Quantitative prediction of variant effects on alternative splicing using endogenous pre-messenger RNA structure probing

Authors: Jayashree Kumar[a,b], Lela Lackey[a,c], Justin M. Waldern[a], Abhishek Dey[a], David H. Mathews[d], Alain Laederach[a,b,1]

Affiliations: [a] Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC

[b] Curriculum in Bioinformatics and Computational Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC

[c] Department of Genetics and Biochemistry, Center for Human Genetics, Clemson University, Greenwood, SC

[d] Department of Biochemistry & Biophysics and Center for RNA Biology, 601 Elmwood Avenue, Box 712, School of Medicine and Dentistry, University of Rochester, Rochester, NY 14642

Abstract:

Splicing is a highly regulated process that depends on numerous factors. It is particularly challenging to quantitatively predict how a mutation will affect precursor messenger RNA (mRNA) structure and the subsequent functional consequences. Here we use a novel Mutational Profiling (-MaP) methodology to obtain highly reproducible endogenous precursor and mature mRNA structural probing data in vivo. We use these data to estimate Boltzmann suboptimal ensembles, and predict the structural consequences of mutations on precursor mRNA structure. Together with a structural analysis of recent cryo-EM spliceosome structures at different stages of the splicing cycle, we determined that the footprint of the B^{act} complex on precursor mRNA is best able to predict splicing outcomes for exon 10 inclusion of the alternatively spliced *MAPT* gene. However, structure alone only achieves 74% accuracy. We therefore developed a β-regression weighting framework that incorporates splice site strength, structure and

exonic/intronic splicing regulatory elements which together achieves 90% accuracy for 47 known and six newly discovered splice-altering variants. This combined experimental/computational framework represents a path forward for accurate prediction of splicing related disease-causing variants.

Introduction

Precursor messenger RNA (pre-mRNA) splicing is a highly regulated process in eukaryotic cells (Z. Wang and Burge 2008). Numerous factors control splicing including *trans*-acting RNA-binding proteins (RBPs), components of the spliceosome, and the pre-mRNA itself. Pre-mRNA structure is a key attribute that directs splicing, particularly alternative splicing, but we have only a poor understanding of pre-mRNA structure-mediated splicing mechanisms (Taylor and Sobczak 2020). In addition, it is particularly challenging to develop quantitative models capable of predicting splicing outcome, specifically the Percent Spliced In (PSI) for alternatively spliced junctions. This difficulty is especially true for predicting the effects of genetic variation at exon-intron junctions. Indeed, mutations may affect not only the binding specificity of RBPs but also may alter pre-mRNA structure (Tazi, Bakkour, and Stamm 2009).

Similar to the challenge of predicting PSI outcomes, the consequences of mutations on pre-mRNA structure are difficult to predict. First and foremost, little is known about native pre-mRNA structure because pre-mRNAs are relatively short-lived in cells (Herzel et al. 2017). Only recently has pre-mRNA structure determination become amenable to high resolution in vivo experimental characterization (Mustoe et al. 2018). Second, it is not clear what structures of a pre-mRNA control spliceosome assembly and activity. Finally, we lack quantitative measures for the relative weighting of RBPs' affinity for specific motifs in pre-mRNA to the importance of pre-mRNA structure. Several technical developments address these issues and enable us to propose an integrated, RNA structure based-framework that accurately predicts the percent of splicing. In this study, we used a combination of endogenous pre-mRNA chemical

structure probing (Homan et al. 2014), an RNA structure model that considers multiple alternative structures in equilibrium (Dethoff et al. 2012; Lai et al. 2018), quantitative analysis of exonic and intronic splicing enhancers/silencers (Fairbrother et al. 2002; Z. Wang et al. 2004; Yang Wang, Ma, et al. 2012; Yang Wang, Xiao, et al. 2012), and a β-regression weighting (Ferrari and Cribari-Neto 2004).

In this work we measure endogenous pre-mRNA structure in vivo by combining recent developments in RNA structure Mutational Profiling (so-called -MaP approaches) with targeted amplification of specific exon-intron junctions. This novel approach enables us to obtain single-nucleotide RNA structure probing data for endogenous pre- and mature mRNAs in the same cell. The high reproducibility of these data also makes it possible to use Boltzmann suboptimal sampling guided by the data (Spasic et al. 2018) to predict free energies of unfolding for an ensemble of structures. In addition, we can now leverage recent high resolution cryo-Electron Microscopy (cryo-EM) structures of various stages of the spliceosome during the splicing cycle to reveal the effective spliceosomal footprint on pre-mRNA (L. Zhang et al. 2019).

