Tissue-specific regulation of gene expression via unproductive splicing

Alexey Mironov[1], Maria Vlasenok[1], Sergei Margasyuk[1], Andrei A. Mironov[2], and Dmitri D. Pervouchine[1,*]

[1]Skolkovo Institute of Science and Technology, Center of Molecular and Cellular Biology, Bolshoy blv. 30, Moscow 121205, Russia

[2]Moscow State University, Faculty of Bioengineering and Bioinformatics, ul. Kolmogorova 1, Moscow 119991, Russia

[*]e-mail: d.pervouchine@skoltech.ru

July 4, 2022

Abstract

Eukaryotic gene expression is regulated post-transcriptionally by a universal mechanism called unproductive splicing, in which mRNA is triggered to degradation by the nonsense-mediated decay (NMD) pathway as a result of alternative splicing (AS). Only a few dozen unproductive splicing events (USEs) are documented, and many more remain to be identified. We performed an exhaustive search to catalog experimentally-validated and annotated USEs and monitored them in a panel of RNA-seq experiments from the Genotype-Tissue Expression Consortium by detecting concordant tissue-specific changes of gene expression levels and NMD isoform splicing rates. A juxtaposition of these data with transcriptomic responses to RNA-binding protein (RBP) expression perturbations from ENCODE Consortium and RBP footprinting data revealed a number of remarkable cases

of regulated USEs with strong tissue-specific signatures, including previously proposed models of brain-specific regulation of *PTBP2*, *DLG4*, and *GABBR1* mediated by *PTBP1*. Additionally, we identified a high-confidence set of 86 novel AS events generating NMD isoforms that are associated with tissue-specific downregulation of gene expression. We present a gallery of novel predicted USEs, including a *PTBP1*-controlled event in the *DCLK2* gene, which may be responsible for its inhibition in non-neural tissues. These results greatly expand current knowledge on tissue-specific unproductive splicing and its regulation.

Introduction

Nonsense mutations and frame-shifting splicing errors give rise to transcripts with premature termination codons (PTC), which encode truncated, deleterious proteins. In eukaryotes, such transcripts are selectively degraded by a translation-dependent mRNA surveillance pathway called the Nonsense-Mediated mRNA Decay (NMD) [1]. However, NMD serves not only as mRNA quality control system but also as a regulatory homeostatic mechanism to maintain the abundance of a broad class of physiological transcripts [2, 3, 4, 5, 6]. In this mechanism, referred to as regulated unproductive splicing and translation (RUST) [7, 8] or simply unproductive splicing [9, 10], the cell employs alternative splicing (AS) to produce PTC-containing transcript isoforms in order to post-transcriptionally downregulate the expression level of the gene [11, 9]. Unproductive splicing plays an essential role in normal and disease conditions including early embryonic development [4], granulocyte differentiation [12], stabilization, and repression of oncogenic expression [13, 14].

The majority of unproductive splicing events (USEs) are described in genes encoding RNA-binding proteins (RBPs) [15]. Many of them use unproductive splicing to autoregulate their expression levels in a negative feedback loop, in which the protein product binds its own pre-mRNA and causes alternative splicing to induce a PTC. This autoregulation takes place in almost all serine-arginine-rich (SR) proteins [16], many hnRNP proteins [17, 18, 19, 20], spliceosome components [21, 22], and even in ribosomal proteins [23, 24]. Autoregulatory unproductive splicing is found in almost all eukaryotes studied to date and exhibits a high degree

of evolutionary conservation [25, 26, 27].

Cross-regulatory unproductive splicing networks have a different hierarchical organization compared to transcriptional networks, with a few master regulators and many more regulatory connections among RBPs than between RBPs and other genes [8]. These connections have been characterized for many splicing factors, particularly for SR proteins, which coordinate their expression in a dense unproductive splicing network [16]. Cross-regulatory circuits among paralogs such as *PTBP1*/*PTBP2* [28, 17], *SRSF3*/*SRSF7* [29, 30], *RBM10*/*RBM5* [21], *RBFOX2*/*RBFOX3* [31], *hnRNPD*/ *hnRNPDL* [20], and *hnRNPL*/*hnRNPLL* [32] are particularly abundant, but unproductive splicing also extends beyond RBPs and shapes the transcriptional landscape in other gene classes [15]. Remarkably, the relationship between the expression of the NMD isoform and the mRNA or protein levels in many cases is complex due to indirect connectivities in the regulatory network [19, 33, 34, 35]. Tissue specificity of unproductive splicing has been studied only for a handful of cases [15], including the regulation of neural-specific expression of the postsynaptic proteins *DLG4* and *GABBR1* that is controlled by *PTBP1* and *PTBP2* [36, 37, 38], which are also discussed here.

In this work, we interrogated tissue-specific unproductive splicing using a large panel of human tissue transcriptomes from the Genotype-Tissue Expression (GTEx) project [39]. Towards this goal, we analyzed AS events that lead to the generation of the NMD isoform and identified cases with a strong negative association between gene expression level and NMD isoform splicing rate. We further characterized the tissue-specificity of this association and combined it with protein expression data assessed for non-GTEx human tissue samples from the ProteomicsDB database [40]. To identify RBPs that regulate unproductive splicing, we superimposed these results with RBP footprinting data from the POSTAR3 database [41] and publicly available experiments on transcriptomic responses to RBP expression perturbations, including those from the ENCODE Consortium [42]. Finally, we obtained additional support for tissue-specific regulation via unproductive splicing by detecting concordant changes between the expression of the predicted regulator and that of its target across GTEx tissues.

MATERIALS AND METHODS

Unproductive splicing events

The GRCh37 (hg19) assembly of the human genome was downloaded from the UCSC genome browser website [43] and used along with the comprehensive gene annotation (v35lift37) from GENCODE [44]. The coordinates of exons, introns, and stop codons were extracted from transcripts labeled as protein-coding and NMD to identify AS event types shown in Figure 1A. Among them, USEs were characterized by PTCs that were located within or downstream of the AS event, with the exception of poison exons and intron detention that can trigger NMD by inducing splice junctions in 3'-UTR (Figure S1). As a result, we obtained a stringent set of 5,309 USEs (Table S1) and 38,396 AS events that didn't generate an NMD isoform.

Quantification of AS

STAR v2.4.2a alignments [45] of short reads in 8551 samples from the Genotype-Tissue Expression (GTEx) Consortium v7 [46] (excluding samples from transformed cells and testis) were obtained from dbGaP website in BAM format under the accession number phs000424/GRU. Split reads supporting splice junctions as well as non-split reads supporting intron retention were extracted from BAM files using the IPSA package with the default settings [47].

In the GTEx samples, the splicing rate was estimated from split-read counts computed by the IPSA pipeline using the ψ (percent-spliced-in, PSI) metric, defined as the number of reads supporting the NMD isoform as a fraction of the combined number of reads supporting the NMD isoform and the protein-coding isoform (Figure S1). For uniformity, the ψ metric was defined with respect to the NMD isoform, i.e., $\psi = 1$ for a poison exon assumes that it is 100% included, but $\psi = 1$ for an essential exon assumes that it is 100% skipped. Only ψ values with the denominator of at least 15 were considered reliable. Further, we discarded AS events if the lower and the upper quartiles of the ψ distribution in GTEx samples coincided, or if ψ values were missing in more than 80% of samples of a particular tissue.

The results of RBP perturbation experiments followed by RNA-seq (Table S2), including knockdown (KD), knockout (KO), and overexpression (OE), and the results of NMD in-

activation experiments (Table S3) were downloaded from the ENCODE portal and the SRA archive in BAM and FASTQ formats [42]. Short reads in FASTQ files were mapped to the GRCh37 assembly of the human genome using STAR-2.7.7a aligner in two-pass mode with the following options: --outSJfilterOverhangMin 20 8 20 20 --alignSJDBoverhangMin 8 --outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax 0.1. In differential splicing analysis, BAM files were processed with rMATS v.4.1.1 [48] with the following parameters: --novelSS --mil 10 --mel 10000 --libType fr-unstranded --variable-read-length. Only JC output files were used, i.e. differential splicing analysis did not involve reads mapped within exons. In NMD inactivation experiments, we adjusted rMATS P-values for testing multiple USEs using the Benjamini-Hochberg method (FDR) and recognized an USE as targeted by NMD if $\Delta\psi$ was significantly greater than zero at FDR < 0.05.

Gene expression quantification and analysis

In the GTEx samples, the local gene expression level of an USE was defined as the denominator of its ψ ratio, which reflects the abundance of short reads in the vicinity of the AS event and allows estimation of gene expression changes that are independent of AS acting globally in the gene, however at the cost of higher stochasticity. In contrast, the global gene expression level was defined as the total number of reads mapped to a gene that harbors an USE (counts obtained from the GTEx portal). Both local and global gene expression values were normalized according to the DESeq2 methodology with a pseudocount of 8 [49]. Namely, each row of the expression matrix, whose columns correspond to samples and rows correspond to genes, was divided by the row median. The size factor sf_k of the sample k was estimated as the median of the k-th column. Each gene expression value was divided by sf_k and $\log_2$-transformed. The resulting expression values were centered by subtracting the median over all samples. The normalized local and global gene expression levels are denoted by e_l and e_g, respectively.

To evaluate differential gene expression in the RBP perturbation and NMD inactivation experiments, the read counts were extracted from the respective BAM files using RNASeQC utility [50] and processed with DESeq2 package [49] with apeglm shrinkage correction [51]. In each RBP perturbation experiment, we checked that RBP expression changed upon perturbation

in the intended direction, i.e. decreased upon KD or KO and increased upon OE; otherwise, the experiment was excluded from the analysis.

Unproductive splicing and gene expression

To detect the decrease of gene expression that was accompanied by the activation of the NMD isoform in GTEx, we compared gene expression levels in the 25% of samples with the highest ψ value (the upper quartile) vs. the 25% of samples with the lowest ψ value (the lower quartile) for each USE. Namely, we estimated the medians of e_l and e_g in these sample groups and compared them using the Mann-Whitney sum of ranks test with continuity correction. The statistical significance of the differences in e_l and e_g between the two groups (denoted as Δe_l and Δe_g) was interpreted by z-scores associated with the U-statistic of the test.

