Received: 18 October 2021 | Revised: 2 March 2022 | Accepted: 2 April 2022

DOI: 10.1002/humu.24379

RESEARCH ARTICLE

Human Mutation HGVS HUMAN GENOME VARIATION SOCIETY WILEY

Bioinformatics detection of modulators controlling splicing factor-dependent intron retention in the human brain

Steven X. Chen[1,2] | Ed Simpson[1,2] | Jill L. Reiter[1,2] | Yunlong Liu[1,2]

[1]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, Indiana, USA

[2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana, USA

Correspondence

Jill L. Reiter and Yunlong Liu, Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA.
Email: jireiter@iupui.edu and yunliu@iu.edu

Funding information

Indiana Clinical and Translational Sciences Institute: Eli Lilly-Stark Neurosciences Pre-Doctoral Research Fellowship in Neurodegeneration

Abstract

Alternative RNA splicing is an important means of genetic control and transcriptome diversity. However, when alternative splicing events are studied independently, coordinated splicing modulated by common factors is often not recognized. As a result, the molecular mechanisms of how splicing regulators promote or repress splice site recognition in a context-dependent manner are not well understood. The functional coupling between multiple gene regulatory layers suggests that splicing is modulated by additional genetic or epigenetic components. Here, we developed a bioinformatics approach to identify causal modulators of splicing activity based on the variation of gene expression in large RNA sequencing datasets. We applied this approach in a neurological context with hundreds of dorsolateral prefrontal cortex samples. Our model is strengthened with the incorporation of genetic variants to impute gene expression in a Mendelian randomization-based approach. We identified novel modulators of the splicing factor SRSF1, including *UIMC1* and the long noncoding RNA *CBR3-AS1*, that function over dozens of SRSF1 intron retention splicing targets. This strategy can be widely used to identify modulators of RNA-binding proteins involved in tissue-specific alternative splicing.

KEYWORDS

alternative splicing regulation, bioinformatics, brain, intron retention, Mendelian randomization, SRSF1

1 | INTRODUCTION

Alternative RNA splicing (AS) promotes transcriptome diversity and is especially widespread in the brain, where it is necessary for the differentiation of neurons and may have played a role in the development of the vertebrate brain (Barbosa-Morais et al., 2012; Merkin et al., 2012; Yeo et al., 2004). In addition to selective inclusion or exclusion of certain exons, intron retention (IR) events can also be an important AS mechanism. Transcripts harboring unspliced introns may serve as stable intermediates that can be stored temporarily until the appropriate signal is received. In neurons, transcripts with select retained introns accumulate in the nucleus and undergo splicing upon cellular activation to rapidly mobilize a pool of mRNAs at the precise time they are needed (Mauger et al., 2016). While AS in the brain is associated with aging and neurodegeneration (Raj et al., 2018; Tollervey et al., 2011), the mechanisms regulating AS pathways have yet to be fully explored.

Alternative splicing is regulated by *cis*-acting elements (e.g., 5′-donor sites, 3′-acceptor sites, branch sites, and polypyrimidine tracts) and *trans*-acting splicing factors that exert combinatorial

control of short, degenerate RNA motifs known as exonic or intronic splicing enhancers or silencers (Black, 2003; Z. Wang & Burge, 2008; Xiao et al., 2007). These elements together constitute the splicing code and allow for tissue- and cell-specific RNA splicing (Fu & Ares, 2014). The two major classes of splicing factors are heterogenous nuclear ribonucleoproteins (hnRNPs) and serine/arginine-rich (SR) proteins. In addition to the role of splicing factors, AS programs may substantially change in the presence or absence of modulating proteins that target RNA-binding proteins (RBPs). Other regulatory layers, including transcription and chromatin, may also modulate AS, as do posttranscriptional and signaling pathways.

Current understanding of AS regulation remains limited. The ENCODE consortium has profiled over 150 RBPs to generate RBP splicing regulatory maps and has enabled researchers to query RBP binding and activation/repression of splicing sites (Yee et al., 2019). Much of our understanding of AS derives from proteins characterized through cross-linked immunoprecipitation (CLIP) studies (Ule et al., 2005). For example, we used CLIP followed by next-generation sequencing (CLIP-seq) to demonstrate that overexpression of HNRNPA1 promotes recognition of noncanonical 3′-splice sites by splicing factor U2AF2 (Howard et al., 2018). Additional approaches include attempts to systematically characterize functional splicing networks (Papasaikas et al., 2015). However, while Papasaikas et al. were able to knock down over 200 splicing factors, their evaluation was limited in scope to 36 splicing events involved in cell proliferation or apoptosis.

Overexpression or knocking down individual splicing regulators followed by both CLIP-seq to analyze RBP binding profiles and RNA-seq to measure splicing levels is time-consuming and costly. Apart from coexpression networks and motif discovery, there exists no exhaustive transcriptome-wide computational method to dissect splicing networks. Thus, a scalable, general-purpose computational screen is needed to identify tissue- and cell type-specific modulators of splicing factors and their associated AS modifications. Similar methods have been used to identify modulators of transcription factors, but no one has identified modulators of splicing to date (Babur et al., 2010; K. Wang et al., 2009).

To address the limited understanding of AS regulation, we developed a bioinformatics approach to identify modulators of splicing factors using the biological perturbation found in RNA-seq datasets. The varied gene expression levels in each sample serve as a proxy for experimental modifications. The relationship of a splicing factor with its target events can generally be understood in the way that as the expression level of an RBP changes, corresponding changes in splicing outcomes occur. This suggests that the introduction (or withdrawal) of a modulator affects the existing RBP-splicing relationship such that the role of the RBP can be enhanced, attenuated, or reversed. Modulator is used here as a general term that may represent another splicing factor, a signaling protein, a noncoding RNA, or other genetic element. Simply, an RBP's underlying splicing function is modulated by an external factor.

We applied our method to study modulators of the serine and arginine rich splicing factor 1 (SRSF1) in the human brain. SRSF1 is a prototypical SR protein that functions in constitutive and alternative splicing. While not part of the core spliceosome, SRSF1 is essential for splicing and also plays roles in nonsense-mediated mRNA decay (NMD), mRNA export, and translation (Aznarez et al., 2018; Li & Manley, 2005). CLIP-seq analysis in human cells revealed widespread preferential binding of SRSF1 to exonic regions and a consensus binding motif of GAAGAA (Sanford et al., 2009). SRSF1 generally promotes exon definition and the use of proximal alternative 5′ splice sites or 3′ splice sites in a concentration-dependent manner, in part through recognition of degenerate exonic splicing enhancer (ESE) sequence elements on its pre-mRNA targets.

The advent of large-scale consortia RNA-seq datasets have recently allowed us to interrogate AS at a transcriptome-wide level. The use of multiple sequencing modalities in these consortia provides additional dimensions of data to analyze. In our case, we apply a Mendelian randomization (MR) approach that allows us to bypass confounding and environmental effects in conventional association studies. Single nucleotide polymorphisms derived from whole-genome sequencing data are used to impute gene expression, and results are verified with RNA-seq-derived gene expression. Our method is scalable and with the incorporation of genetic variants, prioritizes causal relationships. While this study discusses the modulators of SRSF1 IR splicing targets in a neurological context, we also provide a general framework to identify modulators of other RBPs.

2 | MATERIALS AND METHODS

2.1 | RNA sequencing data

Brain dorsolateral prefrontal cortex (DLPFC) RNA sequencing data from 890 samples in BAM format were downloaded from the Religious Order Study and Memory and Aging Project (ROSMAP) (Bennett et al., 2018). The nongapped aligner Bowtie was used to align reads to the transcriptome reference assembly GRCh37 (Langmead et al., 2009). Gene expression counts were called using featureCounts (Subread release 2.0.0) with GENCODE Release 19 (GRCh37.p13) reference annotation (Frankish et al., 2019; Liao et al., 2014). Lowly expressed genes were removed using the default filterByExpr function, and gene counts were normalized by library size using TMM normalization through edgeR (R version 4.0.2, Bioconductor version 3.11). Downstream analyses used genes expressed in counts per million (CPM) mapped reads.

From the CommonMind Consortium (CMC), brain DLPFC RNA sequencing data from 991 samples in BAM format were downloaded (Hoffman et al., 2019). Sequence reads were aligned to reference assembly GRCh38 using STAR 2.7.2a (Dobin et al., 2013). The reference assembly version differed from ROSMAP because the raw data from each study corresponded to different genome assemblies. Gene expression counts were called using featureCounts (Subread release 2.0.0) with GENCODE Release 33 (GRCh38.p13) reference annotation (Frankish et al., 2019; Liao et al., 2014). Genes were normalized in the same manner as above.

2.2 | Whole-genome sequencing data

Whole-genome sequencing (WGS) data from 1200 DLPFC samples were downloaded from ROSMAP. Of the available samples, 791 had matching RNA-seq data. Genomic variants were then used to impute the individual-level gene expression with PrediXcan (MetaXcan v0.7.3, https://github.com/hakyimlab/MetaXcan) (Gamazon et al., 2015). Transcription prediction weights were computed using an elastic net prediction model built on matched genotyped and RNA-seq DLPFC samples from CMC (Huckins et al., 2019).