As a model system to validate our framework, we study the effects of 47 experimentally measured mutations at the Exon10-Intron10 junction of the human Microtubule-Associated Protein Tau gene, *MAPT* (Park, Ahn, and Gallo 2016; Catarina Silva and Haggarty 2020). Exon 10 is a cassette exon that is alternatively spliced resulting in a Tau protein with either four microtubule binding repeats (4R) or three repeats (3R). The ratio of 3R to 4R isoforms is approximately 1:1 (Hefti et al. 2018). This is highly unusual for a splicing event as single-cell RNA-seq analysis demonstrates that this type of event, where alternative isoforms are expressed equally, comprises less than 20% of all splicing events (Song et al. 2017). The Exon10-Intron10 junction has 29 clinically validated disease-causing mutations (Stenson et al. 2003) that impair the function of Tau protein and are implicated in many neurodegenerative diseases (Spillantini et al. 1998; Hutton et al. 1998; Clark et al. 1998; Rizzu et al. 1999; Goedert et al. 1999). Although some mutations alter the Tau protein sequence (Mirra et al. 1999; Iseki et al.

2001), 20 disease-associated mutations are known that deregulate *MAPT* pre-mRNA splicing by altering the 1:1 ratio of 3R to 4R *MAPT* isoforms (Hutton et al. 1998; D'Souza et al. 1999; Hasegawa et al. 1999; Jiang et al. 2000). An additional 27 mutations were previously experimentally tested to measure Exon 10 PSI with splicing assays (D'Souza and Schellenberg 2000; Tan et al. 2019; Grover et al. 1999), making this junction the most experimentally characterized junction of clinical importance in the human genome and an excellent system for developing forward predictive models of splicing. Our work thus provides a framework for integrating endogenous pre-mRNA structure probing data with our current structural understanding of spliceosome assembly and *trans*-acting RBPs to achieve unprecedented quantitative prediction accuracy of the effect of mutations at structured exon-intron junctions.

Results

Median ratio of individual and tissue 3R to 4R MAPT mRNA isoforms is 1:1

Splicing of *MAPT* Exon 10 yields a 1:1 ratio of alternatively spliced isoforms (Goedert et al. 1989; Andreadis 2005). To corroborate the 1:1 isoform ratio among tissues and individuals, we analyzed RNA-sequencing data from the Genotype-Tissue Expression (GTEx) database (Lonsdale et al. 2013). We selected tissue types with median *MAPT* transcripts per million greater than 10 (Figure 1-figure supplement 1A) and calculated the Percent Spliced In (PSI) value for Exon 10 for each sample (Figure 1A-source data 1; Materials and methods). We examined the distribution of PSIs for each tissue type over 2,315 tissue samples in 375 individuals of median age 61 (Figure 1A; Figure 1-figure supplement 1B). A PSI of 0 indicated that none of the *MAPT* transcripts in a sample had Exon 10 spliced in (3R isoform), whereas a PSI of 1 corresponded to all *MAPT* transcripts having Exon 10 spliced in (4R isoform). We found variation in Exon 10 PSI both within and between different tissue types; the pituitary gland had the largest variation among brain tissues, and the cerebellum had the least variation but the difference between the two standard deviations was 0.04. Also, while the pituitary gland and caudate had the lowest and highest median Exon 10 PSI respectively among

individual samples, the distance between the two values was only 0.25. Interestingly, although *MAPT*'s function in breast tissue is not understood compared with its function in the brain, for breast tissue, individuals had greater variation in Exon 10 PSI and a lower median PSI compared with the pituitary gland (Figure 1-figure supplement 1B). We also discovered a large amount of variation within tissues of an individual (Figure 1-figure supplement 1C), although there was significantly greater variation between individuals than within a single individual (see Supplementary file 1 for ANOVA table). Overall, 75% of samples were within a standard deviation of the median PSI of 0.54, which confirmed that the 3R to 4R isoform ratio was approximately 1:1 among individuals and within different tissue types. The consistency of this isoform ratio, despite the likely presence of different levels of RBPs, suggest that inherent sequence and structural features regulate splicing at this exon-intron junction. RNA structure regulates alternative splicing around exon-intron junctions (Warf and Berglund 2010; Buratti and Baralle 2004) and a hairpin structure at the exon 10-intron 10 junction is implicated in establishing the 3R to 4R 1:1 isoform ratio (Hutton et al. 1998; Varani et al. 1999; Grover et al. 1999; Donahue et al. 2006). Hence, we next used high-throughput chemical mapping techniques to interrogate the endogenous in vivo structure of the *MAPT* junction.

