulatory modules with all regulatory elements, apart from AREs, serving as splice acceptors (Fig. 2D). We also observed a subset of reporters that utilized a 3′ acceptor site in the constant spacer sequence (Fig. 2D), although we cannot distinguish whether these spliced products originate from normal or insertion-containing plasmids. All 3′ splice sites featured an AG dinucleotide and an upstream U-rich region (Fig. 2D). By contrast, ARE modules lack an AG dinucleotide motif, explaining why they are not used as acceptors. Similar to 5′ donors, these 3′ splice acceptors are substantially weaker than annotated human 3′ donors (1st to 28th percentile, Fig. 2E). Acceptor site strength also poorly correlated with observed splicing fraction (Figure S3E). The restriction of splicing to the 3rd and 4th regulatory modules but not the 1st and 2nd module positions can be explained by the 70-nucleotide minimum intron size needed for spliceosome assembly (Figure S3F)[30]. Altogether, our data indicate that factors beyond 5′ and 3′ splice site availability govern splicing in the PTRE-seq library.

To validate the function of the identified GFP splice sites, we used site-directed mutagenesis to synonymously recode the GFP C-terminus in several highly spliced reporters. Removal of the primary GFP donor via an AAG→AAA substitution at the -2 codon suppressed most splicing, but residual splicing was still observed at the secondary alt-GFP site (Fig. 2F). Full suppression was achieved with an additional CTG→CTA substitution at the -4 codon. Notably, these recoded GFP proteins were expressed at equivalent or greater levels than standard GFP (Figure S3G), indicating that these splice-suppressing mutations do not significantly impact translation. Overall, these results emphasize that common sequences can serve as cryptic splice sites and drive complex splicing behaviors.

### AU-rich elements activate cryptic splicing and determine splice site choice in the PTRE-seq library

Our analyses indicate that every reporter contains both feasible 5′ and 3′ splice sites, yet observed splicing fraction varied dramatically across reporters. Further analysis revealed that splicing was almost entirely determined by the presence and locations of ARE modules. Splicing strictly depends on the presence of at least one 3′ UTR ARE module within the cryptic intron (Fig. 3A, B, S4A). For some splice acceptor sites, such as B and P, a single ARE module is sufficient to activate strong splicing (Fig. 3C, S4B). For example, reporter BBBB is unspliced, but reporters BAPB and BBAB are >80% spliced (Figure S4C). An increased number of intronic ARE modules also drives increased observed splicing fraction at all acceptors, including activating weaker splice acceptors (Fig. 3C, S4B). For instance, L, S, and spacer acceptors are only used when there are at least two upstream AREs (Fig. 3C, S4B). This ability of multiple AREs to increase observed splicing fraction was observed regardless of whether AREs were contiguous or non-contiguous (Figure S4D).

Our data also reveal that AREs play a major role in dictating both 5′ and 3′ splice site choice. At the 5′ donor site, an increased number of “intronic” AREs promotes selection of P donors over GFP donors (Figure S4E, S4F). For instance, reporters PSAB and PAAB both contain a potential GFP donor and 1st module P donor and share identical 3′ acceptor sites. PSAB is predominantly spliced at the GFP donor, but PAAB is almost exclusively spliced at the P donor (Figure S4E). However, AREs that are 5′-adjacent to potential donors (i.e. are exonic) are repressive of splice site selection. For example, reporter BPAA is exclusively spliced using a P donor whereas the closely related APAA

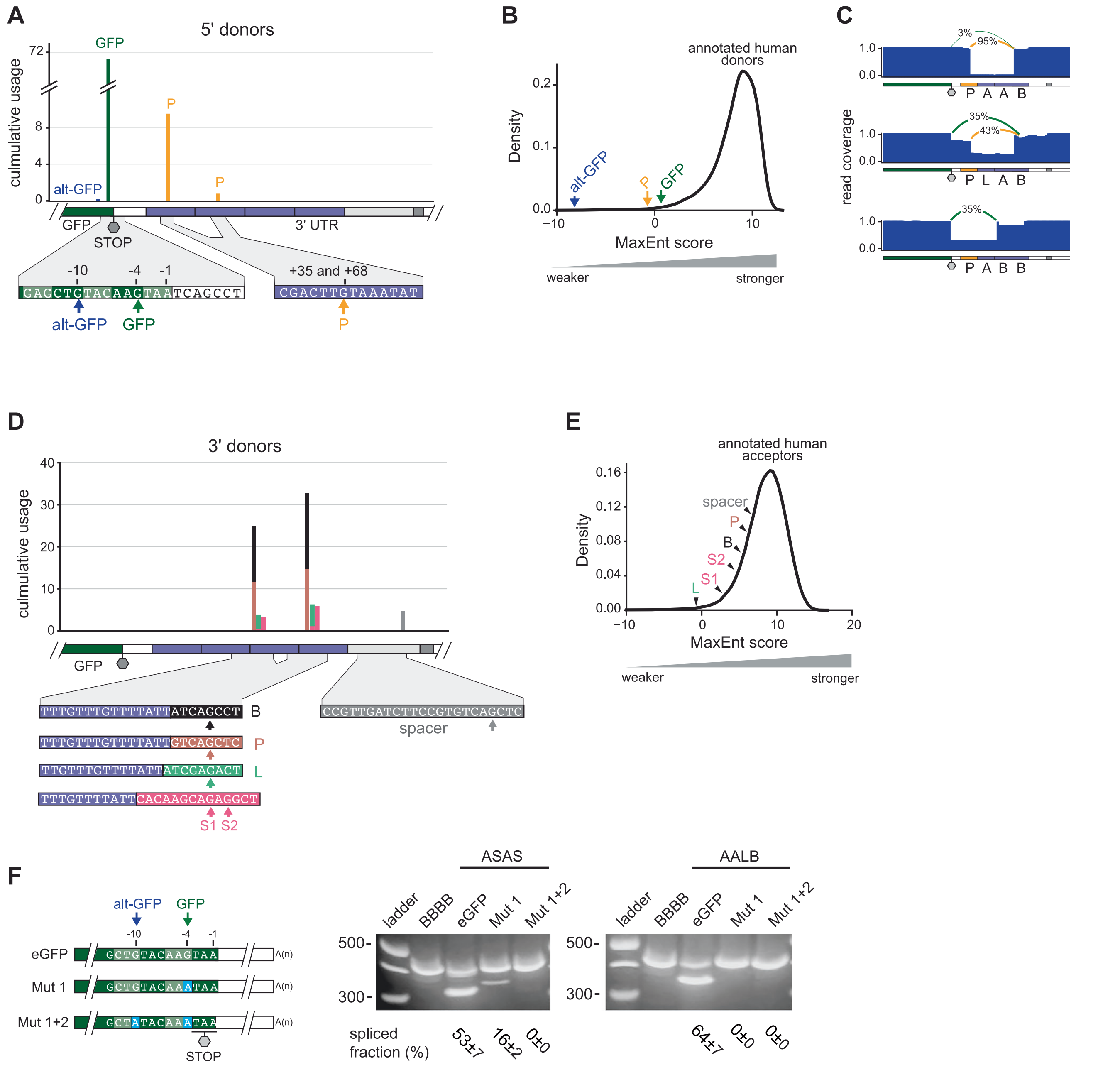

**Fig. 2 | Diverse weak splice donors and acceptors are used in the PTRE-seq library. A** Location and usage frequency of splice donors in PTRE-seq reporters. Cumulative usage was computed as the sum of observed splicing fractions for each splice donor across all reporters. Nucleotide sequences are detailed in the diagram below. Note that usage of GFP (−10, −4) and 1st module ( + 35) donor sites are not expected to trigger non-sense mediated decay, whereas 2nd module ( + 68) sites may. **B** 5′ donor site strengths of PTRE-seq splice donors and annotated human splice donors. **C** Representative read coverage tracts illustrating alternative usage of the GFP and P donors. **D** Location and usage frequency of splice acceptors in PTRE-seq reporters. Cumulative usage was computed as the sum of observed splicing fractions at each corresponding splice acceptor across all reporters. Sequence details of each acceptor are shown at bottom. **E** 3′ acceptor site strengths of PTRE-seq splice acceptors and annotated human splice acceptors. **F** Synonymous mutations that abolish GFP cryptic splice donors abrogate splicing in representative reporters. Schematics of mutations are shown at left. RT-PCR analysis of recoded ASAS and AALB reporters is shown at right. BBBB reporter is included as an unspliced control. Reporters were individually transfected into HeLa cells, followed by RT-PCR and resolved on an agarose gel. Quantification shown below the gel represents the median and standard deviation over 3 biological replicates. For (**B**, **E**), splice site strength is quantified by using Maximum Entropy score[58]. Uncropped gels are provided in the Source Data file.

reporter switches to the GFP donor site (Fig. 3D). This observation generalizes across reporters, with an ARE in the 1st position blocking usage of 2nd position P sites (Fig. 3E). Thus, AREs function as exonic silencers and intronic enhancers of 5′ splice sites.