To accurately identify single nucleotide polymorphisms (SNPs) from WGS data, genetic variants were mapped to corresponding RS IDs from dbSNP 151 using a reference table from GTEx Analysis Release V8 (https://gtexportal.org/home/datasets). PrediXcan models were built upon inverse-rank normalized expression data, where a negative value means that an individual is predicted to have lower expression values than expected in the model population.

2.3 | Quantification of alternative splicing

Splicing events were called from GENCODE Release 19 (GRCh37.p13) reference annotation using a script modified from the rMATS program (Shen et al., 2014). We then applied the splicing annotation to the RNA-seq data with the filters: ≥10 reads supporting the event and ≥1 read supporting the exclusion isoform. The percent spliced-in (PSI) values were calculated for each event in every sample as such:

$$\text{PSI} = \frac{\#\text{ inclusion reads}}{\#\text{ inclusion reads} + 2 \times \#\text{ exclusion reads}}.$$

Only events with PSI values with an interquartile range (IQR) ≥ 0.1 were included for downstream analysis. Splicing events from CMC data were identified using a GENCODE Release 33 (GRCh38.p13) reference annotation. The number of splicing events before filtering did not differ using this annotation.

2.4 | SRSF1 binding sites

Crosslinking and immunoprecipitation sequencing (CLIP-seq) data for the SRSF1 were downloaded from the Encyclopedia of DNA Elements (ENCODE) Consortium data portal (https://www.encodeproject.org/). CLIP-seq data were from HepG2 (ENCSR989VIY, 1781 peaks) and K562 (ENCSR432XUP, 2155 peaks) cell lines. Peaks were pre-processed using standard ENCODE pipelines, including removal of blacklisted regions and irreproducible discovery rate filtering with two isogenic replicates. Peak coordinates were intersected with alternative splicing annotations. Splice events were counted if peak coordinates were within 300 base pairs upstream from the proximal exon and 300 base pairs downstream of the distal exon.

2.5 | Modulator identification

To identify modulators of RBP-mediated splicing, we required inputs representing each factor. We assessed the input variables in a generalized linear model with the identity link function:

$$Y_t = \beta_o + \beta_1 X_r + \beta_2 X_m + \beta_3 X_r X_m + \varepsilon, \quad (1)$$

where X_r is the gene expression of the RBP, X_m is the expression level of a candidate modulator, and Y_t is the PSI value of a target splicing event. Nonzero outcomes of β_3 represent interactions between the gene expression levels of the modulator and the RBP on the given splicing event. Accurate estimation of β_3 requires large datasets. To infer β_3 in an unbiased, nonparametric manner, we adapted the gene expression modulation (GEM) algorithm (Babur et al., 2010). A nonparametric approach eliminates the assumption of a probability distribution (e.g., normal distribution) among the predictor variables. To parameterize the inputs, they are rank ordered and divided into tertiles. The top and bottom tertile values are then transformed into 1 and 0, respectively, while the middle 33% of values are discarded:

$$x' = \begin{cases} 1 & \text{if } x \text{ is in upper tertile} \\ \text{null} & \text{if } x \text{ is in middle tertile} \\ 0 & \text{if } x \text{ is in lower tertile} \end{cases} \quad (2)$$

After discretization, each model falls into one of the 27 possible bins based on the ternary state of X_r', X_m', and Y_t'. Only the eight bins where no variables have a null value are considered. Observed frequencies of values in each of the eight bins are then used to calculate the proportions of $Y_t' = 1$ for each combination of states of X_r' and X_m'. Estimation and significance of the coefficients β_o, β_1, β_2, β_3 were calculated as published previously (Y. Wang et al., 2020). We evaluated significant nonzero values of β_3 ($p < 0.05$) to denote modulator effects on RBP-mediated splicing. False discovery rate (FDR) was calculated by the Benjamini–Hochberg method (Benjamini & Hochberg, 1995).

3 | RESULTS

3.1 | Overview of the model for identifying modulators of splicing activity

RBP activity is estimated by the splicing levels of its target events, which are measured as percent spliced-in (PSI). In our model, we hypothesize that the splicing activity of an RBP will change with respect to the expression level of putative modulators. Intuitively, if an RBP regulates the splicing outcome of a target event, we expect the expression levels of the RBP and the PSI levels of the target events would be correlated (positively or negatively) across multiple samples. In addition, we expect that such a correlation may be dependent on the expression level of a modulator. Figure 1a depicts an example of how high expression of a modulator might affect the correlation between RBP expression and PSI of the target gene:

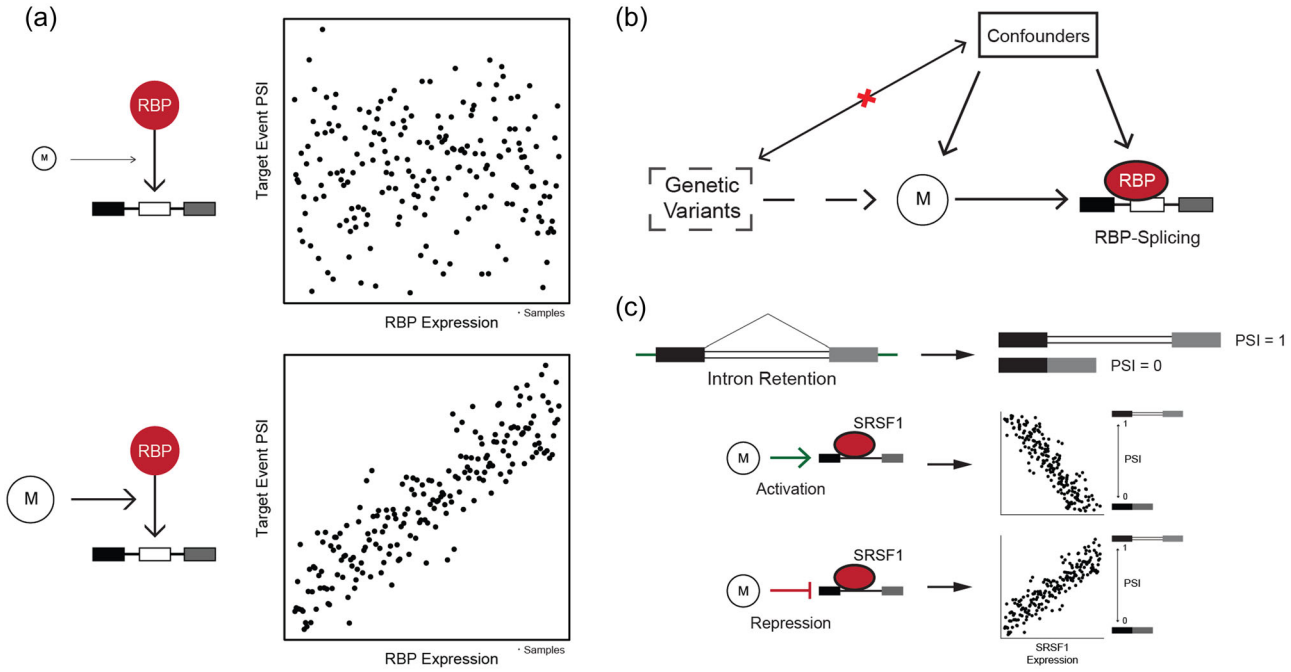

FIGURE 1 Modulation of the RNA-binding protein (RBP)-splicing relationship. (a) Schematic illustrating how modulator expression levels influence RNA splicing. RBP activity on splicing levels of its target genes, as measured by percent spliced-in (PSI), depends on the expression levels of a modulator (M). When an RBP binds a target splicing event, we expect to observe a relationship between the expression levels of the RBP and the PSI levels of the target events across multiple samples (upper panel). In this scenario, the expression of a modulator is low, and the correlation is stochastic. In the presence of high modulator expression (lower panel), the RBP-splicing relationship is strengthened or activated. (b) Schematic representation of the Mendelian randomization approach. Causal effects of modulators on the RBP-splicing target relationship can be identified while avoiding confounders. From matched genotype data, genetic variants within 1 MB of the candidate modulator are used to impute their gene expression. The approach uses only the genetically regulated component of gene expression to avoid environmental and feedback effects on the transcriptome. It is used to determine causal modulators of the RBP-splicing target association rather than modulators that may be correlated in expression to one another. (c) Schematic of the expected results of a modulator on serine and arginine rich splicing factor 1 (SRSF1) splicing activity. SRSF1 is known to promote splicing out of introns. In this study, a PSI value of 0 indicates that the intron has been spliced out of all transcripts, and a PSI value of 1 denotes that the intron is retained in all transcripts. SRSF1 and PSI values are negatively correlated in the absence of any external factors. A positive modulator (activator) would strengthen this negative correlation, and a negative modulator (repressor) would reverse this correlation

No correlation is observed when a modulator is expressed at a low level, while a positive correlation is detected when modulator expression is high. Many scenarios exist for how a modulator might influence the relationship between RBP expression levels and PSI of target genes, but the common element is that they are dependent upon the expression level of the modulator.