Structure of 3R and 4R MAPT mature mRNA isoforms is open and accessibility of exons is similar for the two isoforms

Although the structure of the *MAPT* pre-mRNA was previously studied computationally and in vitro (Varani et al. 1999; Lisowiec et al. 2015; Tan et al. 2019; Chen et al. 2019), the structures of the mature 3R and 4R isoforms and *MAPT* pre-mRNA have not been assessed in their endogenous in vivo context. We used dimethyl sulfate (DMS) to chemically probe RNA structure in T47D and neuronal SH-SY5Y cells and primer-amplified the Exon 9-Exon 11 and Exon 9-Exon 10-Exon 11 junctions during library preparation for Mutational-Profiling (-MaP) (Figure 1B; Materials and methods). This approach leverages the read-through aspect of MaP technology to probe the structure of two alternatively spliced isoforms in the same cells. DMS reactivities for replicates,

and between T47D and SH-SY5Y MAPT mRNAs were highly correlated (Figure 1-figure supplement 2A; Figure 1-figure supplement 2B; Figure 1-figure supplement 2D; Figure 1-figure supplement 2E).

We also collected in vivo DMS data for the small ribosomal RNA (SSU) whose secondary structure is known from X-ray crystallography (Petrov et al. 2014) (Figure 1-figure supplement 3A). The DMS reactivities for unpaired nucleotides in the SSU were significantly higher than for paired nucleotides (Figure 1-figure supplement 3B), confirming our probing strategy accurately recapitulates RNA secondary structure. We used the SSU in vivo data to calibrate the estimation of equilibrium ensembles as guided by MaP technology (Methods and materials), and we validated that structure prediction guided by experimental DMS reactivities yielded more accurate estimation of the SSU structure (Figure 1-figure supplement 3C). The median DMS reactivity of the mature *MAPT* isoforms was 0.22, significantly greater than the median DMS reactivity of the SSU, 0.0083 (Figure 1-figure supplement 3D); these results suggested that the nucleotides of the mature *MAPT* isoforms were more accessible and unpaired compared with the highly structured SSU, indicating that our endogenous in vivo probing strategy reveals important differences in the structure of cellular RNAs.

Reactivities of Exon 9 and Exon 11 were highly correlated between the 3R and 4R isoforms (Figure 1-figure supplement 2C). Additionally, computed base-pairing probabilities guided by the experimental data for the two isoforms revealed that, although there were some long-range interactions, 66% of base pairs spanned less than 50 nucleotides and were contained within the exon units (Figure 1B). This result suggested that the mature exons function as their own structural unit. However, the mature isoform structures did not suggest how they might regulate splicing of Exon 10. Hence, we next chemically probed the MAPT pre-mRNA.