For the majority of reporters, the 3′ splice site is located immediately downstream of the 3′-most ARE, reflecting that the ARE module likely functions as a polypyrimidine tract (Fig. 3F). When AREs are located in the 2nd or 3rd module, splicing occurs at the most proximal AG dinucleotide in the 3rd or 4th position, respectively (Fig. 3F). Similarly, in reporters where the 3′-most ARE is at the 4th position, the acceptor site shifted to the most proximal AG site in the adjacent spacer sequence (Fig. 3F). As exceptions to this rule, reporters with 3′-most AREs in the 1st position go unspliced, likely because the resulting intron would be less than 70 nucleotides (Figure S3F). The relative

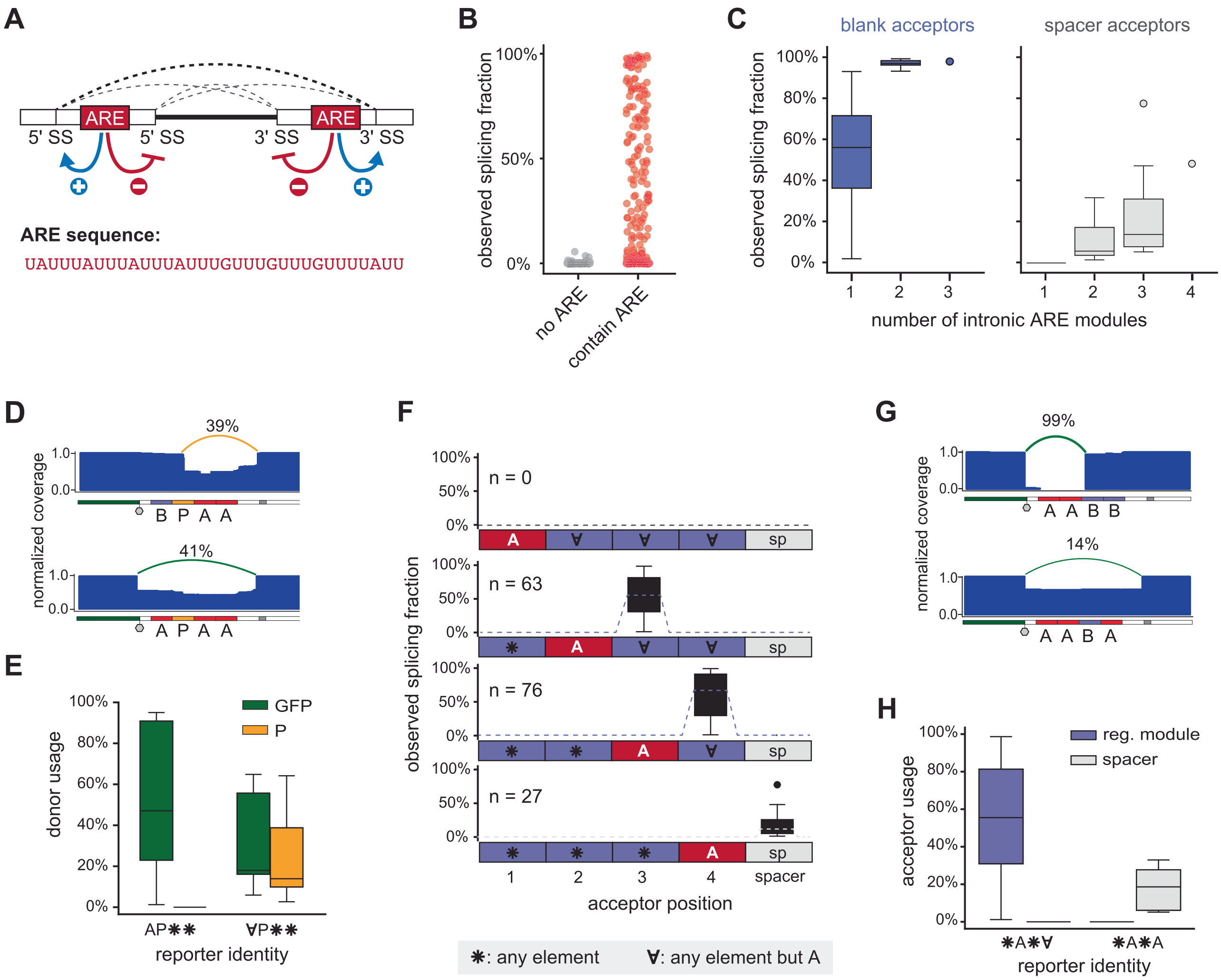

**Fig. 3 | AU-rich elements (AREs) are potent splicing regulators in the PTRE-seq library.** **A** The nucleotide sequence of the PTRE-seq ARE element and a cartoon illustrating its impact on splice site selection. **B** Observed splicing fractions of ARE-containing ($n = 371$) and non-ARE-containing ($n = 271$) PTRE-seq transcripts. **C** Relationship between the number of intronic AREs and observed splicing fraction for blank ($n = 51$) and spacer ($n = 27$) splice acceptors. **D** Example read coverage tracts illustrating ability of 5’ AREs to block usage of adjacent P donors. **E** Comparison of donor usage in reporters with ($n = 7$) and without ($n = 16$) AREs 5’ to a 2nd position P donor. **F** Splice acceptor usage based on location of the 3’-most ARE module. Box plots represent the distribution of observed splicing fractions observed at each potential acceptor site (1st, 2nd, 3rd, or 4th module sites, or spacer site) for different combinations of regulatory elements denoted along the bottom. **G** Example read coverage tracts illustrating ability of 3’ AREs to block usage of a strong upstream splice acceptor. **H** Comparison of acceptor usage in reporters without ($n = 60$) or with ($n = 8$) an ARE in the 4th regulatory module. For (**B**, **C**, **E**, **F**, **H**), splicing fraction was computed as the median of 10 internal barcodes measured across two biological replicates (20 total measurements). For (**C**, **E**, **F**, **H**), only acceptors spliced at fractions above 1% are included. 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. ⁎ symbol represents all regulatory elements, ∀ symbol represents all regulatory elements except ARE.

spacing between the ARE polypyrimidine tract and the downstream AG site also explained the difference in observed splicing fraction of different acceptor modules[26]: B and P modules feature AG sites only 5 nucleotides downstream from an adjacent ARE, whereas AG sites are located 6 and 9 nucleotides downstream in L and S modules (Fig. 2D). For the poorest splicing spacer acceptor, the nearest AG site is 30 nucleotides downstream of a 4th position ARE module. Spacer acceptors thus likely use a shorter and less U-rich polypyrimidine tract encoded within the spacer (Fig. 2D).

Interestingly, the rule that splicing occurs downstream of the 3’-most ARE applied even when superior splicing sites were available upstream (Fig. 3G, H). For instance, reporter AABB is >98% spliced using the favorable 3rd position B acceptor. By comparison, reporter AABA, which contains the same 3rd position B site, instead switches to the weaker spacer acceptor site and is spliced at only 14% splicing fraction (Fig. 3G). In the “no spacer” library, we similarly observed that upstream acceptors go unused when the 4th position is an ARE (Figure S4G). Thus, AREs can suppress usage of upstream acceptor sites, inducing switching to weaker downstream acceptors. Overall, our data demonstrate that “intronic” AREs are potent activators of cryptic splicing, whereas “exonic” AREs can suppress both 5’ and 3’ splice site selection (Fig. 3A).

## Cryptic splicing modulates reporter expression via multiple mechanisms

We next explored how cryptic splicing impacts PTRE-seq expression measurements. PTRE-seq was designed to measure changes in gene expression caused by post-transcriptional regulation by 3’ UTRs, but cryptic splicing may modulate expression and consequently convolute result interpretation. Indeed, reporter expression was strongly positively correlated with observed splicing fraction ($R = 0.74$), with the most highly spliced reporters exhibiting a > 4× higher expression than baseline “blank” 3’ UTRs (Fig. 4A).

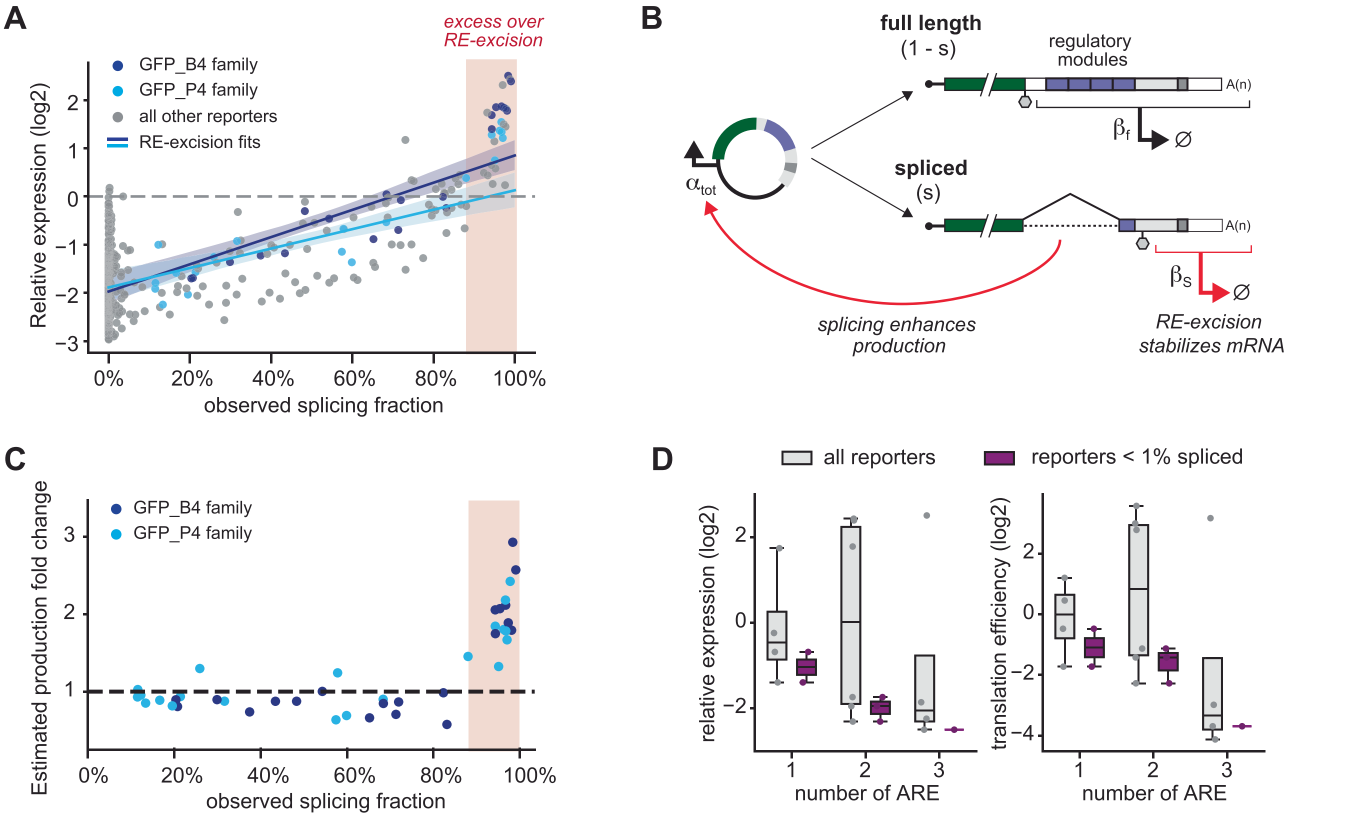

**Fig. 4 | Cryptic splicing enhances PTRE-seq reporter expression via RE-excision and enhanced production. A** Relationship between observed splicing fraction and reporter expression. GFP_B4 and GFP_P4 families of spliced transcripts are shown in dark and light blue, respectively. Regression lines of respective colors show the fits of RE-excision model to GFP_B4 and GFP_P4 reporters. Translucent bands show the 95% confidence interval for the regression estimates. The RE-excision model assumes constant production and is computed via Eq. 4 (Methods). Expression measurements are from the original PTRE-seq study[24] using barcode primers and normalized to the BBBB reporter. **B** Illustration of the mechanisms through which splicing impacts PTRE-seq reporter expression. **C** Estimated production rate for GFP_B4 and GFP_P4 reporters as a function of observed splicing fraction. Production rate is computed via Eq. 6 (Methods). **D** Effect of ARE module copy number on RNA steady state expression and translation efficiency. Shown are reporters consisting of all arrangements of blank modules with the indicated number of ARE modules. The original PTRE-seq study was unaware of cryptic splicing, making it appear that AREs have widely divergent impacts on mRNA stability and translation (grey boxplots). When spliced reporters are excluded, AREs have a uniformly destabilizing impact on reporter expression and translation efficiency (purple boxplots). Relative expression and translation efficiency were calculated as the median RNA and polysome sequencing measurements across 4 biological replicates reported in the original PTRE-seq study[24]. 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.