Next, we determined the features that distinguish USEs from protein-coding AS events. First, we selected the events, in which both e_l and e_g were positive (positive set) and those, in which both e_l and e_g were negative (negative set). In most cases, the negative set was substantially larger among USEs compared to protein-coding AS events (Figure S2A). Next, we compared the distributions of five parameters in the negative set for USEs and protein-coding AS events using the positive set as a control (Figure S2B-I). We found that the most remarkable feature that distinguished USEs and protein-coding AS events was the z-score of the global gene expression level. Additionally, we checked that gene expression levels don't show a significant dependence on ψ value for AS events in protein-coding genes (Figure S3).

Tissue specificity of USEs

Each USE was characterized by three parameters, ψ, e_l, and e_g. In each sample, we computed the deviations of these three metrics from their respective medians across all GTEx samples (pooled medians) and performed a sign test ($H_0 : p = 0.5$) in each tissue to statistically evaluate the number of positive and negative deviations. As a result, we obtained P-values adjusted for testing multiple tissues using the Benjamini-Hochberg false discovery rate (FDR) and the sign of the deviation of the tissue median from the pooled median for ψ, e_l, and e_g. After discarding the tissues, in which e_l and e_g had opposite signs, we selected tissues, in which the deviations

of ψ, e_l, and e_g were significant (FDR < 0.05) and categorized them into two groups by ψ and e_g having the same or the opposite signs. One-sided departures from the null hypothesis that deviations with the same and the opposite signs are equally likely were assessed using a binomial test, and the resulting P-values were adjusted for testing multiple USEs using the Benjamini-Hochberg FDR. USEs with FDR < 0.1 were considered tissue-specific.

Identification of regulators in tissue-specific USEs

The assessment of differential splicing by rMATS was carried out separately for each RBP perturbation experiment (Table S2). For uniformity, we used $\Delta\psi$ values reflecting the difference in the NMD isoform splicing rates between the high RBP condition and the low RBP condition (Control vs. KD, Control vs. KO, and OE vs. control). Only ψ estimates with the read support of at least 15 and $|\Delta\psi| \geq 0.1$ were considered. Consequently, each USE was characterized by $\Delta\psi$ values, one for each perturbation experiment, and their respective P-values corrected for multiple testing using the Benjamini-Hochberg FDR. Similarly, each USE was characterized by the global gene expression changes Δe_g, one for each perturbation experiment, and their associated P-values adjusted for multiple testing.

To account for the fact that some RBPs had only one perturbation experiment, while others had many, the values of $\Delta\psi$ and Δe_g were converted to scores ± 1.5, ± 1, and 0, where the sign corresponds to the sign of $\Delta\psi$ and Δe_g, and the absolute values 1.5, 1, and 0 correspond to significant differences (FDR < 0.05), insignificant differences (FDR≥ 0.05), and discarded values, respectively. These scores were averaged over all perturbation experiments for each RBP-USE pair. An RBP was recognized as a candidate regulator of an USE if the average $\Delta\psi$ and Δe_g scores had opposite signs. Additionally, we tested whether the tissue-specific expression profile of the predicted regulator was consistent with that of the USE using the same strategy as in the analysis of tissue specificity, but comparing ψ changes with the regulator expression changes instead of the changes in expression of the host gene.

RBP footprinting data

The complete POSTAR3 database in tabular format was kindly provided by the authors [41]. An RBP-USE pair was considered as having CLIP support in the gene if the RBP had at least one peak obtained by Piranha or eCLIP pipelines within the gene that harbors the USE, the boundaries of which were determined by the GENCODE annotation. An RBP-USE pair was further considered as having local CLIP support if a Piranha or eCLIP peak occurred within the intron containing the USE that was extended by 20 nts into the flanking exons.

Proteomic data

The data on protein expression in human tissues, which was measured by the precursor intensity in mass spectrometry (MS1-level), was downloaded from the Proteomics DB portal [52]. To test for negative associations between the NMD isoform splicing rate and the expression of the host gene product at the protein level, we matched the tissues from Proteomics DB to GTEx tissues (Table S4), computed the median ψ values for Proteomics DB tissues, and sorted the tissues in ascending order by the median ψ. The one-sided Jonckheere trend test was applied to check whether protein expression levels follow a descending trend. Rejection of the null hypothesis in such a test indicated a negative association between ψ and the protein expression level.

Statistical analysis

The data were analyzed using python version 3.8.2 and R statistics software version 3.6.3. Non-parametric tests were performed with the scipy.stats python package using normal approximation with continuity correction. The binomial (sign) test was performed by direct computation of the binomial tail probability. Jonckheere trend test was performed with the clinfun R package. In all figures, the significance levels 0.05, 0.01, and 0.001 are denoted by *, **, and ***, respectively.

RESULTS

Poison and essential AS events

In what follows, we consider four major types of AS events, namely cassette (skipped) exons, alternative 5'- and 3'-splice sites, and retained introns. They are classified according to the poison-essential dichotomy as in [53] (Figure 1A). An USE is referred to as poison (essential) if the NMD isoform represents an insertion (deletion) with respect to the protein-coding isoform. For instance, poison exons are normally skipped, but they trigger degradation by NMD when included in the mRNA. Numerous poison exons have been described in genes encoding SR proteins [9]. The reciprocal case is essential exons, which are normally included in the mRNA, but induce a PTC when skipped. Examples of essential exons are found in genes encoding hnRNP proteins, including *PTBP1* [17], *PTBP2* [18], and *FUS* [19]. Similarly, an alternative splice site or intron is referred to as poison (essential) if PTC occurs as a result of insertion (deletion) of a mRNA fragment. Examples of these USEs have also been described [54, 53, 30, 55].

The characterization of USEs and their regulation was based on the following strategy (Figure 1B). In cross-regulation, we expect a negative association between ψ (see Methods) and gene expression level. Therefore, an RBP that inhibits the NMD isoform must upregulate the expression level, and an increase in the expression level of such RBP must be accompanied by an increase in the expression level of the target gene and a decrease in ψ. Conversely, an RBP that activates the NMD isoform must downregulate the expression level, and an increase in the expression level of such RBP will lead to a decrease in the expression level of the target gene and an increase in ψ. In auto-regulation, we consider only negative feedback loops, in which an RBP activates the NMD isoform, increases ψ, and downregulates its own expression level.

Validated and annotated USEs

We performed an exhaustive literature search to catalog experimentally validated USEs (Table S5). In particular, we collected information on experimental outcomes and indicated whether the USE is reactive to NMD inactivation, whether RBP perturbations affect the NMD isoform

inclusion or gene expression on protein and mRNA levels, and whether there is evidence of binding of the targeted mRNA by RBPs. We obtained 237 RBP-USE pairs containing 48 RBPs and 57 USEs, including 203 cross-regulatory and 34 auto-regulatory cases for which the experimental validation was reported in human cell lines, mouse cell lines, or tissues. In what follows, these USEs will be referred to as validated. A snapshot of these findings is shown in Figure 2 and also in Figure S4, which illustrates a particularly dense subnetwork for SR proteins.

Among four considered AS types (Figure 1A), we identified 5,309 AS events that switch from protein-coding to NMD isoform (Table S1). These USEs are referred to as annotated. As in previous works [53], we used publicly available NMD inactivation experiments followed by RNA-seq to check whether the annotated USEs indeed produce NMD-sensitive transcripts. In accordance with the literature, 36 of 45 (80%) validated USEs with sufficient read support demonstrated a significant increase of ψ upon NMD inactivation in at least one experiment. However, only 1,435 of 3,196 (45%) annotated USEs did so, indicating that the evoked response to NMD inactivation may not always be a good predictor of unproductive splicing. Therefore, we treated the reactivity to NMD inactivation as an additional evidence and didn't use it as a constraint.

Association between unproductive splicing and gene expression

To study the association between unproductive splicing and gene expression at the mRNA level, we analyzed RNA-seq data from about 8.5 thousand GTEx samples, first disregarding their tissue attribution. Samples from testis were excluded since the action of the NMD system in this tissue is different from that in other tissues [56, 57, 58]. After the removal of the USEs with low variability and low read support, 2,831 out of 5,309 annotated cases and 48 out of 57 validated USEs remained (Table 1). To detect concordant changes between the NMD isoform splicing rate and the gene expression level, we characterized each USE by $\psi_H - \psi_L$, the difference of median ψ values between the upper and the lower quartiles of the ψ distribution, Δe_g and Δe_l, the difference of medians of the global and local expression levels, and their respective z-scores (see Methods). A comparison with AS events that don't generate NMD isoforms revealed a set of 579 USEs with z-score below -5, which we refer to as significant (Table S6). Among them,

there were 11 validated cases, most of which showed a remarkable magnitude of ψ changes (Figure 3A).

Table 1: The number of unproductive splicing events.

USE group	Validated	Annotated	Total
All	48	2,831	2,879
Significant	11	568	579
Tissue-specific	5	86	91
Regulated	3	47	50
CLIP in the gene	3	31	34
Local CLIP support	3	14	17

The strongest association was observed for the well-documented *PTBP1* targets such as *GABBR1* and *DLG4*, in which $\psi_H - \psi_L$ exceeded 50% and was accompanied by more than a twofold decrease of the expression level (Figure 3B). Other examples include USEs in *RBM10*, an important component of the spliceosome that is associated with genetic diseases and cancer [59, 60, 61], and *TRA2A*, a member of the SR protein family also known as an oncogene [62, 63, 64]. In *RBM10*, skipping of the essential exon 6 leads to a decrease in expression, which promotes proliferation in lung adenocarcinoma [65, 21]. In *TRA2A*, suppression of expression occurs through the stimulation of the poison exon 2 in its mRNA by the *TRA2A* gene product and its paralog *TRA2B* [16].