Our strategy to evaluate candidate modulators integrates four categories of inputs: a large-scale RNA-seq data set paired with genotype data, an RBP of interest, corresponding RBP-binding sites, and alternative splicing annotations. The RNA-seq data were used to calculate the expression levels of the RBP and the PSI values of the target events. The RBP-binding sites were derived from CLIP-seq data in the public domain, and splicing annotations were downloaded from GENCODE. The expression levels of the candidate modulators were imputed from the genotypes of the flanking regions in 791 dorsolateral prefrontal cortex profiles from the ROSMAP study (Bennett et al., 2018) using a prediction model built from 538 CommonMind Consortium samples with matched genotype and RNA-seq data (Huckins et al., 2019).

The candidate modulators were identified by applying the generalized linear model shown in Equation (1). A potential complication to an association-based approach is that any analysis based on gene expression levels of the candidate modulators does not guarantee a causal relationship. For example, expression levels of several top candidate modulators were strongly correlated amongst themselves, indicating they were likely correlated in expression to a true causal modulator (Figure S1). Similar quantitative relationships will be observed for all genes whose expression levels are correlated with a causal modulator. To address this issue, we adopted a MR approach by substituting gene expression levels derived from RNA-seq data with imputed gene expression calculated from the genotype within 1 MB upstream and downstream of the candidate modulator gene locus (Figure 1b). While RNA-seq data may provide more accurate measurements of gene expression, it is subject to confounders, cofactors, and feedback effects. The MR-based approach utilizes only the genetically regulated component of gene expression for candidate modulators; that is in Equation (1), X_m is now the imputed gene expression level of a candidate modulator.

Moreover, the use of DNA-level sequence information to impute gene expression also bypasses any environmental influences so that only causal modulators are inferred.

In the 791 ROSMAP WGS samples, PrediXcan was used to impute the expression levels of 10,677 genes. The PrediXcan-imputed expression output ranged from −1.9 to 2.7. Both prediction models and imputation output were measured as rank-based inverse normal transformed values. The mean Spearman correlation (ρ) between the imputed and RNA-seq-derived expression levels for 4007 genes was 0.16 ($p < 0.05$). We also found that imputation using ROSMAP genotype array data was comparable to using WGS with a mean Spearman correlation of 0.19 ($p < 0.05$) for 7974 genes in 582 samples (data not shown). Our findings compare favorably to the benchmark for cross-validation ($R^2 > 0.10$) in transcriptomic imputation methods (Gamazon et al., 2015). We used the imputed WGS expression levels in this study since the WGS data was higher quality and there were more samples with matched RNA-seq data. Since our model requires data to be highly variable among samples, we applied a filter requiring the imputed expression standard deviation be greater than 10%. Additionally, we ran the model concurrently using RNA-seq-derived expression levels and applied the same filters. Candidate modulators denoting 3014 genes remained in both sets (Figure S2).

3.2 | Mapping of SRSF1-targeted IR events

To test our hypothesis, we used our model in the genome-wide identification of modulators of the well-studied SRSF1 splicing factor. Because *SRSF1* expression is negatively correlated to IR levels, and shRNA knockdown of *SRSF1* increased the number of retained introns (Ullrich & Guigó, 2020), we presumed that SRSF1 promotes mRNA isoforms with the intron spliced out. That is, *SRSF1* gene expression levels and PSI values of the targets should be negatively correlated in the absence or low expression of any modulators. Using this assumption, a positive SRSF1 modulator (i.e., activator) would be expected to strengthen the negative correlation of SRSF1 with IR, whereas a negative modulator (i.e., repressor) would reverse a negative SRSF1-IR association (Figure 1c). For any given modulator-SRSF1-IR triplet, the interaction term β_3 from Equation (1) indicates the direction and effect of interaction of the modulator on SRSF1 function. Accordingly, positive or negative modulators can be determined by considering the correlation between SRSF1 and IR levels in aggregate.

To map the SRSF1-targeted IR events, we first annotated SRSF1 binding regions by integrating enhanced CLIP (eCLIP) sequencing data from ENCODE with annotated alternative splicing events (GENCODE release 19, GRCh37.p13). In total, there were 38,133 skipped exon, 5558 IR, 7728 alternative 3′ splice site, 4927 alternative 5′ splice site, and 2363 mutually exclusive exon events. By limiting our study to IR events, we identified 411 SRSF1-binding IR events, which represented 7% of annotated IR events and 0.7% of total annotated splicing events. PSI levels were calculated for each of the 411 SRSF1-binding IR events in the 791 ROSMAP RNA-seq samples. A PSI value of 0 indicates that all transcripts were missing the intron, and a PSI value of 1 denotes that all transcripts that include the neighboring exons retained the intron. To ensure PSI levels of splice events had enough variability at the population level, we required that the interquartile range (IQR) of the PSI values should be ≥ 10%. After applying this filter, 198 IR events remained (Table S1).

3.3 | Modulators of SRSF1-IR events

We generated causal gene-based associations between candidate modulators and SRSF1-mediated intron splicing. Imputed modulator expression, RBP expression, and target splicing PSI values were used as inputs for Equation (1). Figure 2 provides an overview of how the data were integrated. We interrogated 3014 candidate modulators on the set of 198 SRSF1-mediated IR events and identified 27,302 interactions between modulator expression and SRSF1 expression (β_3 $p < 0.05$) (Table S2). Figure 3a shows a histogram of the number of modulators for each of the affected IR targets.

To prioritize modulator-SRSF1-targets for further analysis, we reasoned that biologically relevant modulators were likely to affect multiple splicing targets. We presumed that candidate modulators might influence 9 of the 198 (5%) IR events by random chance. After applying Fisher's exact test to determine if a candidate modulator affected more IR events than random, we prioritized 82 candidate modulators that regulated at least 31 targets at FDR < 0.01 (Figure 3b).

The above analyses were done on modulators whose expression values were imputed from genetic data. To ensure the relationship derived from the imputed gene expression levels was maintained at the RNA level. The same set of candidate modulators and splicing targets were analyzed using the model with only RNA-seq gene expression. For each candidate modulator, we required 60% of significant targets identified initially using imputed gene expression be replicated at FDR < 0.05 and with the same directional effect as determined by the sign of the interaction term. This analysis resulted in 13 modulators of SRSF1 IR activity (Figure 3C, Table 1).

3.4 | Validation of modulators from CommonMind Consortium data

To verify our SRSF1 modulator results, we tested our model using CommonMind Consortium RNA-seq data from 985 DLPFC samples. After imputing gene expression with PrediXcan, we found a Spearman correlation (ρ) of 0.16 ($p < 0.05$) for 7943 genes. Splicing PSI, *SRSF1* expression, and candidate modulator expression were computed from the RNA-seq data and inputted into Equation (1). We identified 212 SRSF1-mediated IR events with PSI values that fit the constraint of IQR ≥ 0.10 (Table S3). Of these, 151 events were originally tested in the ROSMAP data. Differences in the number of

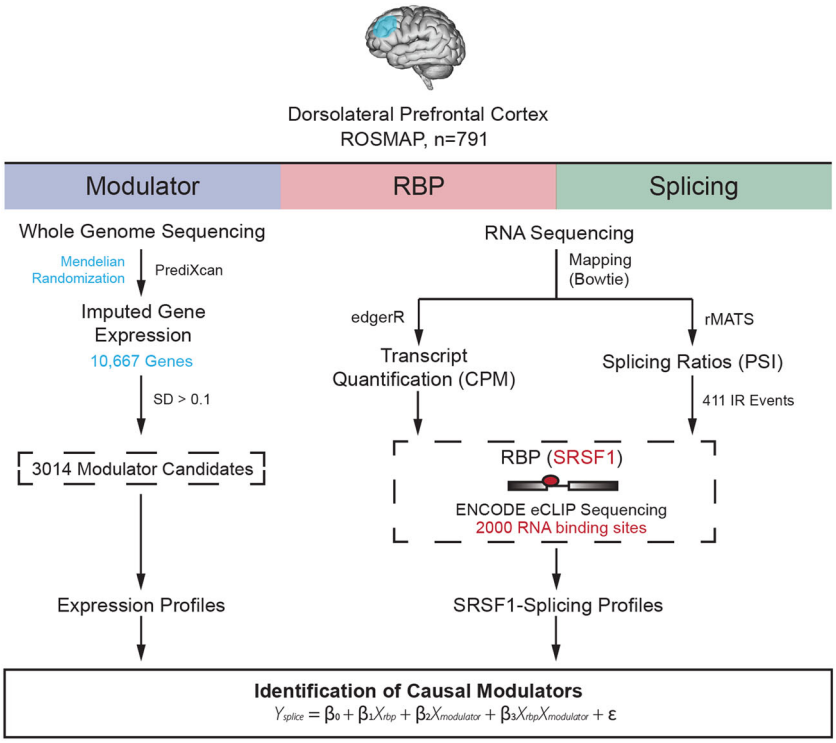

FIGURE 2 Integration of whole-genome sequencing (WGS), RNAseq, and CLIPseq data to identify splicing modulators. Imputed gene expression profiles of the dorsolateral prefrontal cortex were derived using PrediXcan from Religious Order Study and Memory and Aging Project (ROSMAP) WGS data. Splicing profiles were computed from matched RNA-seq data from the same ROSMAP samples. CLIP-seq was used to identify splicing events targeted by the RNA-binding protein of interest. For serine and arginine rich splicing factor 1 (SRSF1), each candidate modulator and intron retention splicing target (dashed boxes) were tested according to Equation (1) shown at the bottom. A similar pipeline was used on 538 samples from the CommonMind Consortium to validate our model

SRSF1-mediated IR events between the two studies likely arose because of data set differences in both SRSF1-binding and alternative splicing.