Figure 1

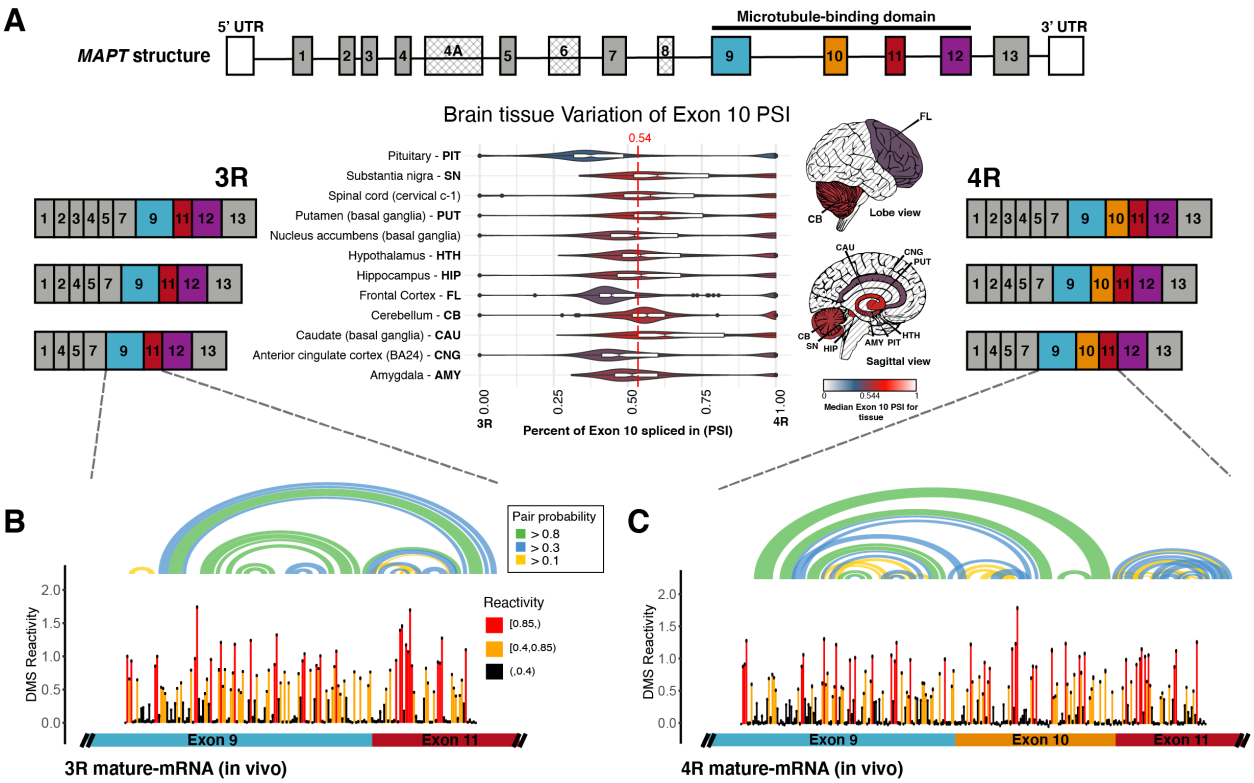


Figure 1: In vivo DMS-MaP structure probing data for 3R and 4R mature *MAPT* transcripts that are expressed in a 1:1 ratio.

A) Ratio of 3R and 4R MAPT transcripts is approximately 1:1 among brain tissues. There are 14 exons alternatively spliced in *MAPT*. Exons 4A, 6, and 8 are not found in brain mRNA. The four exons highlighted in color are repeat regions that form the microtubule binding domain in the Tau protein. Exon 10 is alternatively spliced to form the 3 repeat (3R) or 4 repeat (4R) isoform. The six canonical transcripts found in the central nervous system can be separated into 3R and 4R transcripts. Percent Spliced In (PSI) of Exon 10 was calculated from RNA-seq data for 2315 tissue samples representing 12 central nervous system tissue types and collected from 375 individuals in GTEx v8 database. The violin plot for each tissue type and the corresponding region on the brain diagram is colored by the median PSI for all samples of a given type. The patterned regions on the brain diagram indicate tissue types with no data. Tissue types Spinal cord and

Nucleus accumbens are not visualized on the brain diagram. The median PSI of 0.54 among all tissue samples is indicated by the red dotted line.

B) In vivo DMS-MaP structure probing data across exon9-exon11 junction of 3R mature MAPT transcript. T47D cells were treated with DMS. Structure probing data for junctions of interest were obtained using primers (Supplementary file 4) following RT of extracted RNA. DMS reactivity is plotted for each nucleotide across spliced junctions. Each value is shown with its standard error and colored by reactivity based on color scale. High DMS reactivities correspond to unstructured regions, whereas low DMS reactivities correspond to structured regions. The base pairs of the predicted secondary structure guided by DMS reactivities are shown in the arcs colored by pairing probabilities.

C) In vivo DMS-MaP structure probing data across exon9-exon10-exon11 junction of 4R mature MAPT transcript

MAPT pre-mRNA Exon 10-Intron 10 junction is more structured compared with the mature isoforms