Tissue-specific regulation of USEs

To explore tissue-specificity of the association between splicing rate and gene expression for significant USEs, we estimated the deviations of the median ψ value and the median e_g in each tissue from the respective pooled medians and compared the number of tissues, in which these deviations had opposite signs, with the number of tissues, in which they had the same sign (see Methods). Under the null hypothesis that deviations with the same and the opposite signs are equally likely, this procedure yielded 91 USEs with significantly negative tissue-specific associations (Table S7). In what follows, we refer to these USEs as tissue-specific (Table 1).

On the other hand, a number of large-scale functional assays have assessed the response of the cellular transcriptome to perturbations of RBP expression levels [42]. We collected a panel of 419 such experiments followed by RNA-seq for 248 RBPs (Table S2). For each RBP, we assessed the concordance of changes in the NMD isoform splicing rate and changes in the target gene expression using multiple experiments and selected RBP-USE pairs, in which these changes had opposite signs (see Methods). In 124 out of 203 validated cross-regulatory cases, for which this assessment was possible, we identified 30 RBP-USE pairs with the expected regulatory outcome and only one pair, in which the direction of the regulation was opposite to that reported in the literature (Table S8). In application to tissue-specific USEs, it yielded at least one regulator in 74 cases, in 50 of which tissue-specific expression of the regulator was accompanied by tissue-specific ψ changes. In what follows, we refer to these USEs as regulated (Table 1).

To further confine the list of candidates, we required that the RBP bind the target pre-mRNA. To check this, we used high-throughput crosslinking and immunoprecipitation (CLIP) experiments provided in the POSTAR3 database [41] and found evidence of binding by at least one predicted regulator for 34 regulated USEs, including 17 cases with footprints in the immediate proximity of the USE (Table 1, Table S9).

We reconfirmed brain-specific USEs in the *PTBP2*, *DLG4*, and *GABBR1* genes, in which the NMD isoform is induced by *PTBP1*-mediated skipping of an essential exon (Figure 4). These USEs are known to upregulate gene expression by inhibiting the NMD isoform expression in the brain and activating it in the other tissues [18, 38, 36]. In accordance with this, we observed a statistically significant association with brain-specific expression of *PTBP1* in all three cases and a significant association with the expression of *SRRM4* for *PTBP2*. To further confirm the observed patterns, we used proteomic data to demonstrate that protein expression is increased in the brain, with a particularly strong trend in the case of *DLG4*.

We next considered 34 regulated USEs with CLIP support in the gene and clustered them according to their AS rates (Figure 5A). The clustering revealed three USE groups of approximately equal sizes, which were characterized by a decreased ψ in the brain (cluster 1), a decreased ψ in skeletal muscle and heart (cluster 2), and an increased ψ in the brain (cluster 3).

These USEs form a regulatory network, in which *PTBP1* controls the expression of 15 targets (Figure 5B). In cluster 1, *PTBP1* stimulates brain-specific gene expression by activating the NMD isoform in non-neural tissues, as it also does in the case of *PTBP2*, *GABBR1*, and *DLG4*.

A previously unknown example of USE regulation was found in the *DCLK2* gene, the product of which is necessary for the development of the hippocampus and the regulation of dendrite remodeling [66, 67]. We predict that *PTBP1* stimulates the inclusion of the poison exon 16, which leads to the suppression of *DCLK2* expression with a significant effect observed both at the mRNA and protein levels (Figure 6A). In this case, the presence of CLIP peaks of *PTBP1* both upstream and downstream of the poison exon indicates a non-canonical role of *PTBP1* as a splicing activator, consistently with the reports on the dual action of *PTBP1* that depends on its binding position relative to the exon [68].

PTBP1 may cooperate with other regulators such as *SRSF3*, *RBM15* and *AQR* to inhibit NMD isoforms in genes from cluster 3, e.g. *IQGAP1* (Figure 6B). The expression of *IQGAP1* is maintained at low levels in healthy brain tissues, but in brain tumors it is increased [69]. We confirm a brain-specific decrease of *IQGAP1* expression that is associated with increased poison exon 29 inclusion and reduced expression of *PTBP1* and *SRSF3*, most remarkably in the cerebellum. *PTBP1* may also regulate cluster 2 targets such as *ACSF3*, where it promotes the inclusion of an essential exon together with *DDX52* and *XRCC6* (Figure 6C). The product of *ACSF3* is necessary for mitochondrial metabolism, and hence it is upregulated in tissues where cells are rich in mitochondria, including liver, muscle, and heart [70, 71], where *PTBP1*, *DDX52*, and *XRCC6* are upregulated.

Tissues with frequent USE regulation

We next asked whether unproductive splicing is more active in some tissues than in the others. Concordant changes between ψ and gene expression level could arise by pure chance due to large fractions of upregulated NMD isoforms and downregulated genes. For instance, tibial nerve, thyroid, and prostate are characterized by simultaneous upregulation of NMD isoforms and downregulation of expression in many genes, while in skeletal muscle and heart, the pattern is the opposite (Figure S5).

To address this, we performed a χ^2-test for association between up- and downregulation of the NMD isoform and up- and downregulation of gene expression level in each tissue and compared the observed and the expected number of USEs (Figure S6). The observed frequencies of negative concordant changes between NMD isoform splicing rate and gene expression level significantly exceeded the expected frequencies in cerebellum, most brain subregions, and ovary, but not in thyroid, prostate, or tibial nerve. These findings indicate a particularly strong association between unproductive splicing and gene expression in the brain and support the importance of NMD in brain development and disease [72].

DISCUSSION

A coordinated interaction between AS and NMD that leads to the degradation of specific mRNAs, which constitutes unproductive splicing, is a widespread phenomenon that has been documented in almost all eukaryotes [73, 74, 11, 15]. Compared to other post-transcriptional regulatory mechanisms such as endogenous RNA interference [75, 76, 77, 78] and the control of mRNA stability by RNA modifications [79, 80, 81], unproductive splicing appears to act pervasively at the transcriptome level, as evidenced by the fact that nearly a third of human protein-coding genes have at least one annotated NMD transcript isoform [82].

The examples of *PTBP2*, *DLG4*, and *GABBR1* demonstrate as a proof of principle that tissue-specific unproductive splicing can be inferred from the transcriptomic data. Although we observed a significant association between NMD isoform splicing rate and gene expression level in many cases (Figure 3B), the majority of validated USEs exhibit no convincing evidence of such association. In part, this lack of response is explained by a bias in observing splicing changes that are masked by rapid degradation of unproductive isoforms by the NMD pathway and constant influx of pre-mRNAs due to ongoing transcription [27]. The unproductive splicing could also be inactive in fully differentiated tissues and operate only in specific conditions such as differentiation [83, 12], neurogenesis [38, 18], or hypoxia [84]. In fact, most splicing factors demonstrate only minor differences in expression between tissues [85] in comparison to KD or OE experiments, in which the validated USEs have been originally described. The regulatory

potential of unproductive splicing requires a relatively large dynamic range of AS changes, which may be restricted in the mature tissues. All these factors inherently limit the sensitivity of the proposed method.

A fundamental challenge in studying unproductive splicing is the existence of feedback loops and indirect connections in the network. For instance, in the case of autoregulation, the product of a gene that harbors a poison exon counteracts its own upregulation by promoting exon inclusion, thus leading to the downregulation of gene expression level and, consequently, to suppression of the NMD isoform that was initially upregulated. Splicing regulatory networks may contain circuits that serve as proxies and interfere with direct interactions, e.g., *PTBP1* upregulates *SRSF3*, which in turn upregulates *PTBP2*, but *PTBP1* itself directly downregulates *PTBP2* (Figure 2). As a result, the sign and the magnitude of the association between NMD isoform splicing rate and RBP expression may vary depending on the connectivity in the network potentially leading to both false positive and false negative predictions.

The methodology outlined here aims at the discovery of tissue-specific USEs. The core part of it is based on a statistically robust test that compares gene expression levels in the extremes of ψ distribution rather than on the analysis of correlations, thus estimating not only significant but also substantial and explicit differences. It doesn't depend on the tissue attribution in GTEx and can be applied to other panels of RNA-seq experiments. The other parts measuring the response of the transcriptome to the perturbations of RBP expression levels in combination with RBP footprinting provide the evidence of causality for the proposed regulatory circuits and can be used independently of the first part. Modular organization of this workflow allows discovery of regulated USEs based on their splicing and expression signatures in other transcriptomic datasets, although not all aspects of these signatures are currently well understood.

For instance, the response of the validated USEs to the inactivation of NMD system components and perturbations of RBP expression levels in cell lines doesn't always agree with the signatures that are observed in tissues. In some cases, the activation of unproductive splicing in tissues is concordant with the changes in mRNA expression levels but inconsistent with protein expression levels and vice versa. The analysis of RBP perturbation experiments consistently demonstrated nearly the same number of positive and negative associations between ψ

and e_g in the validated USEs implying the existence of positive feedback loops, the physiological relevance of which is debatable [30916337]. Further analysis of the molecular mechanisms underlying unproductive splicing and larger volumes of multi-omics data will be required to address these concerns in future studies.

CONCLUSION

This study presents an integrative bioinformatic workflow for the identification of tissue-specific unproductive splicing and its potential regulators. It confirms the regulatory mechanisms in the validated cases and predicts novel candidates including unproductive splicing evens in the *DCLK2*, *IQGAP1*, and *ACSF3* genes that are controlled by *PTBP1*. These results greatly expand current knowledge on tissue-specific regulation of unproductive splicing and open new avenues for future research.

AVAILABILITY OF DATA AND MATERIALS

The datasets generated during the current study are available online at `https://doi.org/10.5281/zenodo.6786643`. The source code used for the analysis is available at `https://github.com/magmir71/unproductive_splicing`.