We found that 8 of the 13 modulators from Table 1 were recapitulated in the CommonMind Consortium data set, including UIMC1, CBR3-AS1, LRRC27, PRIMPOL, POLDIP2, ALG8, PMS2, and TRANK1. In addition, these modulators regulated a higher proportion of IR targets than were found in the ROSMAP analysis (Figure 4). These findings demonstrate that our model can be replicated in independent samples from the same tissue type.

3.5 | UIMC1 represses SRSF1 splicing activity

To confirm that the predicted modulators influenced SRSF1-mediated splicing, we chose to examine ubiquitin interaction motif containing 1 (UIMC1), also known as BRCA1-A Complex Subunit RAP80. UIMC1 is a ubiquitin-binding protein that plays a central role in the BRCA1 complex to repair DNA lesions. SRSF1 is a known proto-oncogene that influences splicing of the *BRCA1* tumor suppressor gene, resulting in variants that lack important functional domains (Karni et al., 2007; Raponi et al., 2014; Silipo et al., 2015). To test whether the correlation between *SRSF1* expression and IR PSI was dependent on *UIMC1* expression levels, we divided the ROSMAP samples into high and low groups based on imputed *UIMC1* expression levels (top and bottom tertiles). With UIMC1 selected as a potential modulator of SRSF1 in our model, we found that 52 of 198 SRSF1-targeted IR events were affected. We calculated the Spearman correlation coefficients between *SRSF1* expression and IR PSI for each of these IR events in the high and low *UIMC1* expression groups. There was a stronger correlation in the high *UIMC1* group for 48 events with a mean difference in the correlation coefficient of 0.25 (Figure 5a). Of these 52 events, 44 were confirmed when *UIMC1* gene expression was based on RNA-seq data. These results indicate that when UIMC1 activity is high, increasing *SRSF1* expression leads to more retained introns, and implies a reduction in SRSF1 splicing efficiency.

To further analyze the effect of *UIMC1* expression on specific SRSF1 targets, we examined the correlation between *SRSF1* expression and IR PSI in ROSMAP samples with either high or low *UIMC1* imputed expression. We selected two targets, *MRNIP (C5orf45) intron 4* (chr5:179264275-179267959, minus strand) and *DDX39A intron 6* (chr19:14520553-14521146, minus strand), as examples. Our analysis indicated that *UIMC1* expression modulated SRSF1-mediated splicing of *MRNIP* (Figure 5b). In the *UIMC1*-low group, *SRSF1* expression and *MRNIP* PSI showed little discernable correlation ($\rho = -0.032$). However, in the *UIMC1*-high group, *SRSF1* expression was positively correlated with *MRNIP* PSI ($\rho = 0.337$), which is consistent with IR and repression of normal SRSF1 function. A similar relationship was observed with *DDX39A* where we observed that in the high *UIMC1* expression group, SRSF1 expression was positively correlated with *DDX39A intron 6* inclusion (Figure 5c). Taken together, we conclude that UIMC1 functions as a repressor of SRSF1 splicing.

3.6 | *CBR3-AS1* activates SRSF1 splicing activity

While all but one of the modulators in Table 1 are protein-coding genes, *CBR3-AS1* is a lncRNA for which very little is known other than

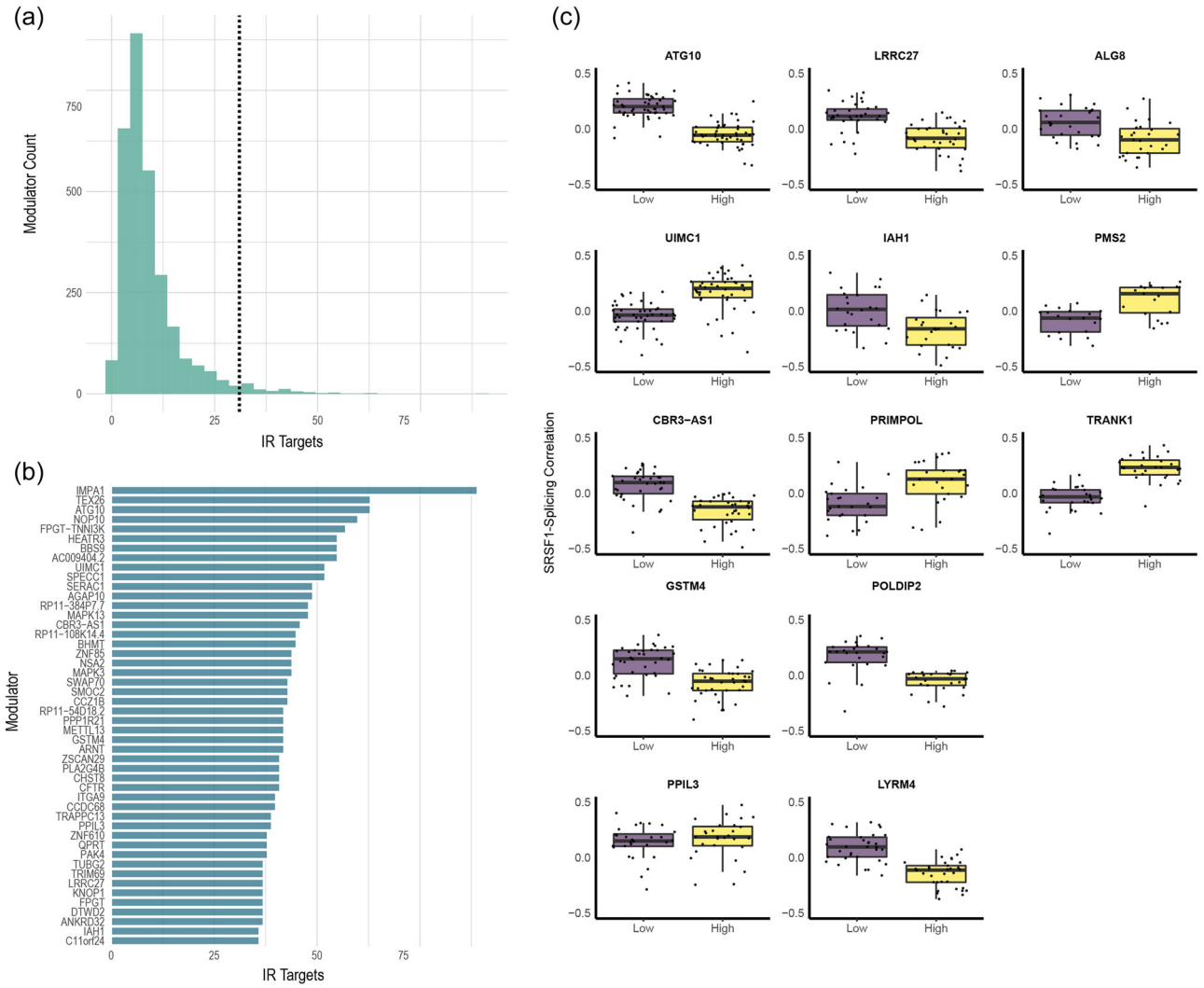

FIGURE 3 Modulator expression changes serine and arginine rich splicing factor 1 (SRSF1)-intron retention correlation. (a) A histogram of the number of modulators for each of the SRSF1 splicing events (β_3 $p < 0.05$) identified in the Religious Order Study and Memory and Aging Project (ROSMAP) cohort samples ($n = 198$, bin size = 3). Vertical dashed line indicates the number of affected intron retention (IR) targets that would be expected by chance for any modulator. A total of 82 modulators affected IR targets above this threshold. (b) Predicted SRSF1 modulators and the number of IR targets. Chart showing the 31 modulators remaining after filtering using Fisher's exact test and false discovery rate (FDR) < 0.01 on 5% of the possible SRSF1 targets. (c) Boxplots of 13 modulators that were found to significantly alter the SRSF1-percent spliced-in (PSI) correlation between low (purple) and high (yellow) activity states. Modulator activity is represented by their gene expression levels as imputed from genetic variants. Modulator expression was verified by RNA-seq data and SRSF1 modulators were selected when >60% of its targets were confirmed. *y*-axes show Spearman correlation coefficients for SRSF1-splicing event comparisons. Each datapoint represents an IR target that was modulated. While data in this graph are paired, visualization does not reflect pairing. Boxes show the median and first and third quartiles; whiskers extend from the hinges to 1.5 × interquartile range (IQR)