Existence of a hairpin at the *MAPT* Exon 10-Intron 10 junction, implicated in regulating Exon 10 splicing, was established by NMR and in vitro chemical probing (Varani et al. 1999; Lisowiec et al. 2015); however, the endogenous in vivo structure of this region has yet to be determined. While collecting data for mature *MAPT* isoform junctions, we simultaneously obtained data for the pre-mRNA Exon 10-Intron 10 junction (Figure 2A; Materials and methods). Replicates were highly correlated (Figure 2-figure supplement 1A). Surprisingly, although Exon 10 was still being spliced, the reactivities for Exon 10 in pre-mRNA and the mature 4R isoform were highly correlated (Figure 2-figure supplement 1B). Again, base pairing between nucleotides appeared to be contained within exons, independent of introns. The reactivities were highly correlated between data collected in SH-SY5Y and T47D cells (Figure 2-figure supplement 1C); thus, despite likely differences in RBP concentrations, the structure of the pre-spliced region is the same between cell lines. Additionally, we found lower DMS reactivities for the pre-mRNA Exon 10-Intron 10 junction compared with the mature isoform junctions (Figure 2-figure supplement 1D), which suggests that pre-mRNA is more structured than mature mRNA. We uncovered strong evidence for the previously in vitro identified hairpin structure in the DMS reactivity data; pairing probabilities were greater than 0.8 for the entire hairpin stem (Figure 2A).

Shifts in structural ensemble of MAPT Exon 10-Intron 10 junction associated with disease mutations correlate with changes in splicing level of Exon 10

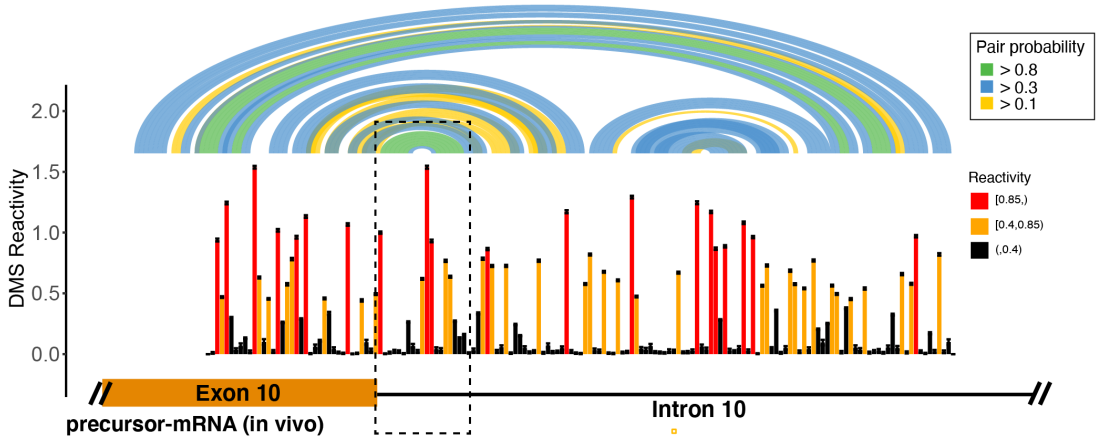
Many RNAs inhabit multiple conformations in vivo to form a structural ensemble instead of a single rigid structure (Halvorsen et al. 2010; Adivarahan et al. 2018). We posit that a structural ensemble at the *MAPT* Exon 10-Intron 10 junction regulates Exon 10 splicing and disease mutations alter the composition of the structural ensemble to disrupt splicing.

We used Boltzmann sampling of RNA structures guided by DMS reactivity data (Spasic et al. 2018) (Materials and methods) to sample 1000 structures for the wild type and two mutant intronic sequences, +15A>C and +19C>G. The two mutations alter, in opposite directions, the isoform ratio at this junction (Tan et al. 2019). We visualized the structural ensemble for the 3000 structures using t-Distributed stochastic neighbor embedding (t-SNE) (Van Der Maaten and Hinton 2008) (Figure 2B; Materials and methods). Each structure is a dot and is colored by the $\Delta G°$ of unfolding of the 5' splice site defined as the last three nucleotides of Exon 10 and the first six nucleotides of Intron 10 (Yeo and Burge 2004). The lower the unfolding free energy, the easier to unfold the structure. Overall, although there was a range of unfolding free energies for the three ensembles, there were three distinct populations of free energies for the three sequences (Figure 2-figure supplement 2A). We used k-means clustering to identify representative structures for each cluster (Figure 2B; Figure 2-figure supplement 2B; Materials and methods). We quantified and visualized the density of the clusters (Figure 2C; Materials and methods) and revealed distinct regions in the structure space occupied by each sequence. More than 55% of structures in the ensemble of the +19C>G mutation, which shifts the isoform balance entirely 3R (3R mutation) (Figure 2C inset), clustered in the lower left quadrant with larger unfolding free energies for the splice site. This result was evidenced by the highly base-paired exon-intron junction in the representative structure for the cluster. Hence, in the presence of the 3R mutation, the structural ensemble of the junction shifted towards more closed structures. Conversely, greater than 50% of structures in the ensemble of the +15 A>C mutation, which shifts the isoform balance entirely 4R (4R mutation) (Figure 2C inset), were clustered in the upper left quadrant with lower unfolding free energies for the splice site. The representative structure for this region was more open and accessible around the exon-intron junction. Correspondingly, the wild-type sequence had structures distributed across the entire space consistent with an ensemble of structures. The exon-intron junction of the representative structure for this region was not as accessible with the 4R mutation, but it had fewer base-pairs than with the 3R mutation, a result recapitulated by