ACKNOWLEDGEMENTS

The authors thank Drs. Roderic Guigó, Silvia Perez Lluch, Dmitry Skvortsov, and Olga Donstova for insightful discussions about molecular mechanisms of unproductive splicing. A.M., M.V., S.M., A.A.M., and D.D.P. acknowledge Russian Science Foundation grant 22-14-00330.

Conflict of interest statement.

The authors declare no competing interests.

References

[1] Lykke-Andersen, S. and Jensen, T. H. (Nov, 2015) Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat Rev Mol Cell Biol,* **16**(11), 665–677.

[2] Medghalchi, S. M., Frischmeyer, P. A., Mendell, J. T., Kelly, A. G., Lawler, A. M., and Dietz, H. C. (Jan, 2001) Rent1, a trans-effector of nonsense-mediated mRNA decay, is essential for mammalian embryonic viability. *Hum Mol Genet,* **10**(2), 99–105.

[3] Weischenfeldt, J., Damgaard, I., Bryder, D., Theilgaard-Mönch, K., Thoren, L. A., Nielsen, F. C., Jacobsen, S. E., Nerlov, C., and Porse, B. T. (May, 2008) NMD is essential for hematopoietic stem and progenitor cells and for eliminating by-products of programmed DNA rearrangements. *Genes Dev,* **22**(10), 1381–1396.

[4] McIlwain, D. R., Pan, Q., Reilly, P. T., Elia, A. J., McCracken, S., Wakeham, A. C., Itie-Youten, A., Blencowe, B. J., and Mak, T. W. (Jul, 2010) Smg1 is required for embryogenesis and regulates diverse genes via alternative splicing coupled to nonsense-mediated mRNA decay. *Proc Natl Acad Sci U S A,* **107**(27), 12186–12191.

[5] Li, T., Shi, Y., Wang, P., Guachalla, L. M., Sun, B., Joerss, T., Chen, Y. S., Groth, M., Krueger, A., Platzer, M., Yang, Y. G., Rudolph, K. L., and Wang, Z. Q. (Jun, 2015) Smg6/Est1 licenses embryonic stem cell differentiation via nonsense-mediated mRNA decay. *EMBO J,* **34**(12), 1630–1647.

[6] Lou, C. H., Shao, A., Shum, E. Y., Espinoza, J. L., Huang, L., Karam, R., and Wilkinson, M. F. (Feb, 2014) Posttranscriptional control of the stem cell and neurogenic programs by the nonsense-mediated RNA decay pathway. *Cell Rep,* **6**(4), 748–764.

[7] Lewis, B. P., Green, R. E., and Brenner, S. E. (Jan, 2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci U S A,* **100**(1), 189–192.

[8] Desai, A., Hu, Z., French, C. E., Lloyd, J. P. B., and Brenner, S. E. (2020) Networks of Splice Factor Regulation by Unproductive Splicing Coupled With Nonsense Mediated mRNA Decay. *bioRxiv,*.

[9] Lareau, L. F., Inada, M., Green, R. E., Wengrod, J. C., and Brenner, S. E. (Apr, 2007) Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature,* **446**(7138), 926–929.

[10] Filichkin, S. A. and Mockler, T. C. (Jul, 2012) Unproductive alternative splicing and nonsense mRNAs: a widespread phenomenon among plant circadian clock genes. *Biol Direct,* **7**, 20.

[11] Ni, J. Z., Grate, L., Donohue, J. P., Preston, C., Nobida, N., O'Brien, G., Shiue, L., Clark, T. A., Blume, J. E., and Ares, M. (Mar, 2007) Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev,* **21**(6), 708–718.

[12] Wong, J. J., Ritchie, W., Ebner, O. A., Selbach, M., Wong, J. W., Huang, Y., Gao, D., Pinello, N., Gonzalez, M., Baidya, K., Thoeng, A., Khoo, T. L., Bailey, C. G., Holst, J., and Rasko, J. E. (Aug, 2013) Orchestrated intron retention regulates normal granulocyte differentiation. *Cell,* **154**(3), 583–595.

[13] Dewaele, M., Tabaglio, T., Willekens, K., Bezzi, M., Teo, S. X., Low, D. H., Koh, C. M., Rambow, F., Fiers, M., Rogiers, A., Radaelli, E., Al-Haddawi, M., Tan, S. Y., Hermans, E., Amant, F., Yan, H., Lakshmanan, M., Koumar, R. C., Lim, S. T., Derheimer, F. A., Campbell, R. M., Bonday, Z., Tergaonkar, V., Shackleton, M., Blattner, C., Marine, J. C., and Guccione, E. (Jan, 2016) Antisense oligonucleotide-mediated MDM4 exon 6 skipping impairs tumor growth. *J Clin Invest,* **126**(1), 68–84.

[14] Barbier, J., Dutertre, M., Bittencourt, D., Sanchez, G., Gratadou, L., de la Grange, P., and Auboeuf, D. (Oct, 2007) Regulation of H-ras splice variant expression by cross talk between the p53 and nonsense-mediated mRNA decay pathways. *Mol Cell Biol,* **27**(20), 7315–7333.

[15] García-Moreno, J. F. and Romão, L. (Dec, 2020) Perspective in Alternative Splicing Coupled to Nonsense-Mediated mRNA Decay. *Int J Mol Sci,* **21**(24).

[16] Leclair, N. K., Brugiolo, M., Urbanski, L., Lawson, S. C., Thakar, K., Yurieva, M., George, J., Hinson, J. T., Cheng, A., Graveley, B. R., and Anczuków, O. (11, 2020) Poison Exon Splicing Regulates a Coordinated Network of SR Protein Expression during Differentiation and Tumorigenesis. *Mol Cell,* **80**(4), 648–665.

[17] Spellman, R., Llorian, M., and Smith, C. W. (Aug, 2007) Crossregulation and functional redundancy between the splicing regulator PTB and its paralogs nPTB and ROD1. *Mol Cell,* **27**(3), 420–434.

[18] Boutz, P. L., Stoilov, P., Li, Q., Lin, C. H., Chawla, G., Ostrow, K., Shiue, L., Ares, M., and Black, D. L. (Jul, 2007) A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes Dev,* **21**(13), 1636–1652.

[19] Zhou, Y., Liu, S., Liu, G., Oztürk, A., and Hicks, G. G. (Oct, 2013) ALS-associated FUS mutations result in compromised FUS alternative splicing and autoregulation. *PLoS Genet,* **9**(10), e1003895.

[20] Kemmerer, K., Fischer, S., and Weigand, J. E. (03, 2018) Auto- and cross-regulation of the hnRNPs D and DL. *RNA,* **24**(3), 324–331.

[21] Sun, Y., Bao, Y., Han, W., Song, F., Shen, X., Zhao, J., Zuo, J., Saffen, D., Chen, W., Wang, Z., You, X., and Wang, Y. (Aug, 2017) Autoregulation of RBM10 and cross-regulation of RBM10/RBM5 via alternative splicing-coupled nonsense-mediated decay. *Nucleic Acids Res,* **45**(14), 8524–8540.

[22] Saltzman, A. L., Pan, Q., and Blencowe, B. J. (Feb, 2011) Regulation of alternative splicing by the core spliceosomal machinery. *Genes Dev,* **25**(4), 373–384.

[23] Cuccurese, M., Russo, G., Russo, A., and Pietropaolo, C. (2005) Alternative splicing and nonsense-mediated mRNA decay regulate mammalian ribosomal gene expression. *Nucleic Acids Res,* **33**(18), 5965–5977.

[24] Takei, S., Togo-Ohno, M., Suzuki, Y., and Kuroyanagi, H. (07, 2016) Evolutionarily conserved autoregulation of alternative pre-mRNA splicing by ribosomal protein L10a. *Nucleic Acids Res,* **44**(12), 5585–5596.

[25] Lareau, L. F. and Brenner, S. E. (Apr, 2015) Regulation of splicing factors by alternative splicing and NMD is conserved between kingdoms yet evolutionarily flexible. *Mol Biol Evol,* **32**(4), 1072–1079.

[26] Kalyna, M., Simpson, C. G., Syed, N. H., Lewandowska, D., Marquez, Y., Kusenda, B., Marshall, J., Fuller, J., Cardle, L., McNicol, J., Dinh, H. Q., Barta, A., and Brown, J. W. (Mar, 2012) Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. *Nucleic Acids Res,* **40**(6), 2454–2469.

[27] Kovalak, C., Donovan, S., Bicknell, A. A., Metkar, M., and Moore, M. J. (05, 2021) Deep sequencing of pre-translational mRNPs reveals hidden flux through evolutionarily conserved alternative splicing nonsense-mediated decay pathways. *Genome Biol,* **22**(1), 132.

[28] Wollerton, M. C., Gooding, C., Wagner, E. J., Garcia-Blanco, M. A., and Smith, C. W. (Jan, 2004) Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. *Mol Cell,* **13**(1), 91–100.

[29] Jumaa, H. and Nielsen, P. J. (Aug, 1997) The splicing factor SRp20 modifies splicing of its own mRNA and ASF/SF2 antagonizes this regulation. *EMBO J,* **16**(16), 5077–5085.

[30] Änkö, M. L., Müller-McNicoll, M., Brandl, H., Curk, T., Gorup, C., Henry, I., Ule, J., and Neugebauer, K. M. (2012) The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. *Genome Biol,* **13**(3), R17.

[31] Jangi, M., Boutz, P. L., Paul, P., and Sharp, P. A. (Mar, 2014) Rbfox2 controls autoregulation in RNA-binding protein networks. *Genes Dev,* **28**(6), 637–651.

[32] Rossbach, O., Hung, L. H., Schreiner, S., Grishina, I., Heiner, M., Hui, J., and Bindereif, A. (Mar, 2009) Auto- and cross-regulation of the hnRNP L proteins by alternative splicing. *Mol Cell Biol,* **29**(6), 1442–1451.

[33] Sun, S., Zhang, Z., Sinha, R., Karni, R., and Krainer, A. R. (Mar, 2010) SF2/ASF autoregulation involves multiple layers of post-transcriptional and translational control. *Nat Struct Mol Biol,* **17**(3), 306–312.