that it might be an important regulator in certain cancers (Jin et al., 2017; Song et al., 2018). One proposed lncRNA function is that they serve as sponges that sequester RNA binding proteins (Salmena et al., 2011). To test whether lncRNAs also have the potential to modulate SRSF1 targets, we examined whether the correlation between *SRSF1* expression and IR PSI was dependent on *CBR3-AS1* expression levels. When *CBR3-AS1* was selected as the potential modulator, our model predicted that 46 of 198 SRSF1-targeted IR events were modulated (Table 1). The Spearman correlations between *SRSF1* expression and IR PSI for 45 of these 46 IR events were stronger in the low compared to the high *CBR3-AS1* expression groups with a mean correlation difference of −0.23 (Figure 6a). Of the 46 events, 34 were confirmed when *CBR3-AS1* expression was based on RNA-seq data. These results indicate that when *CBR3-AS1* activity is high, increasing SRSF1 expression leads to fewer retained introns, and implies an increase in SRSF1 splicing activity.

We also analyzed the effect of *CBR3-AS1* on the correlation between *SRSF1* expression and IR in two targets, *CHKB intron 1+2* (chr22:51020177-51021283, minus strand) and *RFNG intron 4+5* (chr17:80007552-80008431, minus strand). We found that when

TABLE 1 Modulators affecting the serine and arginine rich splicing factor 1 (SRSF1)-intron retention target relationship

Modulator	Type	Mode of action	Targets	Confirmed targets	*p*-value	FDR
ATG10	Protein coding	Activator	63	47	2.85E-13	2.83E-10
UIMC1	Protein coding	Repressor	52	44	6.71E-10	2.00E-07
CBR3-AS1	lncRNA	Activator	46	34	3.34E-08	6.65E-06
GSTM4	Protein coding	Activator	42	36	3.94E-07	4.20E-05
PPIL3	Protein coding	Repressor	39	28	2.33E-06	1.93E-04
LRRC27	Protein coding	Activator	37	34	7.30E-06	4.74E-04
IAH1	Protein coding	Activator	36	24	1.28E-05	7.96E-04
PRIMPOL	Protein coding	Repressor	35	25	2.22E-05	1.33E-03
POLDIP2	Protein coding	Activator	34	25	3.83E-05	2.00E-03
LYRM4	Protein coding	Activator	34	33	3.83E-05	2.00E-03
ALG8	Protein coding	Activator	32	27	1.11E-04	4.34E-03
PMS2	Protein coding	Repressor	31	19	1.85E-04	6.75E-03
TRANK1	Protein coding	Repressor	31	27	1.85E-04	6.75E-03

Note: Modulators were selected according to the number of SRSF1-intron retention targets they impacted. The modulation was confirmed independently using RNA-seq gene expression values of the modulator. Only modulators with more than 60% confirmed targets and false discovery rate (FDR)-corrected (Benjamini-Hochberg) $p < 0.01$ are considered. *p*-values were calculated using Fisher's exact test for 5% significance among 198 SRSF1-intron retention targets.

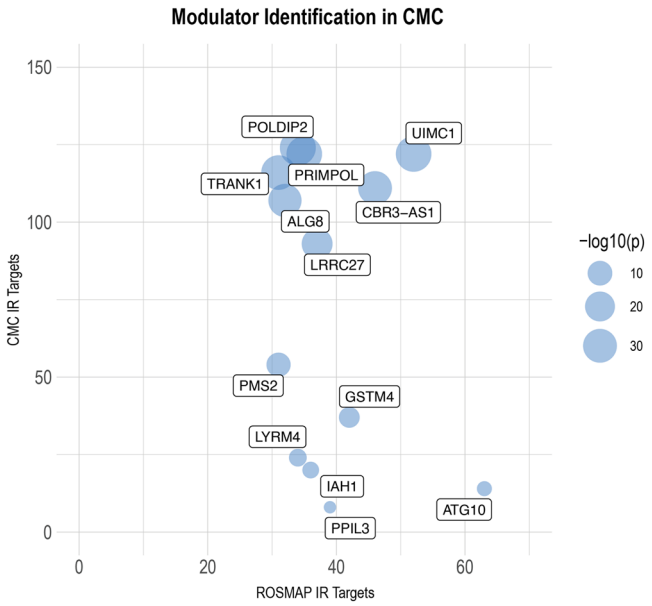

FIGURE 4 Validation of serine and arginine rich splicing factor 1 (SRSF1) modulators using CommonMind Consortium brain data. Modulators of SRSF1 intron splicing activity were confirmed using CommonMind Consortium (CMC) RNA-seq data from 985 dorsolateral prefrontal cortex samples. The 13 modulators prioritized from the ROSMAP data (Table 1) were tested independently using CMC expression and splicing data. The number of intron retention (IR) targets for each modulator are plotted between the two datasets. The size of each circle represents the Fisher's exact test *p*-values that was calculated based on the number of targets found in the CMC models

CBR3-AS1 expression was low, *SRSF1* and *CHKB* PSI were positively correlated ($\rho = 0.237$), indicating that the *CHKB intron 1+2* was retained. In contrast, when *CBR3-AS1* expression was high, the correlation was negative ($\rho = -0.087$), indicating that the intron was spliced out (Figure 6b). A similar pattern was observed with the target *RFNG* (Figure 6c). Together, these results suggest that *CBR3-AS1* functions as an activator of SRSF1 splicing activity.

4 | DISCUSSION

Regulation of alternative splicing is a complex process that is particularly important in normal brain function. In this study, we developed a novel bioinformatics approach to identify causal modulators of splicing activity based on the variation of gene expression in large RNA sequencing datasets. We identified 13 modulators of the splicing factor SRSF1 that alter splicing of intron targets in the context of the aging human brain using sequencing data from ROSMAP and the CommonMind Consortium.

Our method was supported by the inclusion of matched individual DNA-level genotype data as part of a Mendelian randomized (MR) approach to identify gene associations (modulator expression) with phenotypes (RBP-mediated IR). Since DNA cannot be altered after the random assortment of genes at meiosis, it acts as a permanent surrogate to identify causality. In addition, using genetic variants to impute the gene expression of candidate modulators allowed us to isolate only the genetically determined component of gene expression while avoiding confounding,