One mechanism through which splicing may increase reporter expression is by excising otherwise destabilizing 3′ UTR regulatory elements (Fig. 4B). We developed a simple "RE-excision" model that describes reporter expression as a sum of full-length and spliced isoforms that have differing stabilities. To fit this model, we focused on groups of reporters that are spliced at identical 5′ and 3′ sites, yielding identical spliced isoforms but with varied splicing fraction because of different cryptic intronic sequence elements (Figure S5A). For example, 21 reporters are spliced from a GFP donor to a 4th position blank sequence (GFP_B4), with observed splicing fraction ranging from 20% to 99% (Fig. 4A, Figure S5A). The stability of the full-length isoform of each reporter can be estimated from its combination of 3′ UTR regulatory elements (Methods). The stability of the common spliced isoform (GFP_B4) can then be estimated by regressing the observed total expression of each reporter versus its observed splicing fraction (Methods). Fitting this model to two different families of reporters indicated that spliced 3′ UTR isoforms are 40–60% more stable than baseline "blank" 3′ UTRs (Figure S5B–D). Accounting for RE-excision recapitulated the expression pattern observed for reporters spliced at <90% fraction (Fig. 4A).

While the RE-excision model explains the expression of reporters spliced at intermediate splicing fraction, it is unable to explain the dramatic increase of expression above the "blank" baseline for highly spliced reporters (Fig. 4A). We postulated that this increase reflected intron-mediated enhancement of reporter production (Fig. 4B)[31–34]. We solved for the relative production rate needed to explain the expression of each reporter using the full-length and spliced isoform stabilities obtained from RE-excision modeling (Methods). Low-to-moderately spliced reporters feature a constant production rate matching that of unspliced, reference transcripts (Fig. 4C). By comparison, our analysis indicates that reporters with observed splicing fractions >90% exhibit a 2- to 3-fold enhancement in production rate (Fig. 4C). This general trend of production rate increasing non-linearly at observed splicing fraction >90% was consistent across two independent families of reporters (Fig. 4C) and was robust to alternative model fitting strategies (Figure S5E). Similar trends were also observed in the "no spacer" PTRE-seq library (Figure S5F, S5G). Thus, cryptic splicing enhances reporter expression both via relieving RE-mediated destabilization and enhancing the production rate of highly spliced reporters.

## Cryptic splicing explains the position-dependent effects of AREs on PTRE-seq expression

Given the observation that ARE-directed splicing can significantly enhance reporter expression, we revisited our prior conclusion that AREs can sometimes stabilize mRNAs and enhance translational efficiency[24]. Limiting analysis to unspliced 3′ UTRs revealed that AREs induce a uniform reduction in mRNA expression and translation efficiency (TE) that scales with the number of ARE motifs (Fig. 4D). Identical results were observed in the "no spacer" library (Figure S6A). This observation aligns with the canonical role of AREs as destabilizing elements[35,36] and indicates that cryptic splicing likely explains the variable ARE effects observed in our original study. We also reexamined our prior conclusion that AREs can either enhance or antagonize the activity of adjacent P and let-7 regulatory modules depending on their relative positioning[24]. Regression modeling revealed that these variable epistatic interactions are also artifacts of splicing, with AREs weakly antagonizing P and let-7 elements regardless of positioning in unspliced reporters (Figure S6B, S6C). Other conclusions from the PTRE-seq study, which focused on the independent effects of P, let-7, and S sites, are not impacted by splicing[24].

## Cryptic splicing artifacts are common in other published MPRAs

Given that splicing in the PTRE-seq library only requires a GFP coding sequence and an ARE (U-rich) sequence, we investigated whether similar splicing artifacts are present in other published MPRAs. To do so, we developed a strategy to predict cryptic splicing using SpliceAI, a deep learning-based tool for predicting 5′ and 3′ splice sites in the human genome (Figure S7A)[37]. SpliceAI accurately predicted the position of the observed 5′ and 3′ splice sites for the large majority of spliced reporters in the PTRE-seq library (Figure S7B). To predict total splicing probability of each reporter, we computed the product of the maximum 5′ and 3′ SpliceAI site probabilities. This total splicing score strongly correlated with observed splicing fraction in PTRE-seq (Fig. 5A). A splicing probability <0.1 reliably identifies unspliced reporters, whereas a > 0.6 splicing probability provides a good predictor of efficient splicing. However, intermediate splicing probabilities have only moderate predictive value; for example, at predicted splicing probability of 0.3, reporters are equally likely to be highly spliced ( > 90%) as they are unspliced ( < 1%). The reduced predictive power of intermediate splicing probabilities is consistent with previous evaluations of SpliceAI[38]. We thus used total predicted splicing probabilities of >0.3, >0.6, and >0.9 as predictions of moderate, strong, and very strong splicing, respectively (Fig. 5A).

We applied this SpliceAI strategy to evaluate four other representative published MPRAs for which complete sequence information was available (Fig. 5B)[7–10]. Whereas PTRE-seq assayed synthetic 3′ UTRs, these other MPRAs assayed segments of natural human 3′ UTRs: the *Griesemer* library consists of 30,532 132-nt long segments containing putative functional 3′ UTR variants nominated by genome-wide association studies; the *Zhao* library consists of 2,828 highly conserved 200-nt long segments bearing diverse miRNA and RBP sites; the *Siegel* library consists of 41,288 200-nt long segments bearing diverse AU-rich elements; and the *Fu* library consists of 57,568 200-nt long segments containing rare 3′UTR variants. Each of these MPRAs used GFP as the reporter gene, although placed in different plasmid architectures. These MPRAs also used two different strategies to introduce reporters into cells: as exosomal plasmids (*Griesemer* and *Fu*) or integrated into chromatin using lentiviral vectors (*Zhao* and *Siegel*). Reporter expression for each MPRA was quantified by amplicon sequencing.

Our analysis indicates that all four MPRAs are predicted to undergo cryptic splicing to varying degrees (Fig. 5C). In *Griesemer*, ~50% of reporters are predicted to undergo splicing, with 16% having very strong splicing probability ($P > 0.9$). This splicing predominantly occurs from the same stop codon adjacent GFP donor site used in PTRE-seq (Figs. 5D, S7C). By comparison, the *Zhao*, *Siegel*, and *Fu* MPRAs have significantly less predicted splicing, with 1–10% of reporters having moderate to strong splicing probabilities. For these lesser spliced MPRAs, splicing predominantly occurs from donors in the 3′ UTR variable region (Fig. 5D, S7C). This reduction in predicted usage of the GFP donor site can be explained by differences in the plasmid backbone that serve to weaken the GFP donor site (*Zhao* and *Siegel*), or abrogation of the GFP donor site by synonymous substitution (*Fu*) (Figure S7D). Consistent with a critical role for U-rich sequences in driving splicing, motif analysis revealed strong enrichment of intronic U-rich motifs in predicted spliced reporters across all four studies (Fig. 5E).

To test our predictions, we analyzed published expression measurements of the *Griesemer*, *Zhao*, *Siegel*, and *Fu* MPRAs for evidence of gene expression changes linked to splicing. Given the design of these MPRAs, we expected that splicing would partially or completely disrupt the primer binding sites or excise regions used for reporter identification, resulting in sequencing dropout and apparent reduction in gene expression (Fig. 5B, S7C). Consistent with this expected impact, in the *Siegel* MPRA, reporters with predicted splicing probabilities above 0.6 and 0.9 exhibit 37% and 74% mean reduction in steady state expression, respectively, compared to low splicing probability reporters ($p < 10^{-64}$; Mann Whitney U test) (Fig. 5F). Similarly, in the *Zhao* MPRA, mean expression was reduced 43% and 90% for reporters with splicing probabilities above 0.6 and 0.9, respectively ($p < 10^{-4}$; Mann Whitney U test). The *Griesemer* MPRA exhibits less dynamic range in reporter expression (Fig. 5F). Nevertheless, reporters with predicted splicing probabilities >0.9 exhibited 46% mean reduction in expression ($p < 10^{-96}$; Mann Whitney U test). This splicing-associated reduction in expression is greater than the repression induced by AREs, Pumilio binding sites, and miRNA binding sites (8% to 32% reduction) observed in the *Griesemer* study[7]. In the *Fu* MPRA, mean expression was reduced by 74% for reporters with predicted splicing probabilities >0.6 ($p < 10^{-5}$; Mann Whitney U test), and the four >0.9 splicing probability reporters exhibited either anomalously high or low expression (Fig. 5F). Together, these observations support that cryptic splicing is both present and has the potential to convolute interpretation of functional motifs identified by these MPRAs.

The study design of the *Griesemer* MPRA provides an additional opportunity to evaluate the impact of cryptic splicing on measured gene expression. This MPRA was designed to evaluate the functional effects of disease-linked 3′ UTR single-nucleotide polymorphosisms (SNPs) on gene expression[7]. Each assayed 3′ UTR segment was expressed in a pair of reporters bearing either the reference (ref) or alternative (alt) allele. Variants were then considered functional (termed a transcript abundance modulating variant, or tamVar) if they induced a significant difference in expression between the ref and alt reporters. 19 pairs of reporters in the *Griesemer* MPRA exhibit significant differences in predicted splicing probability ( > 0.35; Fig. 5G). For tamVars where the alt allele is predicted to be more highly spliced, the alt allele exhibits lower expression than the ref allele. Conversely, when the ref allele is predicted to be more highly spliced then expression of the ref allele is reduced (Fig. 5G). Thus, predicted splicing explains both increases and decreases in expression caused by SNPs. An additional 266 tamVars were predicted to have high splicing probabilities ( > 0.9) in both alleles, indicating that cryptic splicing likely convolutes these tamVar measurements. Conservatively, our analysis suggests that 285 out of 2,368 SNPs (12%) identified as functional by *Greisemer* may be impacted by cryptic splicing.