[34] Jiménez, M., Urtasun, R., Elizalde, M., Azkona, M., Latasa, M. U., Uriarte, I., Arechederra, M., Alignani, D., Bárcena-Varela, M., Álvarez Sola, G., Colyn, L., Santamaría, E., Sangro, B., Rodriguez-Ortigosa, C., Fernández-Barrena, M. G., Ávila, M. A., and Berasain, C. (04, 2019) Splicing events in the control of genome integrity: role of SLU7 and truncated SRSF3 proteins. *Nucleic Acids Res,* **47**(7), 3450–3466.

[35] Nakano, Y., Kelly, M. C., Rehman, A. U., Boger, E. T., Morell, R. J., Kelley, M. W., Friedman, T. B., and Bánfi, B. (07, 2018) Defects in the Alternative Splicing-Dependent Regulation of REST Cause Deafness. *Cell,* **174**(3), 536–548.

[36] Zheng, S., Gray, E. E., Chawla, G., Porse, B. T., O'Dell, T. J., and Black, D. L. (Jan, 2012) PSD-95 is post-transcriptionally repressed during early neural development by PTBP1 and PTBP2. *Nat Neurosci,* **15**(3), 381–388.

[37] Zheng, S. (Dec, 2016) Alternative splicing and nonsense-mediated mRNA decay enforce neural specific gene expression. *Int J Dev Neurosci,* **55**, 102–108.

[38] Makeyev, E. V., Zhang, J., Carrasco, M. A., and Maniatis, T. (Aug, 2007) The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. *Mol Cell,* **27**(3), 435–448.

[39] Aguet, F., Anand, S., Ardlie, K. G., Gabriel, S., Getz, G. A., Graubert, A., Hadley, K., Handsaker, R. E., Huang, K. H., Kashin, S., Li, X., MacArthur, D. G., Meier, S. R.,

Nedzel, J. L., Nguyen, D. T., Segrè, A. V., Todres, E., Balliu, B., Barbeira, A. N., Battle, A., Bonazzola, R., Brown, A., Brown, C. D., Castel, S. E., Conrad, D. F., Cotter, D. J., Cox, N., Das, S., de Goede, O. M., Dermitzakis, E. T., Einson, J., Engelhardt, B. E., Eskin, E., Eulalio, T. Y., Ferraro, N. M., Flynn, E. D., Fresard, L., Gamazon, E. R., Garrido-Martín, D., Gay, N. R., Gloudemans, M. J., Guigó, R., Hame, A. R., He, Y., Hoffman, P. J., Hormozdiari, F., Hou, L., Im, H. K., Jo, B., Kasela, S., Kellis, M., Kim-Hellmuth, S., Kwong, A., Lappalainen, T., Li, X., Liang, Y., Mangul, S., Mohammadi, P., Montgomery, S. B., Muñoz-Aguirre, M., Nachun, D. C., Nobel, A. B., Oliva, M., Park, Y., Park, Y., Parsana, P., Rao, A. S., Reverter, F., Rouhana, J. M., Sabatti, C., Saha, A., Stephens, M., Stranger, B. E., Strober, B. J., Teran, N. A., Viñuela, A., Wang, G., Wen, X., Wright, F., Wucher, V., Zou, Y., Ferreira, P. G., Li, G., Melé, M., Yeger-Lotem, E., Barcus, M. E., Bradbury, D., Krubit, T., McLean, J. A., Qi, L., Robinson, K., Roche, N. V., Smith, A. M., Sobin, L., Tabor, D. E., Undale, A., Bridge, J., Brigham, L. E., Foster, B. A., Gillard, B. M., Hasz, R., Hunter, M., Johns, C., Johnson, M., Karasik, E., Kopen, G., Leinweber, W. F., McDonald, A., Moser, M. T., Myer, K., Ramsey, K. D., Roe, B., Shad, S., Thomas, J. A., Walters, G., Washington, M., Wheeler, J., Jewell, S. D., Rohrer, D. C., Valley, D. R., Davis, D. A., Mash, D. C., Branton, P. A., Barker, L. K., Gardiner, H. M., Mosavel, M., Siminoff, L. A., Flicek, P., Haeussler, M., Juettemann, T., Kent, W. J., Lee, C. M., Powell, C. C., Rosenbloom, K. R., Ruffier, M., Sheppard, D., Taylor, K., Trevanion, S. J., Zerbino, D. R., Abell, N. S., Akey, J., Chen, L., Demanelis, K., Doherty, J. A., Feinberg, A. P., Hansen, K. D., Hickey, P. F., Jasmine, F., Jiang, L., Kaul, R., Kibriya, M. G., Li, J. B., Li, Q., Lin, S., Linder, S. E., Pierce, B. L., Rizzardi, L. F., Skol, A. D., Smith, K. S., Snyder, M., Stamatoyannopoulos, J., Tang, H., Wang, M., Carithers, L. J., Guan, P., Koester, S. E., Little, A. R., Moore, H. M., Nierras, C. R., Rao, A. K., Vaught, J. B., and Volpi, S. (09, 2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science,* **369**(6509), 1318–1330.

[40] Samaras, P., Schmidt, T., Frejno, M., Gessulat, S., Reinecke, M., Jarzab, A., Zecha, J., Mergner, J., Giansanti, P., Ehrlich, H. C., Aiche, S., Rank, J., Kienegger, H., Krcmar, H., Kuster, B., and Wilhelm, M. (01, 2020) ProteomicsDB: a multi-omics and multi-organism

resource for life science research. *Nucleic Acids Res,* **48**(D1), D1153–D1163.

[41] Zhao, W., Zhang, S., Zhu, Y., Xi, X., Bao, P., Ma, Z., Kapral, T. H., Chen, S., Zagrovic, B., Yang, Y. T., and Lu, Z. J. (01, 2022) POSTAR3: an updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Res,* **50**(D1), D287–D294.

[42] Van Nostrand, E. L., Freese, P., Pratt, G. A., Wang, X., Wei, X., Xiao, R., Blue, S. M., Chen, J. Y., Cody, N. A. L., Dominguez, D., Olson, S., Sundararaman, B., Zhan, L., Bazile, C., Bouvrette, L. P. B., Bergalet, J., Duff, M. O., Garcia, K. E., Gelboin-Burkhart, C., Hochman, M., Lambert, N. J., Li, H., McGurk, M. P., Nguyen, T. B., Palden, T., Rabano, I., Sathe, S., Stanton, R., Su, A., Wang, R., Yee, B. A., Zhou, B., Louie, A. L., Aigner, S., Fu, X. D., Lécuyer, E., Burge, C. B., Graveley, B. R., and Yeo, G. W. (07, 2020) A large-scale binding and functional map of human RNA-binding proteins. *Nature,* **583**(7818), 711–719.

[43] Haeussler, M., Zweig, A. S., Tyner, C., Speir, M. L., Rosenbloom, K. R., Raney, B. J., Lee, C. M., Lee, B. T., Hinrichs, A. S., Gonzalez, J. N., Gibson, D., Diekhans, M., Clawson, H., Casper, J., Barber, G. P., Haussler, D., Kuhn, R. M., and Kent, W. J. (01, 2019) The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res,* **47**(D1), D853–D858.

[44] Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J. E., Mudge, J. M., Sisu, C., Wright, J. C., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Boix, C., Carbonell Sala, S., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., García Girón, C., Gonzalez, J. M., Grego, T., Hardy, M., Hourlier, T., Howe, K. L., Hunt, T., Izuogu, O. G., Johnson, R., Martin, F. J., Martínez, L., Mohanan, S., Muir, P., Navarro, F. C. P., Parker, A., Pei, B., Pozo, F., Riera, F. C., Ruffier, M., Schmitt, B. M., Stapleton, E., Suner, M. M., Sycheva, I., Uszczynska-Ratajczak, B., Wolf, M. Y., Xu, J., Yang, Y. T., Yates, A., Zerbino, D., Zhang, Y., Choudhary, J. S., Gerstein, M., Guigó, R., Hubbard, T. J. P., Kellis, M., Paten, B., Tress, M. L., and Flicek, P. (01, 2021) GENCODE 2021. *Nucleic Acids Res,* **49**(D1), D916–D923.

[45] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (Jan, 2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics,* **29**(1), 15–21.

[46] Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., Johnson, R., Segrè, A. V., Djebali, S., Niarchou, A., Wright, F. A., Lappalainen, T., Calvo, M., Getz, G., Dermitzakis, E. T., Ardlie, K. G., and Guigó, R. (May, 2015) Human genomics. The human transcriptome across tissues and individuals. *Science,* **348**(6235), 660–665.

[47] Pervouchine, D. D., Knowles, D. G., and Guigó, R. (Jan, 2013) Intron-centric estimation of alternative splicing from RNA-seq data. *Bioinformatics,* **29**(2), 273–274.

[48] Shen, S., Park, J. W., Lu, Z. X., Lin, L., Henry, M. D., Wu, Y. N., Zhou, Q., and Xing, Y. (Dec, 2014) rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A,* **111**(51), E5593–5601.

[49] Love, M. I., Huber, W., and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol,* **15**(12), 550.

[50] Graubert, A., Aguet, F., Ravi, A., Ardlie, K. G., and Getz, G. (Mar, 2021) RNA-SeQC 2: Efficient RNA-seq quality control and quantification for large cohorts. *Bioinformatics,*.

[51] Zhu, A., Ibrahim, J. G., and Love, M. I. (06, 2019) Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics,* **35**(12), 2084–2092.

[52] Lautenbacher, L., Samaras, P., Muller, J., Grafberger, A., Shraideh, M., Rank, J., Fuchs, S. T., Schmidt, T. K., The, M., Dallago, C., Wittges, H., Rost, B., Krcmar, H., Kuster, B., and Wilhelm, M. (01, 2022) ProteomicsDB: toward a FAIR open-source resource for life-science research. *Nucleic Acids Res,* **50**(D1), D1541–D1552.