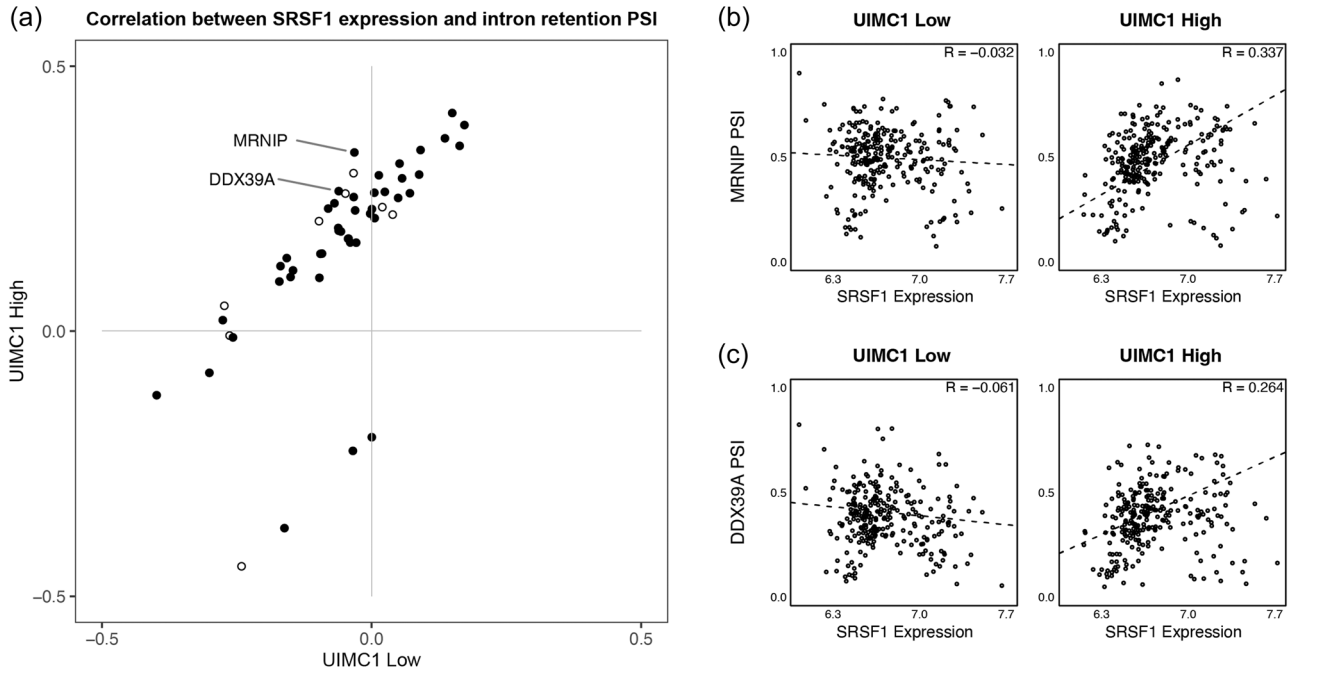

FIGURE 5 Ubiquitin interaction motif containing 1 (UIMC1) represses serine and arginine rich splicing factor 1 (SRSF1)-mediated intron splicing. (a) Correlation between *SRSF1* expression and intron retention PSI dependent on *UIMC1* expression levels. *UIMC1* gene expression in the Religious Order Study and Memory and Aging Project (ROSMAP) samples was imputed from genetic variants and divided into low (bottom tertile, *x*-axis) and high (top tertile, *y*-axis) groups. Each circle represents a splicing event. Spearman correlation coefficients between *SRSF1* expression and splicing PSI for 52 SRSF1-targeted splicing events are plotted. The 44 filled circles represent results corroborated using *UIMC1* expression from RNA-seq data. Spearman coefficient values in the upper left quadrant indicate that when UIMC1 activity is high, increasing *SRSF1* expression leads to more intron retention, which implies a reduction in SRSF1 splicing efficiency. (b,c) UIMC1 modulates the splicing of (b) *MRNIP (C5orf45) intron 4* (chr5:179264275-179267959, minus strand) and (c) *DDX39A intron 6* (chr19:14520553-14521146, minus strand). Each circle represents a sample. When *UIMC1* expression was low (left panels), percent spliced-in (PSI) levels were slightly anticorrelated to *SRSF1* expression. In contrast, when *UIMC1* expression was high (right panels), PSI levels showed positive correlation with *SRSF1* expression, indicating that UIMC1 appears to repress SRSF1 splicing activity. *SRSF1* expression is log2-normalized counts per million. The slope of the correlation coefficient is plotted (dashed line)

feedback, or environmental effects on transcript levels normally present in RNA-seq data. While the computed gene expression levels did not show strong correlation with the true expression values determined from the RNA-seq data, they were comparable to other predictive performances in brain tissue models derived from GTEx data (Huckins et al., 2019) or from combined transcriptomic and epigenetic reference data (Zhang et al., 2019), and models trained on Crohn's disease and type 1 diabetes (Fryett et al., 2018).

We applied multiple filtering steps to decrease the number of false-positive modulator interactions. Preliminary results prioritized possible modulator interactions with an unadjusted significance level. We further prioritized hits under the reasoning that a modulator should affect a significant number of splicing targets. We also maintained that modulators predicted by RNA-seq data should have the same effect on SRSF1 activity. These steps decreased the likelihood that our results were a product of random chance. Increasing the number of WGS samples could improve our model by providing a more accurate estimate of gene expression, but obtaining this data is costly and difficult when using tissue-specific human data such as brain. Integrating epigenetic data, such as chromatin binding sites, may also improve gene expression estimates.

The novelty and strength of an MR-based method lies in its ability to identify causal relationships and bypass confounders. However, a limitation of this approach is that using genetic variants to proxy gene expression may only explain a small proportion of the variance of the true expression as measured by RNA-seq. There can be several reasons for this discrepancy, but the most logical explanation is that gene expression is modulated extensively at the posttranscriptional level. Genetic variants comprise only the *cis*-acting elements contributing to gene expression, whereas several *trans*-acting factors at the transcriptional and posttranscriptional levels also contribute to gene expression and are not accounted for in our imputation model. To minimize this limitation, we also ran our model using RNA-seq-derived gene expression and required at least 60% of targets from the imputed expression to be replicated using the RNA-seq data.

In this study, we focused on the splicing factor SRSF1 due to prior evidence that its activity could be modulated. For example, SR proteins are phosphorylated by kinases such as SRPK1 or CLK1 that are themselves further regulated by molecular chaperones (Aubol et al., 2013; Zhong et al., 2009). We identified several protein modulators and one lncRNA modulator of SRSF1. Interestingly, none of these modulators

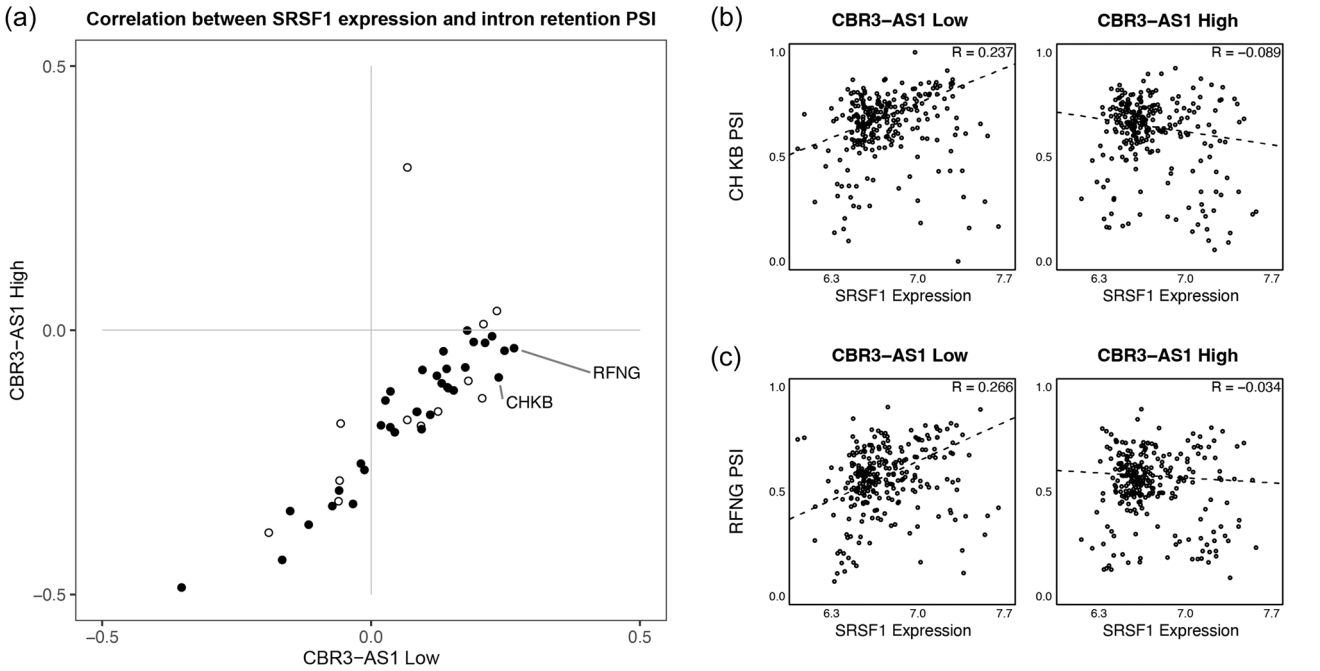

FIGURE 6 *CBR3-AS1* activates serine and arginine rich splicing factor 1 (SRSF1)-mediated intron splicing. (a) Correlation between *SRSF1* expression and intron retention percent spliced-in (PSI) is dependent on *CBR3-AS1* expression levels. *CBR3-AS1* expression in the Religious Order Study and Memory and Aging Project (ROSMAP) samples was imputed from genetic variants and divided into low (bottom tertile, *x*-axis) and high (top tertile, *y*-axis) groups. Each circle represents a splicing event. Spearman correlation coefficients between *SRSF1* expression and splicing PSI for 46 SRSF1-targeted splicing events are plotted. The 36 filled circles represent results corroborated using *CBR3-AS1* expression from RNA-seq data. The Spearman coefficient values in the bottom right quadrant indicate that when *CBR3-AS1* activity is low, there is a higher correlation between *SRSF1* expression and target intron retention; the corollary is that high *CBR3-AS1* expression is associated with higher SRSF1 splicing efficiency. (b,c) *CBR3-AS1* modulates the splicing of (b) *CHKB intron 1+2* (chr22:51020177–51021283, minus strand) and (c) *RFNG intron 4+5* (chr17:80007552–80008431, minus strand). Each circle represents a sample. When *CBR3-AS1* expression was low (left panels), PSI levels were positively correlated with *SRSF1* expression, indicating intron retention. When *CBR3-AS1* expression was high (right panels), PSI levels showed a weakly negative correlation with *SRSF1* expression, indicating that *CBR3-AS1* promotes SRSF1 splicing activity of the targeted intron. *SRSF1* expression is log2-normalized counts per million. The slope of the correlation coefficient is plotted (dashed line)

have been reported to interact with SRSF1 or with the splicing machinery. We did, however, verify 8 of 13 SRSF1 modulators in an independent data set of the same sample type. While it was surprising that SR kinases were not identified as modulators, it is possible that they may not be limiting or dosage-sensitive across the range of expression levels observed in the two datasets we used in this study. Additional experimental data are needed to predict the magnitude of effect each modulator has on alternative splicing outcomes. It would also be informative to further characterize the predicted modulators with CLIP-seq peaks when these data become available.