To independently validate our SpliceAI predictions, we synthesized 20 reporters from the *Griesemer* MPRA comprising 10 ref/alt allele pairs identified as functional tamVars. These tamVars represent SNPs associated with prostate cancer, schizophrenia, anxiety disorder, and human evolution[7]. 2 pairs of reporters were predicted to be both unspliced, 5 were predicted to be both highly spliced, and 3 were predicted to be differentially spliced (one high, one low) (Figs. 5H, S8). Reporters were individually transfected into HEK293 cells and splicing was assessed using semi-quantitative RT-PCR. 12 out of 13 predicted highly spliced reporters demonstrated splicing (30–100% observed splicing fraction), corresponding to a positive predictive value of 92%,

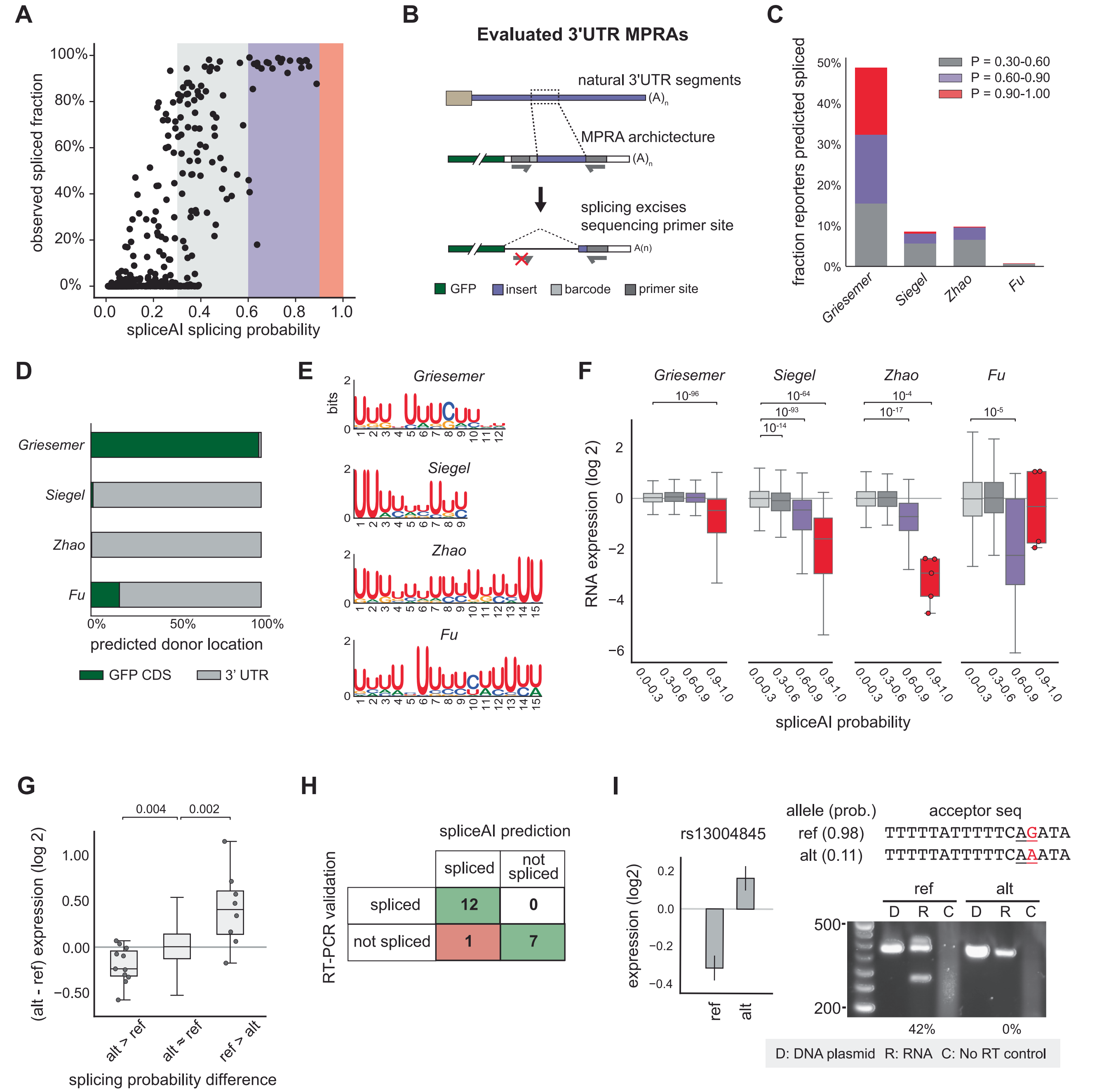

**Fig. 5 | Cryptic splicing is common in published MPRAs. A** SpliceAI[37] total predicted splicing probability positively correlates with observed splicing fraction observed in the PTRE-seq library. **B** Designs of evaluated 3′ UTR-focused MPRAs and predicted effect of cryptic splicing in these libraries. **C** Percentage of reporters predicted to be spliced from MPRAs described in (**B**). **D** Location of predicted splice donors (splicing prob. > 0.30). **E** Top enriched sequence motif in reporter transcripts with >0.60 splicing probability for each MPRA. **F** Reporter expression as a function of predicted splicing probability. For clarity, data points outside the expression range (−6, 3) are not shown. **G** Opposing differences in RNA expression between paired ref and alt single-nucleotide variants in the *Griesemer* MPRA correlate with predicted splicing. alt > ref and ref > alt denote variants in which predicted splicing probability is 0.35 or greater between the ref and alt alleles. ref ≈ alt denotes alleles where the difference in splicing probability is less than 0.35. Expression measurements from HEK293FT cells are shown[7]. Data points outside of the whiskers of the ref ≈ alt group are not shown. **H** Summary of RT-PCR validation of predicted spliced reporters from the *Griesemer* MPRA. See Figure S8 for raw data. **I** Example functional variant from the *Griesemer* MPRA that is explained by cryptic splicing. At right are PCR products from DNA plasmids, RNA, and no RT control. Bar plot shows mean relative expression measured by *Griesemer* with error bars denoting standard error[7]. SpliceAI probability is shown below along with the sequence of the predicted 3′ splice site (underlined). The variant is highlighted red. RT-PCR analysis of individually transfected reporter into HEK293T cells is shown at right. For (**F**, **G**, **I**), RNA expression was obtained from the corresponding MPRAs (mean across 3 replicates for Siegel, Zhao, and Fu, and 6 replicates for Griesemer). 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. *P*-values shown above plots were computed using two-sided Mann-Whitney U tests. Uncropped gels are provided in the Source Data file.

whereas all 7 low splicing probability reporters were unspliced (Figs. 5H, S8). For example, SpliceAI predicted that the rs13004845 tamVar contains a strong 3′ AG splice site in the ref allele which is abolished by a G-to-A substitution in the alt allele (Fig. 5I). Our experiments confirmed that the ref allele was spliced with 42% efficiency, whereas the alt allele is unspliced, consistent with the 28% decrease in ref allele expression measured by the *Griesemer* MPRA. Similarly, variants rs5756095 and rs140761234 create and weaken predicted splicing acceptor sites, respectively, inducing differential splicing that explains the changes in expression measured by MPRA (Figure S8B). 3 of the reporter pairs predicted to be highly spliced in both ref and alt alleles also exhibit significant differences in observed splicing fraction consistent with the differential expressions observed by MPRA (Figure S8C). Overall, these results emphasize the ability of cryptic splicing to cause spurious conclusions across diverse MPRA studies, impacting interpretation of gene regulatory mechanisms and misclassification of human SNPs.

## Discussion

Cryptic splicing is a well-known phenomenon[4], but one that it is generally assumed to be rare. In this study, we report that cryptic splicing likely affects between 1% and 50% of reporter genes bearing functional 3′ UTRs. Using resequencing and reanalysis of the PTRE-seq MPRA, we establish that cryptic splicing is predominantly determined by the presence and arrangement of intronic AU-rich elements that both potentiate splicing and dictate 5′ and 3′ splice site choice. We further show that this cryptic splicing has diverse impacts on reporter assay measurements, with potential to both upregulate and downregulate apparent gene expression depending on splice site locations and measurement strategy. These findings have important implications for the future design and interpretation of MPRAs and reveal new insights into the sequence codes governing splicing activation and splice site choice[23].

A major driver of splicing in the PTRE-seq library is an undocumented cryptic splice donor in the GFP coding sequence. While relatively weak compared to endogenous donors, this site can still drive efficient splicing into the 3′ UTR given the presence of U-rich sequences and viable downstream splice acceptors (Fig. 2). Our analysis of other MPRAs supports that this splicing site is used in other reporter contexts. Since this donor site overlaps the stop codon of GFP, the resulting GFP protein product contained a C-terminal extension and exhibited lowered stability (Figure S3)[28], emphasizing that cryptic splicing can impact gene expression at multiple levels. We show that this GFP-driven splicing can be suppressed using synonymous recoding of the −4 and −2 C-terminal codons (Fig. 2), and the *Fu* MPRA[10] that incorporated a similar synonymous substitution exhibited significantly reduced predicted splicing (Fig. 5C, S7D). We therefore suggest that these synonymous substitutions should be universally incorporated into GFP reporters to mitigate undesirable splicing. However, it is important to emphasize that cryptic splicing may occur from other upstream GFP sites and is unlikely to be restricted to GFP reporters. Notably, SpliceAI analysis predicts that other common reporter genes including fluorescent proteins such as Venus and YFP, and luciferases such as hRluc and Nanoluc, also harbor cryptic donor sites that can support cryptic splicing to 3′ UTR acceptor sites (Figure S7E). Furthermore, splicing can occur from donors located in 3′ UTRs, which we observe in PTRE-seq and predict in other MPRAs (Fig. 2, Fig. 5). Thus, synonymous substitution or use of alternative reporter genes will not completely eliminate cryptic splicing.