[53] Pervouchine, D., Popov, Y., Berry, A., Borsari, B., Frankish, A., and Guigó, R. (06, 2019) Integrative transcriptomic analysis suggests new autoregulatory splicing events coupled with nonsense-mediated mRNA decay. *Nucleic Acids Res,* **47**(10), 5293–5306.

[54] Hamid, F. M. and Makeyev, E. V. (Nov, 2014) Regulation of mRNA abundance by polypyrimidine tract-binding protein-controlled alternate 5' splice site choice. *PLoS Genet,* **10**(11), e1004771.

[55] Boutz, P. L., Bhutkar, A., and Sharp, P. A. (Jan, 2015) Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes Dev,* **29**(1), 63–80.

[56] Shum, E. Y., Jones, S. H., Shao, A., Dumdie, J., Krause, M. D., Chan, W. K., Lou, C. H., Espinoza, J. L., Song, H. W., Phan, M. H., Ramaiah, M., Huang, L., McCarrey, J. R., Peterson, K. J., De Rooij, D. G., Cook-Andersen, H., and Wilkinson, M. F. (Apr, 2016) The Antagonistic Gene Paralogs Upf3a and Upf3b Govern Nonsense-Mediated RNA Decay. *Cell,* **165**(2), 382–395.

[57] MacDonald, C. C. and Grozdanov, P. N. (May, 2017) Nonsense in the testis: multiple roles for nonsense-mediated decay revealed in male reproduction. *Biol Reprod,* **96**(5), 939–947.

[58] Zetoune, A. B., Fontanière, S., Magnin, D., Anczuków, O., Buisson, M., Zhang, C. X., and Mazoyer, S. (Dec, 2008) Comparison of nonsense-mediated mRNA decay efficiency in various murine tissues. *BMC Genet,* **9**, 83.

[59] Hegele, A., Kamburov, A., Grossmann, A., Sourlis, C., Wowro, S., Weimann, M., Will, C. L., Pena, V., Lührmann, R., and Stelzl, U. (Feb, 2012) Dynamic protein-protein interaction wiring of the human spliceosome. *Mol Cell,* **45**(4), 567–580.

[60] Johnston, J. J., Teer, J. K., Cherukuri, P. F., Hansen, N. F., Loftus, S. K., Chong, K., Mullikin, J. C., and Biesecker, L. G. (May, 2010) Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate. *Am J Hum Genet,* **86**(5), 743–748.

[61] Jung, J. H., Lee, H., Cao, B., Liao, P., Zeng, S. X., and Lu, H. (01, 2020) RNA-binding motif protein 10 induces apoptosis and suppresses proliferation by activating p53. *Oncogene,* **39**(5), 1031–1040.

[62] Zhao, X., Chen, Q., Cai, Y., Chen, D., Bei, M., Dong, H., and Xu, J. (2021) TRA2A Binds With LncRNA MALAT1 To Promote Esophageal Cancer Progression By Regulating EZH2/β-catenin Pathway. *J Cancer,* **12**(16), 4883–4890.

[63] Tan, Y., Hu, X., Deng, Y., Yuan, P., Xie, Y., and Wang, J. (10, 2018) TRA2A promotes proliferation, migration, invasion and epithelial mesenchymal transition of glioma cells. *Brain Res Bull,* **143**, 138–144.

[64] Xu, W., Huang, H., Yu, L., and Cao, L. (Apr, 2015) Meta-analysis of gene expression profiles indicates genes in spliceosome pathway are up-regulated in hepatocellular carcinoma (HCC). *Med Oncol,* **32**(4), 96.

[65] Zhao, J., Sun, Y., Huang, Y., Song, F., Huang, Z., Bao, Y., Zuo, J., Saffen, D., Shao, Z., Liu, W., and Wang, Y. (01, 2017) Functional analysis reveals that RBM10 mutations contribute to lung adenocarcinoma pathogenesis by deregulating splicing. *Sci Rep,* **7**, 40488.

[66] Kerjan, G., Koizumi, H., Han, E. B., Dubé, C. M., Djakovic, S. N., Patrick, G. N., Baram, T. Z., Heinemann, S. F., and Gleeson, J. G. (Apr, 2009) Mice lacking doublecortin and doublecortin-like kinase 2 display altered hippocampal neuronal maturation and spontaneous seizures. *Proc Natl Acad Sci U S A,* **106**(16), 6766–6771.

[67] Shin, E., Kashiwagi, Y., Kuriu, T., Iwasaki, H., Tanaka, T., Koizumi, H., Gleeson, J. G., and Okabe, S. (2013) Doublecortin-like kinase enhances dendritic remodelling and negatively regulates synapse maturation. *Nat Commun,* **4**, 1440.

[68] Hamid, F. M. and Makeyev, E. V. (12, 2017) A mechanism underlying position-specific regulation of alternative splicing. *Nucleic Acids Res,* **45**(21), 12455–12468.

[69] McDonald, K. L., O'Sullivan, M. G., Parkinson, J. F., Shaw, J. M., Payne, C. A., Brewer, J. M., Young, L., Reader, D. J., Wheeler, H. T., Cook, R. J., Biggs, M. T., Little, N. S., Teo,

C., Stone, G., and Robinson, B. G. (May, 2007) IQGAP1 and IGFBP2: valuable biomarkers for determining prognosis in glioma patients. *J Neuropathol Exp Neurol,* **66**(5), 405–417.

[70] Bowman, C. E., Rodriguez, S., Selen Alpergin, E. S., Acoba, M. G., Zhao, L., Hartung, T., Claypool, S. M., Watkins, P. A., and Wolfgang, M. J. (Jun, 2017) The Mammalian Malonyl-CoA Synthetase ACSF3 Is Required for Mitochondrial Protein Malonylation and Metabolic Efficiency. *Cell Chem Biol,* **24**(6), 673–684.

[71] Ellis, J. M., Bowman, C. E., and Wolfgang, M. J. (2015) Metabolic and tissue-specific regulation of acyl-CoA metabolism. *PLoS One,* **10**(3), e0116587.

[72] Jaffrey, S. R. and Wilkinson, M. F. (12, 2018) Nonsense-mediated RNA decay in the brain: emerging modulator of neural development and disease. *Nat Rev Neurosci,* **19**(12), 715–728.

[73] Mitrovich, Q. M. and Anderson, P. (Sep, 2000) Unproductively spliced ribosomal protein mRNAs are natural targets of mRNA surveillance in C. elegans. *Genes Dev,* **14**(17), 2173–2184.

[74] Hansen, K. D., Lareau, L. F., Blanchette, M., Green, R. E., Meng, Q., Rehwinkel, J., Gallusser, F. L., Izaurralde, E., Rio, D. C., Dudoit, S., and Brenner, S. E. (Jun, 2009) Genome-wide identification of alternative splice forms down-regulated by nonsense-mediated mRNA decay in Drosophila. *PLoS Genet,* **5**(6), e1000525.

[75] Wight, M. and Werner, A. (2013) The functions of natural antisense transcripts. *Essays Biochem,* **54**, 91–101.

[76] Guo, Z., Maki, M., Ding, R., Yang, Y., Zhang, B., and Xiong, L. (Jun, 2014) Genome-wide survey of tissue-specific microRNA and transcription factor regulatory networks in 12 tissues. *Sci Rep,* **4**, 5150.

[77] Watanabe, T., Totoki, Y., Toyoda, A., Kaneda, M., Kuramochi-Miyagawa, S., Obata, Y., Chiba, H., Kohara, Y., Kono, T., Nakano, T., Surani, M. A., Sakaki, Y., and Sasaki, H.

(May, 2008) Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature,* **453**(7194), 539–543.

[78] Tam, O. H., Aravin, A. A., Stein, P., Girard, A., Murchison, E. P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R. M., and Hannon, G. J. (May, 2008) Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature,* **453**(7194), 534–538.

[79] Yu, J., Chen, M., Huang, H., Zhu, J., Song, H., Zhu, J., Park, J., and Ji, S. J. (02, 2018) Dynamic m6A modification regulates local translation of mRNA in axons. *Nucleic Acids Res,* **46**(3), 1412–1423.

[80] Ke, S., Pandya-Jones, A., Saito, Y., Fak, J. J., Vågbø, C. B., Geula, S., Hanna, J. H., Black, D. L., Darnell, J. E., and Darnell, R. B. (05, 2017) A mRNA modifications are deposited in nascent pre-mRNA and are not required for splicing but do specify cytoplasmic turnover. *Genes Dev,* **31**(10), 990–1006.

[81] Licht, K., Hartl, M., Amman, F., Anrather, D., Janisiw, M. P., and Jantsch, M. F. (01, 2019) Inosine induces context-dependent recoding and translational stalling. *Nucleic Acids Res,* **47**(1), 3–14.

[82] Mudge, J. M., Frankish, A., and Harrow, J. (Dec, 2013) Functional transcriptomics in the post-ENCODE era. *Genome Res,* **23**(12), 1961–1973.

[83] Kim, E., Ilagan, J. O., Liang, Y., Daubner, G. M., Lee, S. C., Ramakrishnan, A., Li, Y., Chung, Y. R., Micol, J. B., Murphy, M. E., Cho, H., Kim, M. K., Zebari, A. S., Aumann, S., Park, C. Y., Buonamici, S., Smith, P. G., Deeg, H. J., Lobry, C., Aifantis, I., Modis, Y., Allain, F. H., Halene, S., Bradley, R. K., and Abdel-Wahab, O. (May, 2015) SRSF2 Mutations Contribute to Myelodysplasia by Mutant-Specific Effects on Exon Recognition. *Cancer Cell,* **27**(5), 617–630.

[84] Gardner, L. B. (Jun, 2008) Hypoxic inhibition of nonsense-mediated RNA decay regulates gene expression and the integrated stress response. *Mol Cell Biol,* **28**(11), 3729–3741.