In this study, we used bulk RNA-seq data from the dorsolateral prefrontal cortex of the human brain. Because splicing regulation often differs between cell types, our findings might be improved if the input data were cell-type specific. A strength of our model is that it was designed as a transcriptome-wide search of candidate modulators, and candidate modulators were not limited to protein-coding genes. Noncoding RNAs have been increasingly implicated in aging and Alzheimer's disease (Abdelmohsen & Gorospe, 2015; Faghihi et al., 2008). We identified the lncRNA *CBR3-AS1* as a causal SRFS1 modulator, providing additional evidence of lncRNA involvement in posttranscriptional regulation.

We were particularly interested in IR due to its underappreciated significance and unexplored molecular consequences. While IR is often associated with nonsense-mediated decay and subsequent downregulation of gene expression, it also plays roles in the coordinated regulation of neuronal mRNA steady-state levels and targeted splicing responses upon neuronal activation (Mauger et al., 2016; Yap et al., 2012). IR-containing transcripts can be stored in the nucleus awaiting signals to be exported and translated. Some IR transcripts can even serve as sponges binding to other RNAs in the nucleus (Schmitz et al., 2017).

Since some IR targets are themselves splicing factor genes, it is tempting to speculate feedback loops in which changes in IR would lead to changes in other splicing factors. One SRSF1 target in our results is the retention of *SRSF7 intron 3*, which has been shown to act as an architectural RNA (arcRNA) that assembles nuclear bodies (Königs et al., 2020). These *SRSF7* bodies sequester intron-retained and fully spliced *SRSF7* isoforms, resulting in the reduction of functional SRSF7 protein in the nucleus. While it has been shown that *SRSF7* overexpression autoregulates this negative feedback mechanism, our results suggest that SRSF1 modulation may also participate in this process.

Finally, our method is generally applicable to other datasets, other RBPs, and other splicing types. Like our approach with ROSMAP, it can be used with other tissue-specific databases for modulators of hundreds of RBPs to facilitate genomic diagnostics with RNA-seq. Type 1 diabetes is an evolving field of interest where splicing alterations can unveil novel immunogenic epitopes (Wu et al., 2021). In addition, IR is a frequent event in many cancer types, and our recent studies have shown that IR-induced neoantigens may be a useful biomarker for predicting survival in multiple myeloma and pancreatic cancer patients (Dong et al., 2021, 2022). With its versatility, our approach can provide new biologically meaningful insights that would facilitate genomic investigations of diseases where RNA splicing plays an important role.

ACKNOWLEDGMENTS

We would like to acknowledge Dr. Dongbing Lai and Dr. Kwangsik Nho for their assistance in assembling input data. This study was supported by the Stark Neurosciences Research Institute, the Indiana Alzheimer Disease Center, Eli Lilly and Company, and by the Indiana Clinical and Translational Sciences Institute, funded in part by grant number UL1TR002529 from the National Institutes of Health, National Center for Advancing Translational Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The code generated for this study can be found in the GitHub repository https://github.com/stevenxchen/splicing-modulators.

ORCID

Steven X. Chen http://orcid.org/0000-0003-3463-5824
Ed Simpson http://orcid.org/0000-0001-9015-9864
Jill L. Reiter http://orcid.org/0000-0001-5460-2355
Yunlong Liu http://orcid.org/0000-0002-2699-626X

REFERENCES

Abdelmohsen, K., & Gorospe, M. (2015). Noncoding RNA control of cellular senescence. *WIREs RNA*, *6*(6), 615–629. https://doi.org/10.1002/wrna.1297

Aubol, B. E., Plocinik, R. M., Hagopian, J. C., Ma, C.-T., McGlone, M. L., Bandyopadhyay, R., Fu, X.-D., & Adams, J. A. (2013). Partitioning RS domain phosphorylation in an SR protein through the CLK and SRPK protein kinases. *Journal of Molecular Biology*, *425*(16), 2894–2909. https://doi.org/10.1016/j.jmb.2013.05.013

Aznarez, I., Nomakuchi, T. T., Tetenbaum-Novatt, J., Rahman, M. A., Fregoso, O., Rees, H., & Krainer, A. R. (2018). Mechanism of nonsense-mediated mRNA decay stimulation by splicing factor SRSF1. *Cell Reports*, *23*(7), 2186–2198. https://doi.org/10.1016/j.celrep.2018.04.039

Babur, Ö., Demir, E., Gönen, M., Sander, C., & Dogrusoz, U. (2010). Discovering modulators of gene expression. *Nucleic Acids Research*, *38*(17), 5648–5656. https://doi.org/10.1093/nar/gkq287

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., Kim, T., Misquitta-Ali, C. M., Wilson, M. D., Kim, P. M., Odom, D. T., Frey, B. J., & Blencowe, B. J. (2012). The evolutionary landscape of alternative splicing in vertebrate species. *Science (New York)*, *338*(6114), 1587–1593. https://doi.org/10.1126/science.1230612

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, *57*(1), 289–300.

Bennett, D. A., Buchman, A. S., Boyle, P. A., Barnes, L. L., Wilson, R. S., & Schneider, J. A. (2018). Religious orders study and rush memory and aging project. *Journal of Alzheimer's Disease*, *64*(Suppl 1), S161–S189. https://doi.org/10.3233/JAD-179939

Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annual Review of Biochemistry*, *72*(1), 291–336. https://doi.org/10.1146/annurev.biochem.72.121801.161720

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dong, C., Cesarano, A., Bombaci, G., Reiter, J. L., Yu, C. Y., Wang, Y., Jiang, Z., Zaid, M. A., Huang, K., Lu, X., Walker, B. A., Perna, F., & Liu, Y. (2021). Intron retention-induced neoantigen load correlates with unfavorable prognosis in multiple myeloma. *Oncogene*, *40*(42), 6130–6138. https://doi.org/10.1038/s41388-021-02005-y

Dong, C., Reiter, J. L., Dong, E., Wang, Y., Lee, K. P., Lu, X., & Liu, Y. (2022). Intron-retention neoantigen load predicts favorable prognosis in pancreatic cancer. *JCO Clinical Cancer Informatics*, *6*, e2100124. https://doi.org/10.1200/CCI.21.00124

Faghihi, M. A., Modarresi, F., Khalil, A. M., Wood, D. E., Sahagan, B. G., Morgan, T. E., Finch, C. E., Iii, G. S. L., Kenny, P. J., & Wahlestedt, C. (2008). Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β-secretase. *Nature Medicine*, *14*(7), 723–730. https://doi.org/10.1038/nm1784

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., & Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, *47*(D1), D766–D773. https://doi.org/10.1093/nar/gky955

Fryett, J. J., Inshaw, J., Morris, A. P., & Cordell, H. J. (2018). Comparison of methods for transcriptome imputation through application to two common complex diseases. *European Journal of Human Genetics*, *26*(11), 1658–1667. https://doi.org/10.1038/s41431-018-0176-5

Fu, X.-D., & Ares, M. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics*, *15*(10), 689–701. https://doi.org/10.1038/nrg3778

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., & Im, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, *47*(9), 1091–1098. https://doi.org/10.1038/ng.3367

Hoffman, G. E., Bendl, J., Voloudakis, G., Montgomery, K. S., Sloofman, L., Wang, Y.-C., Shah, H. R., Hauberg, M. E., Johnson, J. S., Girdhar, K., Song, L., Fullard, J. F., Kramer, R., Hahn, C.-G., Gur, R., Marenco, S., Lipska, B. K., Lewis, D. A., Haroutunian, V., & Roussos, P. (2019). CommonMind Consortium provides transcriptomic and epigenomic data for schizophrenia and bipolar disorder. *Scientific Data*, *6*(1), 180. https://doi.org/10.1038/s41597-019-0183-6

Howard, J. M., Lin, H., Wallace, A. J., Kim, G., Draper, J. M., Haeussler, M., Katzman, S., Toloue, M., Liu, Y., & Sanford, J. R. (2018). HNRNPA1 promotes recognition of splice site decoys by U2AF2 in vivo. *Genome Research*, *28*(5), 689–698. https://doi.org/10.1101/gr.229062.117

Huckins, L. M., Dobbyn, A., Ruderfer, D. M., Hoffman, G., Wang, W., Pardiñas, A. F., Rajagopal, V. M., Als, T. D., Nguyen, H., Girdhar, K., Boocock, J., Roussos, P., Fromer, M., Kramer, R., Domenici, E., Gamazon, E. R., Purcell, S., Demontis, D., Børglum, A. D., & Stahl, E. A. (2019). Gene expression imputation across multiple brain regions provides insights into schizophrenia risk. *Nature Genetics*, *51*(4), 659–674. https://doi.org/10.1038/s41588-019-0364-4