Our results revealed an unexpectedly central role of AU-rich elements as potent activators of cryptic splicing. In PTRE-seq, AREs were capable of activating highly efficient splicing from otherwise weak 5′ and 3′ splice sites. This ability to activate splicing includes but is not limited to the ability of AREs to serve as a polypyrimidine tract; for example, AREs are able to activate splicing 30 nucleotides downstream at the weak "spacer" acceptor site (Fig. 3C). A limitation of our study is that PTRE-seq only assays a single ARE sequence, albeit in many different contexts. Defining a more generalized understanding of the sequence features responsible for splicing activation, including potential roles of other U-rich and C-rich sequences, is an important topic for future studies. Nevertheless, we show that U-rich motifs are associated with cryptic splicing signatures in other MPRAs (Fig. 5E). Our results are also consistent with the well-documented ability of U-rich tracts in *Alu* and other transposable elements to activate weak splice sites and drive evolution of new exons[22,39]. In splicing of evolved (non-cryptic) exons, intronic U-rich elements are associated with high splicing efficiency and with activation of proximal alternative 5′ splice sites[40,41]. Collectively, these data support a model in which U-rich elements can function as principal splicing stimuli.

We also found that AREs exert a powerful position-dependent influence on both 5′ donor and 3′ acceptor splice site choice (Fig. 3). Most notably, placement of AREs in exonic locations relative to potential donor and acceptor sites (5′ and 3′ of sites, respectively) is sufficient to completely suppress splice sites that are efficiently spliced in other contexts. A number of alternative splicing factors, such as various members of the hnRNP family, are known to bind U-rich sequences and exhibit position-dependent effects on splice site selection[22,40,42,43]. However, a dominant role for exonic U-rich sequences in prohibiting splice site selection has not been previously documented, and further investigation is needed to define the mechanisms behind this positional dependence. Overall, these findings suggest that U-rich sequences may play broader roles in defining intronic architecture than previously appreciated.

Our observation that cryptic splicing is activated by relatively simple U-rich motifs also has implications for understanding ubiquitous mis-splicing events in disease. U-rich motifs with appropriately spaced downstream AG sites that could serve as cryptic splice acceptors are prevalent throughout the transcriptome. For example, we find that many natural 3′ UTR sequence elements serve as efficient splice acceptors in MPRA contexts (Fig. 5). These natural 3′ UTR sequences go unspliced in their endogenous contexts, implying that mechanisms normally suppress splicing at these 3′ UTR sites and that dysregulation of these mechanisms could be sufficient to activate cryptic splicing[22,44]. Indeed, a recent study reported widespread upregulation of cryptic splicing in 3′ UTRs in cancer, with these splicing events also exhibiting a strong association with U-rich elements[45]. Some of the disease-associated mutations that activate cryptic splicing in MPRA contexts (Fig. 5) may also be functional in endogenous contexts, inducing cryptic splicing that impacts gene expression.

In the context of MPRAs, cryptic splicing has complex impacts on reporter expression, including reducing apparent expression due to sequencing dropout, altering expression via removal of regulatory elements, and enhancing the production of spliced reporters (Figs. 4B, 5B). Other potential mechanisms through which cryptic splicing may alter reporter expression include enhancing mRNA polyadenylation and export, protecting spliced reporters from m6A modification, and more generally increasing transcript stability[46–48]. The 2–3 fold enhanced production we observed in highly spliced PTRE-seq reporters is consistent with prior measurements of intron-mediated enhancement of transcription[31–34]. Interestingly, our analysis suggests that enhanced production only emerges when reporters are >90% spliced (Fig. 4C). Studies of the likely related phenomenon of exon-mediated activation of transcription starts (EMATS) also observed strong threshold effects, with only introns spliced in at >95% frequency inducing strong EMATS[34]. The mechanisms behind splicing-induced production enhancement and the threshold effects observed in our analyses require further validation, but are reminiscent of threshold effects observed during phase-separation processes and we speculate may be linked to partitioning into transcriptional condensates[49,50].

Our re-analysis of published MPRAs revealed that cryptic splicing and other artifacts can significantly impact MPRA conclusions. In the PTRE-seq MPRA, ARE-induced splicing led to the erroneous conclusion that AREs can stabilize mRNAs and promote translation depending on their 3′ UTR context. Once splicing is accounted for, we find that AREs uniformly reduce mRNA expression, consistent with the prevailing model of ARE function[8,35,36]. In other 3′ UTR-focused MPRAs, we provide evidence that splicing-induced sequencing dropout results in anomalously low mRNA expression, comparable to or greater than the reductions in expression induced by bona fide post-transcriptional regulatory motifs. Because MPRAs are often analyzed with the goal of identifying sequence elements that convey strong regulatory effects, these cryptically spliced sequences are particularly liable to be identified as functionally interesting (Fig. 5). Indeed, we show that cryptic splicing likely impacts more than 10% of functional SNPs identified by a recent MPRA designed to assess human disease-associated variants[7]. Nevertheless, it is important to emphasize that the majority of MPRA measurements remain valid, though they should be interpreted with appropriate caution.

In conclusion, cryptic splicing is unexpectedly common in MPRAs due both to MPRA design features and the apparent ease at which splicing is activated in transgenes. We offer several recommendations for how to protect against such splicing artifacts moving forward. First, MPRAs can be performed by transcribing mRNAs in vitro and then transfecting[51,52] or microinjecting[53] them into cells, which completely circumvents the possibility of splicing. However, RNA-based MPRAs are generally more complex to implement than DNA-based MPRAs. For DNA-encoded MPRAs, future studies should explore whether cryptic splicing can be reduced by improved reporter architectures, such as using different promoters or including designed introns. The PTRE-seq reporter, like many other MPRA reporters, is driven by the CMV promoter and lacks any encoded introns, which is atypical for human genes and may inadvertently promote cryptic splicing. More concretely, we recommend prospectively evaluating MPRA library designs using splicing prediction algorithms such as SpliceAI[37] or Pangolin[54] to identify and edit-out potential cryptic splice sites[37]. Particular attention should be made to remove recurrent cryptic sites in the core reporter, such as the GFP site we characterize here. However, we emphasize that cryptic splicing can occur from idiosyncratic sites throughout reporter mRNAs and will likely not always be predicted by prediction algorithms. Furthermore, we document how cloning artifacts, reverse-transcription, and PCR errors can introduce significant unexpected complexities in MPRA data. Therefore, it is critical to confirm the integrity of MPRA reporters using full-length sequencing, which is now widely available via nanopore sequencing kits and commercial sequencing providers. These quality control steps will ensure more robust and reliable MPRA experimental outcomes and interpretation.

## Methods

### PTRE-seq library design

The design and construction of the pooled PTRE-seq plasmid library was described previously[24]. In brief, a pooled library of synthetic 3′ UTRs was synthesized consisting of a 132-nucleotide variable regulatory array, a 20-nucleotide spacer cloning site, and a 9-nucleotide identifying barcode. Each of the variable regulatory arrays consists of 4 33-nucleotide long regulatory modules encoding a combination of (a) “blank” control sequences, (b) let-7 miRNA binding sites, (c) Pumilio protein recognition sites, (d) Smaug protein recognition sites, and (e) AU-rich elements (Fig. 1). Each regulatory array is present in 10 copies, each bearing a different barcode, which function as internal replicates. In addition, the library contains 40 additional copies of the control sequence (4× “blanks”), 50 copies of a low expression control (4×let-7 perfect complement), and a series of constructs containing natural and synthetic let-7 sites. To obtain the “no spacer” library, this pool was cloned downstream of the eGFP coding sequence in the pCDNA5/FRT/TO plasmid (Addgene 19444), followed by plasmid purification from ~40,000 colonies. To construct the full PTRE-seq library, an additional 25 nucleotide spacer sequence was cloned into the spacer cloning site and plasmid DNA was collected from ~250,000 colonies. The palindromic insertion discovered in the full PTRE-seq library consists of an additional spacer sequence and 33 nt reverse complement to the spacer inserted adjacent to the spacer (Figure S1). Since these palindromic insertions were observed in the full PTRE-seq library but not in the “no spacer” library, we conclude that this insertion arose during cloning or bacterial replication (Figure S1). Plasmid maps for 3 representative plasmids are provided as Supplementary Files.

### Cell culture and transfection of PTRE-seq libraries

HeLa (obtained from the Baylor College of Medicine Tissue Culture Core), HEK293 (R78007, ATCC), SH-S5Y5 (CRL-2266, ATCC), U87 MG (HTB-14, ATCC), and T-REx TM-293 cells (R71007, Thermo Fisher) cells were grown in DMEM (Gibco) supplemented with 10% FBS (Gibco), 1x Penicillin streptomycin and glutamine (Gibco) and 1x MEM non-essential amino acids (Gibco). The full PTRE libraries and the “no spacer” library were transfected using the Neon Transfection System (Invitrogen) per manufacturer protocol. For each transfection, $2.5\times10^6$ cells were electroporated with 8 μg of the PTRE-seq library. RNA was isolated and purified 40 h post-transfection via RNeasy mini kit (Qiagen), treated with Turbo DNase (Invitrogen, AM2238) and purified using Mag-Bind TotalPure NGS beads (Omega Bio-tek). Cells were regularly tested using LookOut Mycoplasma PCR detection kit and confirmed to be mycoplasma free.

Transfection of individual PTRE-seq plasmids and site-directed mutants were performed using Effectene (Promega) or X-tremeGENE™ 9 DNA Transfection Reagent (Roche), respectively. Cells were transfected in a 6-well plate with 1 μg plasmid, then split 24 hr later into two separate 6-well plates. After 40 h, RNA was isolated using RNeasy kit (Qiagen), treated with Turbo DNase (Invitrogen) and purified using Mag-Bind TotalPure NGS beads (Omega Bio-tek).

### Semi-quantitative PCR of individual PTRE-seq reporters

2 μg total RNA from individual transfections was reverse transcribed using Superscript II reverse transcriptase (Invitrogen) using RE_Amp_RT primer or Junction_R (Table S1). cDNA product was PCR amplified (Q5, NEB; 98 °C for 30 s, 35 cycles: 98 °C for 10 s, 71 °C for 20 s, 72 °C for 30 s, and 72 °C for 2 min) using GFP_Amp_F_seq and RE_Amp_R_seq or Junction_R primers (Table S1) and visualized on 1.5% agarose gel stained with ethidium bromide. Agarose gels were imaged using a Bio-Rad ChemiDoc MP Imaging System, and band intensity was quantified using ImageJ (v.1.54p). Uncropped images are provided as Source Data.

### RNA and DNA sequencing of PTRE-seq libraries

RNA sequencing libraries from transfection of the full and “no spacer” PTRE-seq libraries were prepared from 2 μg total RNA. Reverse transcription was performed using Superscript II (Invitrogen) or MarathonRT (Kerafast) using the RE_Amp_RT primer according to manufacturer protocols. Unless otherwise stated, analyses were done using sequencing libraries prepared with Superscript II.