[85] Gerstberger, S., Hafner, M., Ascano, M., and Tuschl, T. (2014) Evolutionary conservation and expression of human RNA-binding proteins and their role in human genetic disease. *Adv Exp Med Biol,* **825**, 1–55.

Figures

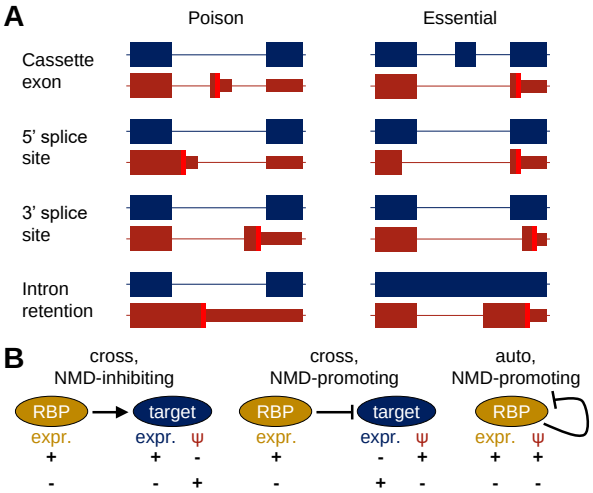

Figure 1: **(A)** Classification of USEs. An USE is referred to as poison/essential if the NMD isoform (red) represents an insertion/deletion compared to the protein-coding isoform (blue). PTCs are colored in bright red. **(B)** The expected changes of AS and gene expression level of the target caused by high (+) and low (−) expression levels of the regulator (RBP).

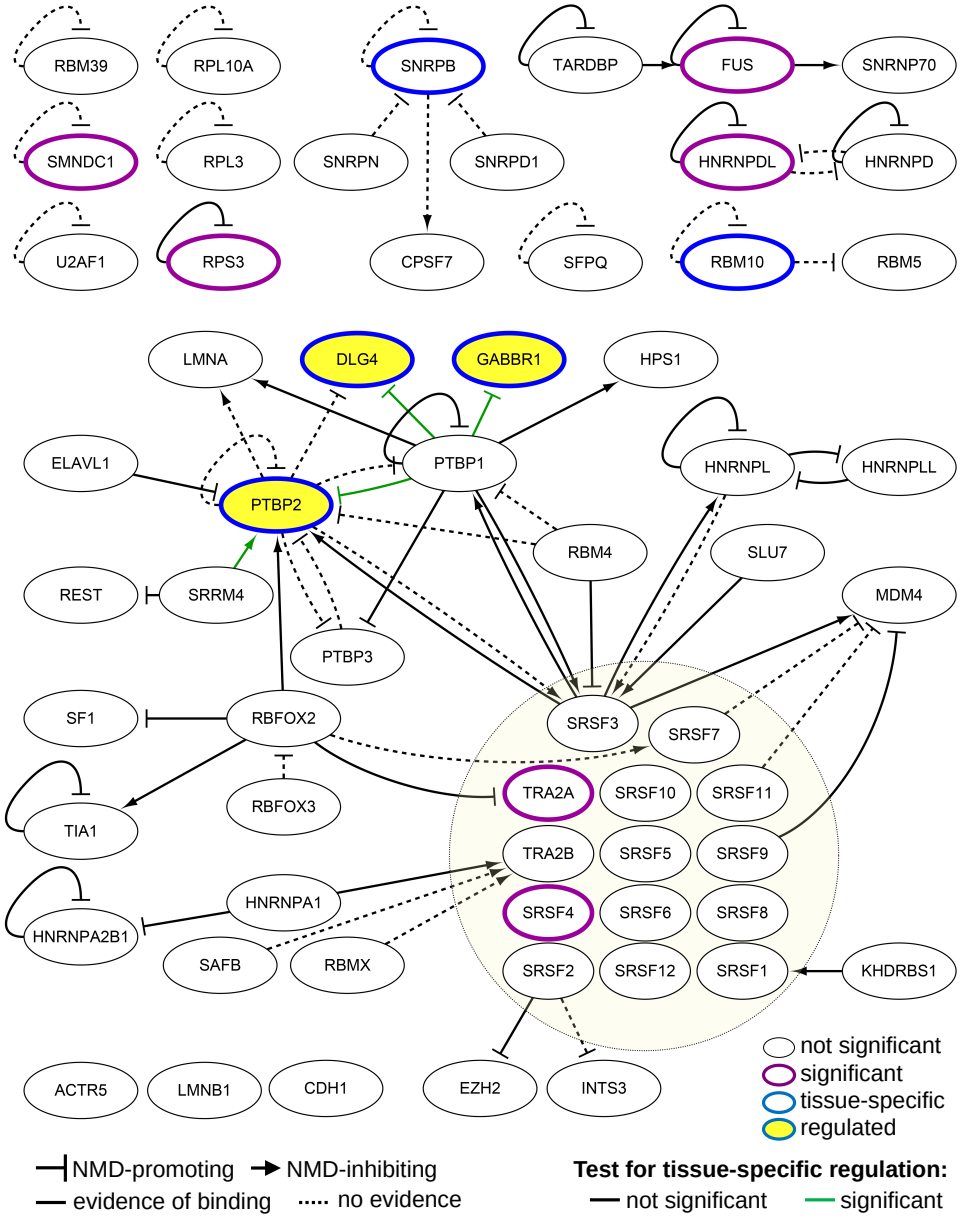

Figure 2: A regulatory network of validated USEs. A subnetwork of SR proteins is exempt to Figure S4. The nodes represent USEs listed in Table S8. The edges represent NMD-promoting and NMD-inhibiting regulatory connections. The color code of the nodes represents the classification of USEs as significant or tissue-specific (Table 1). The color code of the edges represent the result of the tissue-specificity test.

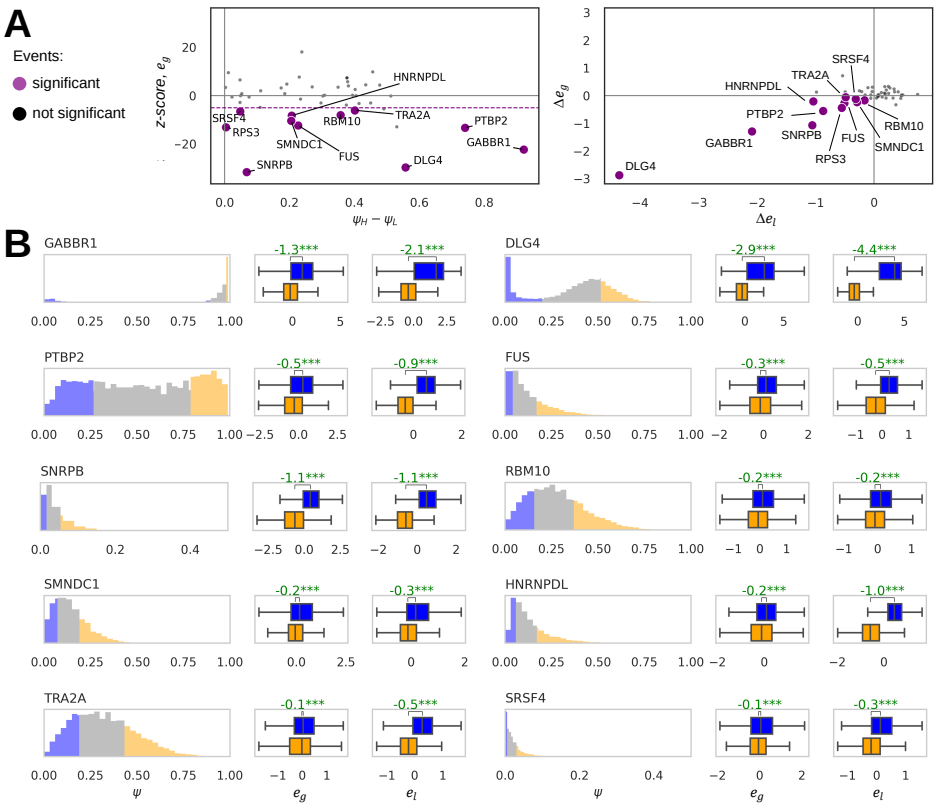

Figure 3: Significance of validated USEs. **(A)** Significant USEs are characterized by the z-score of e_g below -5, and negative Δe_g and Δe_l values. $\psi_H - \psi_L$ denotes the difference of the medians in the upper and the lower quartiles of the ψ distribution. **(B)** The distribution of ψ values for selected USEs from panel (A). The top 25% and the bottom 25% of ψ values are colored in orange and blue, respectively. The box plots represent the distribution of e_g and e_l in the two groups. Shown in green are Δe_g and Δe_l values. Asterisks (***) denote a statistically discernible difference at the 0.1% significance level.