Jin, Y., Cui, Z., Li, X., Jin, X., & Peng, J. (2017). Upregulation of long non-coding RNA PlncRNA-1 promotes proliferation and induces epithelial-mesenchymal transition in prostate cancer. *Oncotarget*, *8*(16), 26090–26099. https://doi.org/10.18632/oncotarget.15318

Karni, R., de Stanchina, E., Lowe, S. W., Sinha, R., Mu, D., & Krainer, A. R. (2007). The gene encoding the splicing factor SF2/ASF is a proto-oncogene. *Nature Structural & Molecular Biology*, *14*(3), 185–193. https://doi.org/10.1038/nsmb1209

Königs, V., de Oliveira Freitas Machado, C., Arnold, B., Blümel, N., Solovyeva, A., Löbbert, S., Schafranek, M., Ruiz De Los Mozos, I., Wittig, I., McNicoll, F., Schulz, M. H., & Müller-McNicoll, M. (2020). SRSF7 maintains its homeostasis through the expression of Split-ORFs and nuclear body assembly. *Nature Structural & Molecular Biology*, *27*(3), 260–273. https://doi.org/10.1038/s41594-020-0385-9

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25. https://doi.org/10.1186/gb-2009-10-3-r25

Li, X., & Manley, J. L. (2005). Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability. *Cell*, *122*(3), 365–378. https://doi.org/10.1016/j.cell.2005.06.008

Liao, Y., Smyth, G. K., & Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Mauger, O., Lemoine, F., & Scheiffele, P. (2016). Targeted intron retention and excision for rapid gene regulation in response to neuronal activity. *Neuron*, *92*(6), 1266–1278. https://doi.org/10.1016/j.neuron.2016.11.032

Merkin, J., Russell, C., Chen, P., & Burge, C. B. (2012). Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*, *338*(6114), 1593–1599. https://doi.org/10.1126/science.1228186

Papasaikas, P., Tejedor, J. R., Vigevani, L., & Valcárcel, J. (2015). Functional splicing network reveals extensive regulatory potential of the core spliceosomal machinery. *Molecular Cell*, *57*(1), 7–22. https://doi.org/10.1016/j.molcel.2014.10.030

Raj, T., Li, Y. I., Wong, G., Humphrey, J., Wang, M., Ramdhani, S., Wang, Y.-C., Ng, B., Gupta, I., Haroutunian, V., Schadt, E. E., Young-Pearse, T., Mostafavi, S., Zhang, B., Sklar, P., Bennett, D. A., & De Jager, P. L. (2018). Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. *Nature Genetics*, *50*(11), 1584–1592. https://doi.org/10.1038/s41588-018-0238-1

Raponi, M., Smith, L. D., Silipo, M., Stuani, C., Buratti, E., & Baralle, D. (2014). BRCA1 exon 11 a model of long exon splicing regulation. *RNA Biology*, *11*(4), 351–359. https://doi.org/10.4161/rna.28458

Salmena, L., Poliseno, L., Tay, Y., Kats, L., & Pandolfi, P. P. (2011). A ceRNA hypothesis: The Rosetta Stone of a Hidden RNA language? *Cell*, *146*(3), 353–358. https://doi.org/10.1016/j.cell.2011.07.014

Sanford, J. R., Wang, X., Mort, M., VanDuyn, N., Cooper, D. N., Mooney, S. D., Edenberg, H. J., & Liu, Y. (2009). Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Research*, *19*(3), 381–394. https://doi.org/10.1101/gr.082503.108

Schmitz, U., Pinello, N., Jia, F., Alasmari, S., Ritchie, W., Keightley, M.-C., Shini, S., Lieschke, G. J., Wong, J. J.-L., & Rasko, J. E. J. (2017). Intron retention enhances gene regulatory complexity in vertebrates. *Genome Biology*, *18*(1), 216. https://doi.org/10.1186/s13059-017-1339-3

Shen, S., Park, J. W., Lu, Z., Lin, L., Henry, M. D., Wu, Y. N., Zhou, Q., & Xing, Y. (2014). rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences United States of America*, *111*(51), E5593–E5601. https://doi.org/10.1073/pnas.1419161111

Silipo, M., Gautrey, H., & Tyson-Capper, A. (2015). Deregulation of splicing factors and breast cancer development. *Journal of Molecular Cell Biology*, *7*(5), 388–401. https://doi.org/10.1093/jmcb/mjv027

Song, W., Mei, J.-Z., & Zhang, M. (2018). Long noncoding RNA PlncRNA-1 promotes colorectal cancer cell progression by regulating the PI3K/Akt signaling pathway. *Oncology Research*, *26*(2), 261–268. https://doi.org/10.3727/096504017X15031557924132

Tollervey, J. R., Wang, Z., Hortobágyi, T., Witten, J. T., Zarnack, K., Kayikci, M., Clark, T. A., Schweitzer, A. C., Rot, G., Curk, T., Zupan, B., Rogelj, B., Shaw, C. E., & Ule, J. (2011). Analysis of alternative splicing associated with aging and neurodegeneration in the human brain. *Genome Research*, *21*(10), 1572–1582. https://doi.org/10.1101/gr.122226.111

Ule, J., Jensen, K., Mele, A., & Darnell, R. B. (2005). CLIP: A method for identifying protein–RNA interaction sites in living cells. *Methods*, *37*(4), 376–386. https://doi.org/10.1016/j.ymeth.2005.07.018

Ullrich, S., & Guigó, R. (2020). Dynamic changes in intron retention are tightly associated with regulation of splicing factors and proliferative activity during B-cell development. *Nucleic Acids Research*, *48*(3), 1327–1340. https://doi.org/10.1093/nar/gkz1180

Wang, K., Saito, M., Bisikirska, B. C., Alvarez, M. J., Lim, W. K., Rajbhandari, P., Shen, Q., Nemenman, I., Basso, K., Margolin, A. A., Klein, U., Dalla-Favera, R., & Califano, A. (2009). Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. *Nature Biotechnology*, *27*(9), 829–837. https://doi.org/10.1038/nbt.1563

Wang, Y., Chen, S. X., Rao, X., & Liu, Y. (2020). Modulator-dependent RBPs changes alternative splicing outcomes in kidney cancer. *Frontiers in Genetics*, *11*, 265. https://doi.org/10.3389/fgene.2020.00265

Wang, Z., & Burge, C. B. (2008). Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA*, *14*(5), 802–813. https://doi.org/10.1261/rna.876308

Wu, W., Syed, F., Simpson, E., Lee, C.-C., Liu, J., Chang, G., Dong, C., Seitz, C., Eizirik, D. L., Mirmira, R. G., Liu, Y., & Evans-Molina, C. (2021). Impact of proinflammatory cytokines on alternative splicing patterns in human islets. *Diabetes*, *71*(1), 116–127. https://doi.org/10.2337/db20-0847

Xiao, X., Wang, Z., Jang, M., & Burge, C. B. (2007). Coevolutionary networks of splicing cis-regulatory elements. *Proceedings of the National Academy of Sciences United States of America*, *104*(47), 18583–18588. https://doi.org/10.1073/pnas.0707349104

Yap, K., Lim, Z. Q., Khandelia, P., Friedman, B., & Makeyev, E. V. (2012). Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention. *Genes & Development*, *26*(11), 1209–1223. https://doi.org/10.1101/gad.188037.112

Yee, B. A., Pratt, G. A., Graveley, B. R., Nostrand, E. L. V., & Yeo, G. W. (2019). RBP-Maps enables robust generation of splicing regulatory maps. *RNA*, *25*(2), 193–204. https://doi.org/10.1261/rna.069237.118

Yeo, G., Holste, D., Kreiman, G., & Burge, C. B. (2004). Variation in alternative splicing across human tissues. *Genome Biology*, *5*(10), R74. https://doi.org/10.1186/gb-2004-5-10-r74

Zhang, W., Voloudakis, G., Rajagopal, V. M., Readhead, B., Dudley, J. T., Schadt, E. E., Björkegren, J. L. M., Kim, Y., Fullard, J. F., Hoffman, G. E., & Roussos, P. (2019). Integrative transcriptome imputation reveals tissue-specific and shared biological mechanisms mediating susceptibility to complex traits. *Nature Communications*, *10*, 3834. https://doi.org/10.1038/s41467-019-11874-7

Zhong, X.-Y., Ding, J.-H., Adams, J. A., Ghosh, G., & Fu, X.-D. (2009). Regulation of SR protein phosphorylation and alternative splicing by modulating kinetic interactions of SRPK1 with molecular chaperones. *Genes & Development*, *23*(4), 482–495. https://doi.org/10.1101/gad.1752109

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Chen, S. X., Simpson, E., Reiter, J. L., & Liu, Y. (2022). Bioinformatics detection of modulators controlling splicing factor-dependent intron retention in the human brain. *Human Mutation*, 43, 1629–1641. https://doi.org/10.1002/humu.24379