To minimize potential recombination artifacts, emulsion PCR (Micellula DNA Emulsion & Purification Kit, EURx) was used to prepare sequencing libraries. An oil-surfactant mixture was freshly assembled according to manufacturer’s instruction and prechilled at 4 °C. Water phase samples were prepared as described below and mixed via vortexing at 4 °C for 5 m. Emulsion PCR products were column-purified after breaking the emulsions by addition of 1 ml of 2-butanol. DNA was eluted using the kit elution buffer preheated to 65 °C and was further purified using Mag-Bind TotalPure NGS beads.

We adapted a previously described 2-step PCR protocol[55] to amplify the 3′ UTR and attach Illumina adaptors with indexes. We refer to the following protocol as "prep I". For step 1 PCR, each 50 µL reaction consists of 1 µL of cDNA product, 1 unit of Q5 High-Fidelity DNA Polymerase (NEB), 10 µL of Q5 Reaction Buffer (NEB), 200 µM dNTPs (NEB), 2 µM of GFP_Amp_F_seq and RE_Amp_R_seq primers, and 2.5 µg of BSA (Sigma-Aldrich, A9418). PCR1 cycling conditions were as follows: 98 °C for 30 s, 20 cycles (98 °C for 10 s, 71 °C for 20 s, 72 °C for 30 s), and 72 °C for 2 min. For step 2 PCR, each 50 µL reaction consists of 100 pg of PCR step 1 product, 1 unit of Q5 High-Fidelity DNA Polymerase (NEB), 10 µL of Q5 Reaction Buffer, 200 µM dNTPs, 1 µM of GFP_Amp_F_seq and RE_Amp_R_seq primers, and 2.5 µg of BSA. PCR2 conditions are as follow: 98 °C for 30 s, 15 cycles (98 °C for 10 s, 68 °C for 20 s, 72 °C for 30 s), and 72 °C for 2 min. Matching DNA sequencing libraries from the DNA plasmid were prepared by inputting 5 ng of plasmid into the same two-step PCR protocol. DNA and RNA libraries from the "no spacer" library were prepared with the same "prep I" protocol. Unless otherwise stated, analyses were done using sequencing libraries prepared with this "prep I" protocol.

We found emulsion PCR suppressed the amplification of palindrome-containing templates. To better capture the frequency of this artifact, we used standard PCR which we refer to as "prep II". For step 1 PCR, each 25 µL reaction consists of 5 ng of plasmid, 1 unit of Q5 High-Fidelity DNA Polymerase (NEB), 5 µL of Q5 Reaction Buffer (NEB), 200 µM dNTPs (NEB), 10 µM of GFP_Amp_F_seq and RE_Amp_R_seq primers. PCR1 cycling conditions were as follows: 98 °C for 30 s, 20 cycles (98 °C for 10 s, 71 °C for 20 s, 72 °C for 60 s), and 72 °C for 2 min. For step 2 PCR, each 25 µL reaction consists of 100 pg of PCR step 1 product, 1 unit of Q5 High-Fidelity DNA Polymerase (NEB), 5 µL of Q5 Reaction Buffer, 200 µM dNTPs, 1 µM of GFP_Amp_F_seq and RE_Amp_R_seq primers. PCR2 conditions are as follow: 98 °C for 30 s, 15 cycles (98 °C for 10 s, 68 °C for 20 s, 72 °C for 60 s), and 72 °C for 2 min. We found no significant differences in splicing estimates and frequency of anomalous reads between "prep I" and "prep II" protocols (Figure S2C, S2E).

The following libraries were sequenced on an Illumina MiSeq (v2 chemistry, 2 × 250 bp; or v3 chemistry, 2 × 300 bp): DNA plasmid library (2 replicates), HeLa RNA library prepared from Superscript II (2 replicates), HEK293 RNA library (1 replicate), SH-SY5Y RNA library (1 replicate), and U87 RNA library (1 replicate). The following libraries were sequenced on an Illumina NovaSeq 6000 (SP flowcell 2 ×250 bp): HeLa RNA library prepared from MarathonRT (2 replicates), HEK293 RNA library (2 replicates), SH-SY5Y RNA library (2 replicates), and U87 RNA library (2 replicates). Measured splicing fractions were consistent across sequencing platforms.

### Nanopore sequencing

1 µg of the PTRE-seq plasmid library was linearized using EcoRI-HF (NEB) per manufacturer's instruction and linearization confirmed via agarose gel analysis. The linearized product was purified using Mag-Bind TotalPure NGS beads and sequenced using Plasmidsaurus premium PCR sequencing service ("Huge" category). Sequencing was performed for three technical replicates of the plasmid library, and data from these replicates pooled for subsequent analyses.

### Western blot analysis

Protein lysates were collected from cells transfected with individual reporters (cell lysis solution, Cell Signaling Technology), resolved on a 4–12% Bis-Tris gel (Bio-Rad), and transferred onto Immuno-Blot PVDF (Bio-Rad). After blocking (5% milk in 1x PBS with 1% Tween, 1 h minimum), proteins were incubated with primary antibodies against GFP (Clonetech, 632381) at 1:5000 dilution and β-actin-HRP (Cell Signaling, 12261) at 1:1000 dilution for 1 h at room temperature or overnight at 4 °C. Following washing with 1x PBST, proteins were incubated with anti-mouse secondary antibody coupled to horse-radish peroxidase (Cell Signaling, 12262) at 1:10000 dilution. After washing with 1x PBST, imaging was conducted using a Bio-Rad ChemiDoc XRS System and band intensity was quantified using integrated Image Lab software. Uncropped images are provided in the Source Data file.

### Mutation of GFP splice donor sites in PTRE-seq reporters

Site-directed mutagenesis was used to introduce synonymous mutations that ablate the GFP splice donor sites in the ASAS and AALB reporters. PCR was performed following manufacturer's instructions (Q5; NEB) using M1_For and M1_Rev primers for mutant 1 and M2_For and M2_Rev primers for mutant 2 (Table S1). Plasmids were transformed into *E. coli*, purified with Qiagen Plasmid Midi Kit (Qiagen), and confirmed via Sanger sequencing.

### Barcode mapping and splice isoform quantification

Sequencing data were processed by first merging paired-end reads using BBMerge (v.39.19) with default parameters[56]. The pooled MPRA library was then demultiplexed based on internal barcodes using a custom Python script. The edit distance was calculated between the barcode identified in each sequencing read against all 6500 expected barcodes (search sequence = 9-nt barcode + GGTACC flanking sequence). Reads were assigned to a reporter if the edit distance was 1 or less. BBMap[57] (v.39.19) was then used to align assigned reads to the reference sequence of the reporter using default parameters.

Following demultiplexing and mapping, each output bam file was analyzed using a custom Python script to classify reads into one of seven categories:

i. Palindromic indel: affected templates featured a palindrome which is a perfect reverse complement to the spacer (Figure S1C). We observed that this palindrome was often truncated at the 5′ end, presumably because the palindrome-spacer duplex impedes reverse transcription and/or PCR. Reads were assigned as "palindromic indels" if the spacer was immediately preceded by a "CGGTCGG" insertion – the 3′ portion of the palindromic sequence.
ii. Full length: reads were assigned as full length if the edit distance between the read and the reference sequence was less than 20.
iii. Spliced: reads were assigned as spliced if they (a) featured a deletion with GT and AG dinucleotides at the 5′ and 3′ end respectively, and (b) the edit distance between the read and the spliced isoform was less than 20.
iv. Incomplete cloning: reads were assigned as incompletely cloned if they lacked either of the two cloning insert sequences.
v. Mis-priming: the PTRE-seq plasmid backbone contained a sequence in the 3′UTR identical to the "blank" module (Figure S2D). This sequence can anneal to the upstream "blank" module and generate a truncated PCR product. Reads are assigned as mis-priming if they (a) featured a long deletion without a GT-AG junction and (b) miss the spacer and barcode.
vi. PCR recombination: PCR chimeras can arise from annealing of incomplete extension products, which can result in the barcode from one reporter being appended to a different reporter. After barcode demultiplexing, chimeric reads align poorly to their expected reference sequence, resulting in numerous indels. Reads were assigned "chimeric" if they featured more than two indels in the regulatory array region.
vii. Ambiguous: reads were assigned as ambiguous if they featured a > 20 nt deletion not flanked by GT and AG dinucleotides, or feature 2 or more mismatches or indels in the barcode.

The observed splicing fraction for each barcode was calculated as:

$$splicing\ fraction = \frac{spliced\ read\ count}{full\ length\ read\ count + spliced\ read\ count} \times 100\%$$

Each regulatory array is present in 10 copies in the PTRE-seq library, each with a different barcode, resulting in 10 possible splicing fraction measurements per biological replicate. In order to compute a barcode-specific splicing fraction, we required a minimum of 10 total spliced and full-length counts. Final splicing fraction measurements were computed as the median of all measurements passing this minimum count filter (a maximum of 20 measurements: 10 internal replicates x 2 biological replicates).

### Quantification of relative RNA expression and relative TE

Relative RNA expression was obtained either from the original PTRE-seq study[24] or from the "no spacer" library in this study. Relative translation efficiency (TE) values were obtained from the original PTRE-seq study[24]. Briefly, relative RNA expression for each barcode was calculated as the ratio between RNA barcode count and DNA barcode counts. The normalized expression was set relative to the median normalized expression of BBBB control.

$$RNA\ expression = log_2\left[\frac{\frac{RNA}{DNA}}{median(\frac{RNA_{BBBB}}{DNA_{BBBB}})}\right]$$

Relative TE for each barcode was computed similarly, using RNA barcode counts observed in polysome fractions.

### Linear regression model

We adapted the linear regression strategy from the original PTRE-seq study to model reporter expression as a function of 3′ UTR regulatory elements[24]. The parameters of the model represent the impact of each regulatory module and their pairwise interactions on reporter expression relative to the "blank" BBBB reference reporter. The model formula is as follows:

$$E = \rho_0 + \sum_{i}^{4}\sum_{m}^{\{A,L,P,S\}} \rho_{i,m} I(i,m) + \sum_{i,j;i<j}^{4}\sum_{m,n}^{\{A,L,P,S\}} \rho_{i,j,m,n} I(i,m) I(j,n) \quad (1)$$

$$\text{where } I(i,m) = \begin{cases} 1\ if\ R_i == m \\ 0\ otherwise \end{cases}$$

where $E$ denotes relative reporter expression and $\rho$ denote the model parameters. Fitting was done using the Python Statsmodels:glm package with the family set as Gaussian and the link function set as identity. To avoid the confounding effect of splicing, we only fit the model to the 492 reporters exhibiting observed splicing fractions below 10%. The resulting parameters were used to estimate the relative expression of the full-length isoform of the remaining 133 reporters. Low expression controls and reporters with natural and synthetic let-7 sites were excluded from analysis. Fitting for the "no spacer" library was done using the same procedure.