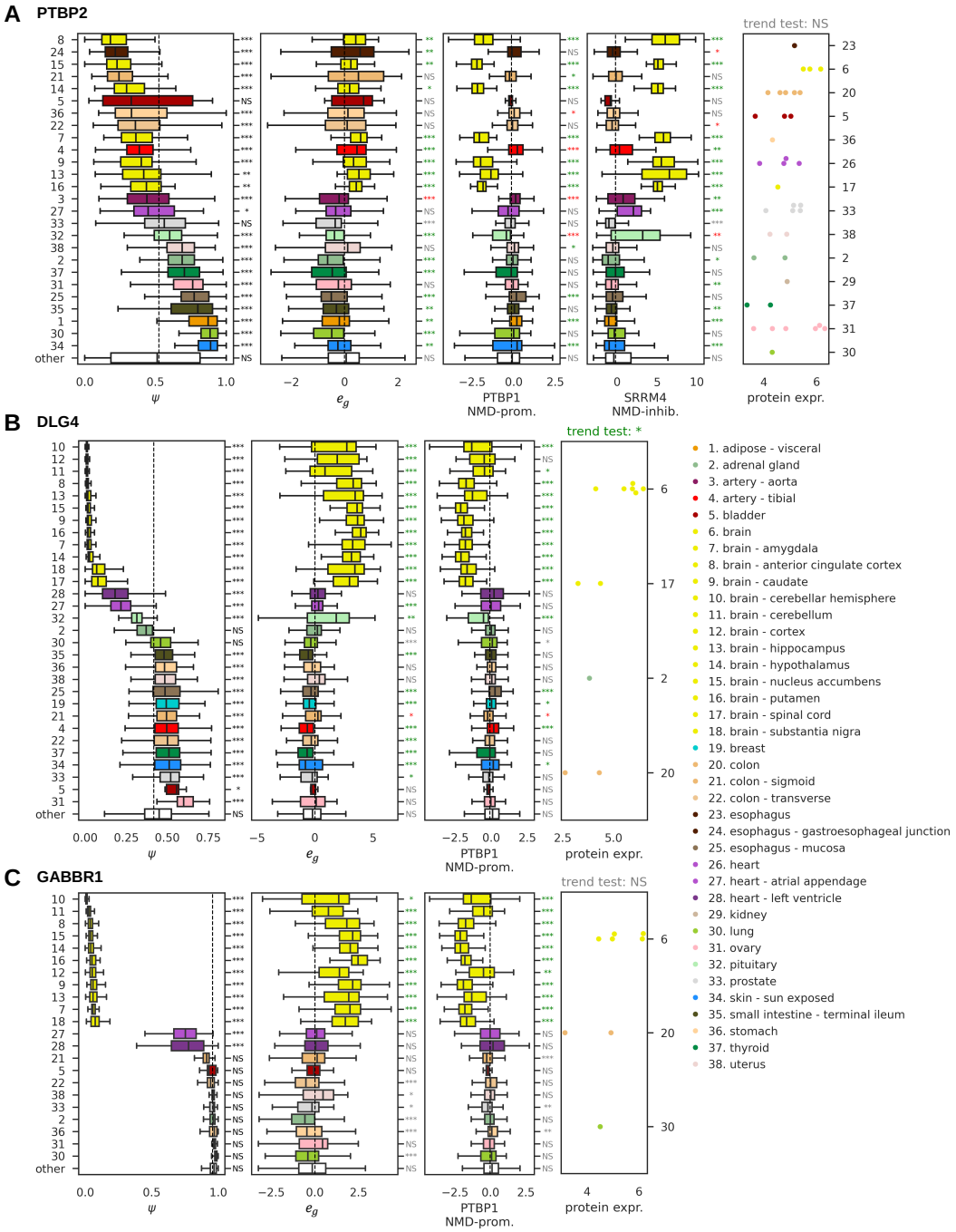

Figure 4: Examples of validated tissue-specific USEs. Panels (A), (B), and (C) show the results for *PTBP2*, *DLG4*, and *GABBR1* genes, respectively. In each case, boxplots show (left to right) the distribution of ψ, global gene expression level (e_g), the expression levels of regulators, and the protein expression level. Statistically significant deviation from the pooled median are marked with asterisks (***FDR < 0.001; **FDR < 0.01; *FDR < 0.05; NS not significant). The color of an asterisk shows whether the deviation is in the expected (green) or the opposite direction. Tissues are denoted by numbers and sorted in ascending ψ order.

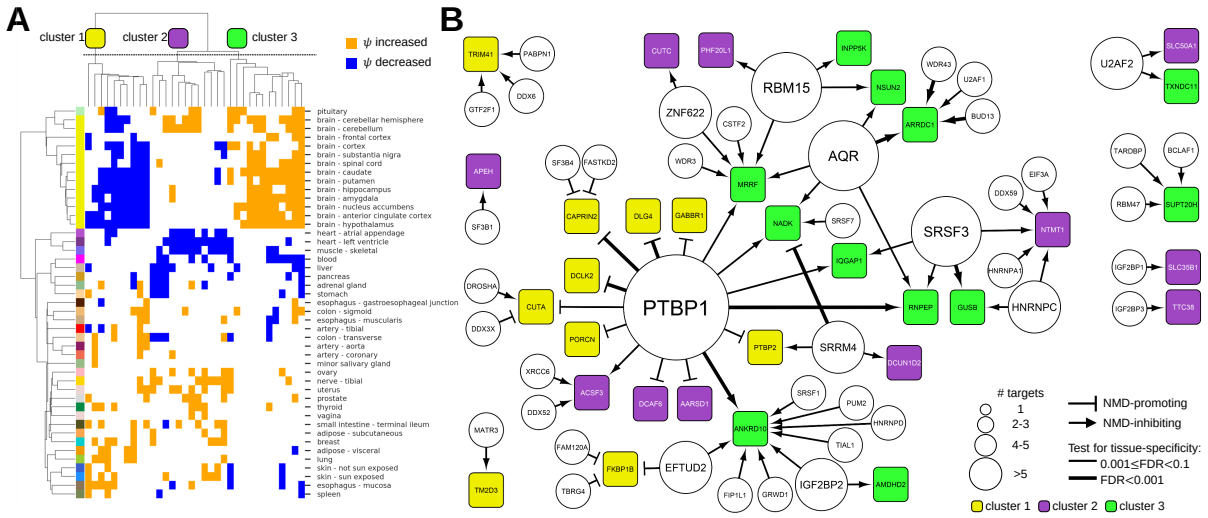

Figure 5: Tissue-specific regulation of USEs. (A) Clustering diagram of 34 regulated USEs with CLIP support in the gene based on tissue-specific deviations of ψ. The clusters are characterized by a decreased ψ in the brain (cluster 1), a decreased ψ in skeletal muscle and heart (cluster 2), and an increased ψ in the brain (cluster 3). **B** The predicted network of regulated USEs with CLIP support in the gene. The nodes and edges are as in Figure 2. The nodes from clusters 1-3 are labeled with the same colors as in panel (A).

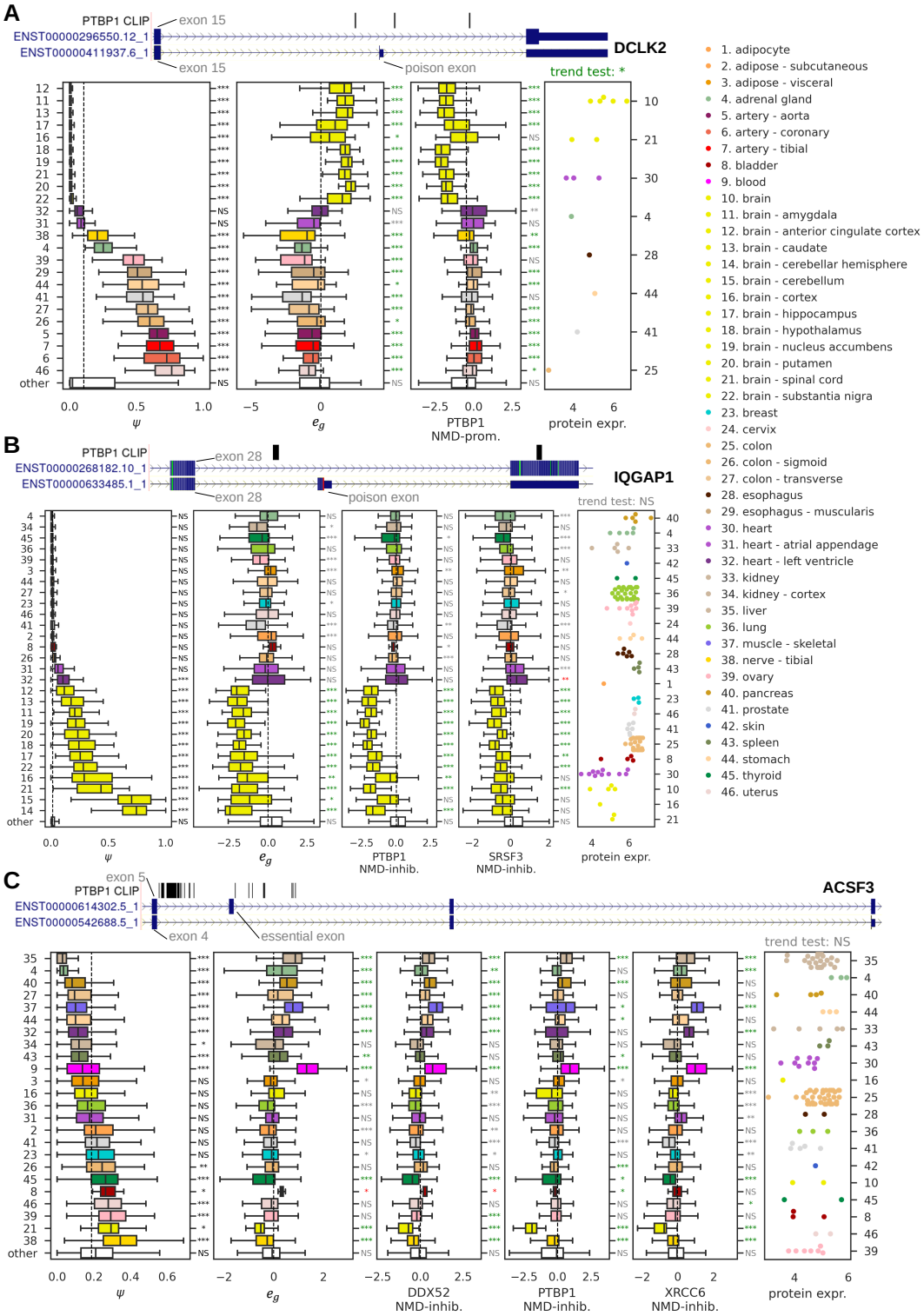

Figure 6: Examples of novel tissue-specific USEs. Panels (A), (B), and (C) show the results for *DCLK2*, *IQGAP1*, and *ACSF3* genes, respectively. The ideograms in each panel show the USE and CLIP peaks of *PTBP1*. The rest of the legend is as in Figure 4.