### Gene expression modeling and transcription rate estimation

We model total expression of a reporter, as measured by barcode-specific sequencing, as a function of the total production rate of reporter mRNA ($\alpha_{tot}$), the degradation rates of the full-length ($\beta_f$) and spliced isoforms ($\beta_s$), and the fraction of mRNAs that are spliced ($s$) (Fig. 4B):

$$\frac{dE_{tot}}{dt} = \alpha_{tot} - \beta_s\ s\ E_{tot} - \beta_f(1-s)E_{tot} \quad (2)$$

In our model, $\alpha_{tot}$ is an aggregate quantity that takes into account transcription rate, polyadenylation, and nuclear export.

At steady state, $E_{tot}$ can be obtained as:

$$E_{tot} = \frac{\alpha_{tot}}{\beta_s\,s + \beta_f\,(1-s)} \quad (3)$$

which can be rewritten as:

$$E_{tot} = \frac{\frac{\alpha_{tot}}{\alpha_0}}{\frac{\beta_s}{\alpha_0}s + \frac{\beta_f}{\alpha_0}(1-s)} \quad (4)$$

where $\alpha_0$ is the production rate of the blank control. The value $\frac{\alpha_0}{\beta_f}$ equals the expected expression of the full-length isoform and can be estimated using the regression model described in Eq. 1. To obtain the expected expression of the spliced isoform, $\frac{\alpha_0}{\beta_s}$, we rearrange Eq. 4 to the form:

$$\frac{\beta_f}{\alpha_0}(1-s)E_{tot} = -\frac{\beta_s}{\alpha_0}sE_{tot} + \frac{\alpha_{tot}}{\alpha_0} \quad (5)$$

Reporters that are spliced using the same 5′ and 3′ splice sites will have identical spliced isoforms and hence identical degradation rates ($\beta_s$). Thus, for families of reporters with identical spliced isoforms, $\frac{\beta_s}{\alpha_0}$ can be obtained as the slope from the regression of $\frac{\beta_f}{\alpha_0}(1-s)E_{tot}$ against $sE_{tot}$. We identified two families of reporters that have sufficient variation in $sE_{tot}$ to permit this regression: GFP_B4, where reporters feature splicing from the primary GFP donor to a 4th position B module, and GFP_P4, where reporters are spliced at the GFP donor to a 4th position P acceptor. For each family, we tested fitting the model using reporters within different ranges of observed splicing fractions (10-70%, 10-80%, and 10-90%) to assess the robustness of the model. Fitting was done using the Python Statsmodels:*glm* function (family: Gaussian; link function: identity) using expression measurements from the original PTRE-seq study[24] and splicing fraction measurements made in the current study. Fitting the model using expression measurements from our current study yielded similar results.

In the RE-excision model, we assume that splicing does not affect reporter production, meaning $\frac{\alpha_{tot}}{\alpha_0} = 1$. Predicted $E_{tot}$ was then computed using the $\frac{\alpha_0}{\beta_f}$ and $\frac{\alpha_0}{\beta_s}$ values via Eq. 4, and the plotted lines of best fit were determined by regressing the predicted $E_{tot}$ against observed splicing fractions (Fig. 4A).

To estimate change in production rate in Fig. 4C, we rearrange Eq. 4 to solve for $\frac{\alpha_{tot}}{\alpha_0}$:

$$\frac{\alpha_{tot}}{\alpha_0} = \frac{\beta_s}{\alpha_0}sE_{tot} + \frac{\beta_f}{\alpha_0}(1-s)E_{tot} \quad (6)$$

### Prediction of cryptic splicing in published MPRAs

We adapted SpliceAI (v. 1.3.1) to predict splicing probabilities for the reporter sequences following the SpliceAI documentation[37]. For the PTRE-seq library, the input sequence for each reporter is a 1172-nt transcript, which begins with the eGFP coding sequence and ends at the poly-A signal. The sequence was padded with 5000 Ns on both sides to a total length of 11172 nts. SpliceAI outputs the probability that each position on the transcript is a donor or acceptor. The product of the donor and downstream acceptor with the highest probabilities is calculated as the splice probability and is used to predict whether a reporter is spliced or not. We explored other potential total splicing scores, including considering only donor or acceptor probabilities, sum of both probabilities, and Maximum Entropy scores[58]; these scores performed comparably or worse than the product at predicting PTRE-seq splicing efficiencies. We also evaluated the Pangolin[54] (v. 1.0.2) tool, which yielded similar results as SpliceAI.

For other MPRAs (*Griesemer*, *Zhao*, *Siegel*, and *Fu*), sequence information of the inserted regions and the backbone plasmids was retrieved directly from electronic supplementary data or from the authors. The transcript sequence for each reporter was then reconstructed beginning with the eGFP coding sequence and ending at the poly-A signal. Barcode sequences were added as 6 N(s) for *Griesemer*

MPRA and 8 N(s) for *Siegel* and *Zhao* MPRA. Splicing probabilities were then computed identically as for PTRE-seq.

To evaluate the potential for cryptic splicing sites in other reporter genes, we appended the strongly spliced ABAB 3′ UTR to the coding sequence of each alternative gene. Splicing probability prediction was then performed as described above.

### Motif enrichment analysis

XSTREME[59] was used to identify enriched motifs in predicted spliced reporters from the *Griesemer*, *Siegel*, *Zhao*, and *Fu* MPRAs. The sequences of reporters above 0.60 splicing probability were uploaded in fasta format to the XSTREME webserver (https://meme-suite.org/meme/tools/xstreme). Sequences of reporters below 0.30 splicing probability were uploaded as controls. All parameters were set to defaults. Motifs shown in Fig. 5E correspond to the most significantly enriched motif for each MPRA. The large majority of other significantly enriched sequences were also U-rich motifs.

### Validation of splicing in *Griesemer* MPRA reporters

10 pairs of alt/ref reporters (20 overall) were selected from the *Griesemer*[7] MPRA for validation: as negative controls, 2 pairs where both alleles were predicted to be not spliced; 3 pairs where the alleles were predicted to be differentially spliced; and 5 pairs where both alleles were predicted to be spliced with high probabilities. Synthesis and cloning were done by the Genetic Design and Engineering Center at Rice University. DNA sequences containing 3′ UTR segments, random 8 nt barcodes, and flanking adaptors matching the final assembled inserts in *Griesemer* were ordered as eBlocks (IDT) (Table S2). Each eBlock was amplified by PCR using the Griesemer_Amp_F and Griesemer_Amp_R primers (Table S1) and then cloned by Gibson Assembly (NEB) into BmtI/XbaI (NEB) digested pmirGLO:Δluc::gfp ΔAmp$^R$::Kan$^R$ vector (gift from Pardis Sabeti; Addgene #176640). The assembled vectors were transformed into 10-beta *E. coli* (NEB) by electroporation, expanded in LB broth supplemented with 50 µg/mL of kanamycin, and plasmids purified via Qiaprep Spin Miniprep Kit (Qiagen). Plasmid sequences were verified via nanopore whole plasmid sequencing using an MinION Mk1C sequencer.

Reporters were transfected using Lipofectamine 2000 (Thermofisher) according to the protocol published by *Griesemer*[7]. HEK293T cells were grown in a 6-well plate to 70% confluency. 3 *µ*g of DNA was combined with 9 *µ*L of Lipofectamine 2000 in a 300 *µ*L DMEM solution and added to a well. After 24 h, RNA was harvested using RNeasy kit. cDNA was prepared using Griesemer_R primer and SuperScript II reverse transcriptase per manufacturer protocol. 4 µL of the purified cDNA product was PCR amplified (98 °C for 30 s, 35 cycles: 98 °C for 10 s, 68 °C for 20 s, 72 °C for 30 s, and 72 °C for 2 min) using Griesemer_F and Griesemer_R primers (Table S1). PCR products were visualized on 1.5% agarose gel. Agarose gels were imaged using a Bio-Rad ChemiDoc MP Imaging System and band intensity was quantified using ImageJ (v.1.54p). Uncropped images are provided as Source Data.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

Sequencing data is available in the NCBI SRA under BioProject accession code PRJNA1116243. Processed sequencing data (expression and splicing quantification) are available at https://github.com/MustoeLab/Publications the under MIT license and have been archived at https://doi.org/10.5281/zenodo.15700943. Full plasmid maps for the 3 representative plasmids are available as Supplementary Data 1–3. Source data are provided with this paper.

## Code availability

Analysis scripts and Jupyter Notebooks are available at https://github.com/MustoeLab/Publications under MIT license and have been archived at https://doi.org/10.5281/zenodo.15700943.

## References

1. Corbett, A. H. Post-transcriptional regulation of gene expression and human disease. *Curr. Opin. Cell Biol.* **52**, 96–104 (2018).
2. Bartel, D. P. Metazoan MicroRNAs. *Cell* **173**, 20–51 (2018).
3. Van Nostrand, E. L. et al. A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).
4. Sibley, C. R., Blazquez, L. & Ule, J. Lessons from non-canonical splicing. *Nat. Rev. Genet.* **17**, 407–421 (2016).
5. Corley, M., Burns, M. C. & Yeo, G. W. How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. *Mol. Cell* **78**, 9–29 (2020).
6. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159–164 (2015).
7. Griesemer, D. et al. Genome-wide functional screen of 3′UTR variants uncovers causal variants for human disease and evolution. *Cell* **184**, 5247–5260.e19 (2021).
8. Siegel, D. A., Le Tonqueze, O., Biton, A., Zaitlen, N. & Erle, D. J. Massively parallel analysis of human 3′ UTRs reveals that AU-rich element length and registration predict mRNA destabilization. *G3 GenesGenomesGenetics* **12**, jkab404 (2021).
9. Zhao, W. et al. Massively parallel functional annotation of 3′ untranslated regions. *Nat. Biotechnol.* **32**, 387–391 (2014).
10. Fu, T. et al. Massively parallel screen uncovers many rare 3′ UTR variants regulating mRNA abundance of cancer driver genes. *Nat. Commun.* **15**, 3335 (2024).
11. Gallego Romero, I. & Lea, A. J. Leveraging massively parallel reporter assays for evolutionary questions. *Genome Biol* **24**, 26 (2023).
12. Muerdter, F., Boryń, M. & Arnold, C. D. STARR-seq — Principles and applications. *Genomics* **106**, 145–150 (2015).
13. Kreimer, A. et al. Massively parallel reporter perturbation assays uncover temporal regulatory architecture during neural differentiation. *Nat. Commun.* **13**, 1504 (2022).
14. Klein, J. C. et al. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat. Methods* **17**, 1083–1091 (2020).
15. Zheng, Y. & VanDusen, N. J. Massively Parallel Reporter Assays for High-Throughput In Vivo Analysis of Cis-Regulatory Elements. *J. Cardiovasc. Dev. Dis.* **10**, 144 (2023).
16. Ansseau, E. et al. Aberrant Splicing in Transgenes Containing Introns, Exons, and V5 Epitopes: Lessons from Developing an FSHD Mouse Model Expressing a D4Z4 Repeat with Flanking Genomic Sequences. *PLOS ONE* **10**, e0118813 (2015).
17. Cheng, Y. et al. Aberrant splicing events caused by insertion of genes of interest into expression vectors. *Int. J. Biol. Sci.* **18**, 4914–4931 (2022).
18. Akirtava, C., May, G. E. & McManus, C. J. False-positive IRESes from Hoxa9 and other genes resulting from errors in mammalian 5′ UTR annotations. *Proc. Natl. Acad. Sci.* **119**, e2122170119 (2022).
19. Baranick, B. T. et al. Splicing mediates the activity of four putative cellular internal ribosome entry sites. *Proc. Natl. Acad. Sci.* **105**, 4733–4738 (2008).
20. Královičová, J. & Vořechovský, I. Global control of aberrant splice-site activation by auxiliary splicing sequences: evidence for a gradient in exon and intron definition. *Nucleic Acids Res* **35**, 6399–6413 (2007).
21. Roca, X., Sachidanandam, R. & Krainer, A. R. Intrinsic differences between authentic and cryptic 5′ splice sites. *Nucleic Acids Res* **31**, 6321–6333 (2003).

22. Zarnack, K. et al. Direct Competition between hnRNP C and U2AF65 Protects the Transcriptome from the Exonization of Alu Elements. *Cell* **152**, 453–466 (2013).
23. Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nat. Rev. Genet.* **17**, 19–32 (2016).
24. Cottrell, K. A., Chaudhari, H. G., Cohen, B. A. & Djuranovic, S. PTRE-seq reveals mechanism and interactions of RNA binding proteins and miRNAs. *Nat. Commun.* **9**, 301 (2018).
25. Mayr, C. What Are 3′ UTRs Doing? *Cold Spring Harb. Perspect. Biol.* **11**, a034728 (2019).
26. Coolidge, C. J., Seely, R. J. & Patton, J. G. Functional analysis of the polypyrimidine tract in pre-mRNA splicing. *Nucleic Acids Res* **25**, 888–896 (1997).
27. Guo, L.-T. et al. Sequencing and structure probing of long RNAs using MarathonRT: a next-generation reverse transcriptase. *J. Mol. Biol.* **432**, 3338–3352 (2020).
28. Kesner, J. S. et al. Noncoding translation mitigation. *Nature* **617**, 395–402 (2023).
29. Lindeboom, R. G. H., Supek, F. & Lehner, B. The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat. Genet.* **48**, 1112–1118 (2016).
30. Zhang, Q. et al. The evolution mechanism of intron length. *Genomics* **108**, 47–55 (2016).
31. Choi, T., Huang, M., Gorman, C. & Jaenisch, R. A Generic Intron Increases Gene Expression in Transgenic Mice. *Mol. Cell. Biol.* **11**, 3070–3074 (1991).
32. Furger, A., Binnie, J. M. O., Alexandra, Lee, B. A. & Proudfoot, N. J. Promoter proximal splice sites enhance transcription. *Genes Dev* **16**, 2792–2799 (2002).
33. Nott, A., Meislin, S. H. & Moore, M. J. A quantitative analysis of intron effects on mammalian gene expression. *RNA* **9**, 607–617 (2003).
34. Uriostegui-Arcos, M., Mick, S. T., Shi, Z., Rahman, R. & Fiszbein, A. Splicing activates transcription from weak promoters upstream of alternative exons. *Nat. Commun.* **14**, 3435 (2023).
35. Barreau, C., Paillard, L. & Osborne, H. B. AU-rich elements and associated factors: are there unifying principles? *Nucleic Acids Res* **33**, 7138–7150 (2005).
36. Otsuka, H., Fukao, A., Funakami, Y., Duncan, K. E. & Fujiwara, T. Emerging Evidence of Translational Control by AU-Rich Element-Binding Proteins. *Front. Genet.* **10**, https://doi.org/10.3389/fgene.2019.00332 (2019).
37. Jaganathan, K. et al. Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535–548.e24 (2019).
38. Wu, H. et al. Combining full-length gene assay and SpliceAI to interpret the splicing impact of all possible SPINK1 coding variants. *Hum. Genomics* **18**, 21 (2024).
39. Sorek, R. The birth of new exons: Mechanisms and evolutionary consequences. *RNA* **13**, 1603–1608 (2007).
40. Aznarez, I. et al. A systematic analysis of intronic sequences downstream of 5′ splice sites reveals a widespread role for U-rich motifs and TIA1/TIAL1 proteins in alternative splicing regulation. *Genome Res* **18**, 1247–1258 (2008).
41. Zeng, Y. et al. Profiling lariat intermediates reveals genetic determinants of early and late co-transcriptional splicing. *Mol. Cell* **82**, 4681–4699.e8 (2022).
42. Erkelenz, S. et al. Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. *RNA* **19**, 96–102 (2013).
43. Llorian, M. et al. Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. *Nat. Struct. Mol. Biol.* **17**, 1114–1123 (2010).
44. Zheng, R. et al. hnRNPM protects against the dsRNA-mediated interferon response by repressing LINE-associated cryptic splicing. *Mol. Cell* **84**, 2087–2103.e8 (2024).
45. Chan, J. J. et al. Pan-cancer pervasive upregulation of 3′ UTR splicing drives tumourigenesis. *Nat. Cell Biol.* **24**, 928–939 (2022).
46. Luo, M. & Reed, R. Splicing is required for rapid and efficient mRNA export in metazoans. *Proc. Natl. Acad. Sci. USA.* **96**, 14937–14942 (1999).
47. Niwa, M., Rose, S. D. & Berget, S. M. In vitro polyadenylation is stimulated by the presence of an upstream intron. *Genes Dev* **4**, 1552–1559 (1990).
48. Uzonyi, A. et al. Exclusion of m6A from splice-site proximal regions by the exon junction complex dictates m6A topologies and mRNA stability. *Mol. Cell* **83**, 237–251.e7 (2023).
49. Guo, Y. E. et al. Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. *Nature* **572**, 543–548 (2019).
50. Ding, F. & Elowitz, M. B. Constitutive splicing and economies of scale in gene expression. *Nat. Struct. Mol. Biol.* **26**, 424–432 (2019).
51. Sample, P. J. et al. Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. *Nat. Biotechnol.* **37**, 803–809 (2019).
52. Niederer, R. O., Rojas-Duran, M. F., Zinshteyn, B. & Gilbert, W. V. Direct analysis of ribosome targeting illuminates thousand-fold regulation of translation initiation. *Cell Syst* **13**, 256–264.e3 (2022).
53. Rabani, M., Pieper, L., Chew, G.-L. & Schier, A. F. Massively parallel reporter assay of 3′UTR sequences identifies in vivo rules for mRNA degradation. *Mol. Cell* **68**, 1083–1094.e5 (2017).
54. Zeng, T. & Li, Y. I. Predicting RNA splicing from DNA sequence using Pangolin. *Genome Biol* **23**, 103 (2022).
55. Smola, M. J., Rice, G. M., Busan, S., Siegfried, N. A. & Weeks, K. M. Selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nat. Protoc.* **10**, 1643–1669 (2015).
56. Bushnell, B., Rood, J. & Singer, E. BBMerge – Accurate paired shotgun read merging via overlap. *PLoS ONE* **12**, e0185056 (2017).
57. Bushnell, B. BBMap: A Fast, Accurate, Splice-Aware Aligner. (2014).
58. Yeo, G. & Burge, C. B. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. *J. Comput. Biol.* **11**, 377–394 (2004).
59. Grant, C. E. & Bailey, T. L. XSTREME: Comprehensive motif analysis of biological sequence datasets. 2021.09.02.458722 Preprint at https://doi.org/10.1101/2021.09.02.458722 (2021).

## Acknowledgements

We thank Anonymous Reviewer 1 for their feedback and suggestions, which were critical for elucidating the palindromic insertion artifact. We thank D. Griesemer, J. Xue, P. Sabeti, W. Zhao, D. Siegel, D. Erle, X. Xiao, T. Fu, and K. Amoah for sharing sequences and data from their studies, and K. Cottrell for helpful discussions. We thank T. Westbrook for gifting us the HEK293T cell line, the Genetic Design and Engineering Center at Rice University for cloning services, and C. Chen and M. Goodell for the use of their electroporator. This work was supported by a CPRIT scholar award to A.M.M. by the Cancer Prevention and Research Institute of Texas (RR190054) and by the National Institutes of Health (R35 GM147010 to A.M.M.). S.D. and C.F.J. are supported by NIH GM112824 and Chan Zuckerberg Neurodegeneration Initiative Award. The Genomic and RNA Profiling Core at Baylor College of Medicine is supported by P30 Cancer Center Support Grant (NCI-CA125123). The Genetic Design and Engineering Center at Rice university is supported by CPRIT RP210116.

## Author contributions

K.D., C.F.J., S.J., and A.M.M. designed experiments. K.D. and C.F.J. performed experiments. K.D. performed analyses. K.D. and A.M.M. wrote the manuscript with input from C.F.J. and S.J.

## Competing interests

A.M.M. is an advisor to and holds equity in RNAConnect, which holds patent rights to MarathonRT, and has consulted for Ribometrix. The remaining authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-62000-9.

**Correspondence** and requests for materials should be addressed to Anthony M. Mustoe.

**Peer review information** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.