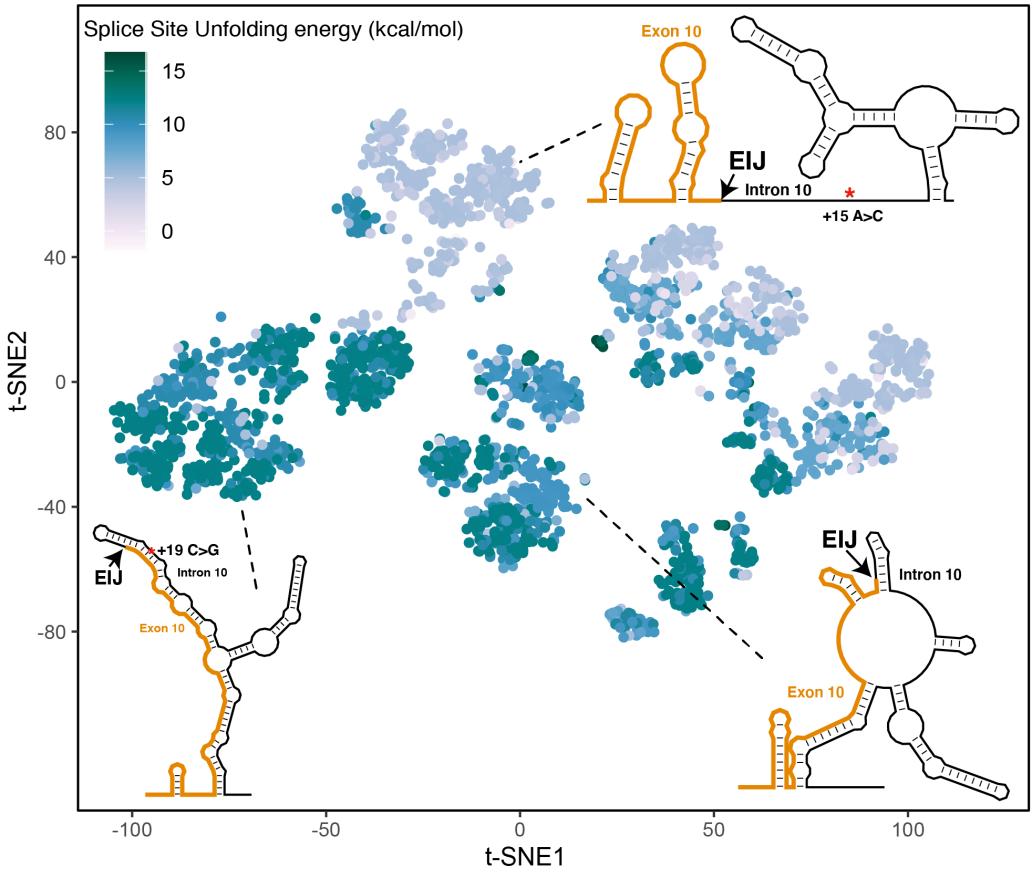
the two other representative structures in the right quadrants (Figure 2-figure supplement 2B).

Figure 2

A


Pair probability
> 0.8
> 0.3
> 0.1
Reactivity
[0.85,)
[0.4,0.85)
(,0.4)
DMS Reactivity
2.0
1.5
1.0
0.5
0.0
Exon 10
Intron 10
precursor-mRNA (in vivo)

B


Splice Site Unfolding energy (kcal/mol)
15
10
5
0
Exon 10
EIJ
Intron 10
+15 A>C
+19 C>G
t-SNE2
t-SNE1
80
40
0
-40
-80
-100
-50
0
50
100

C

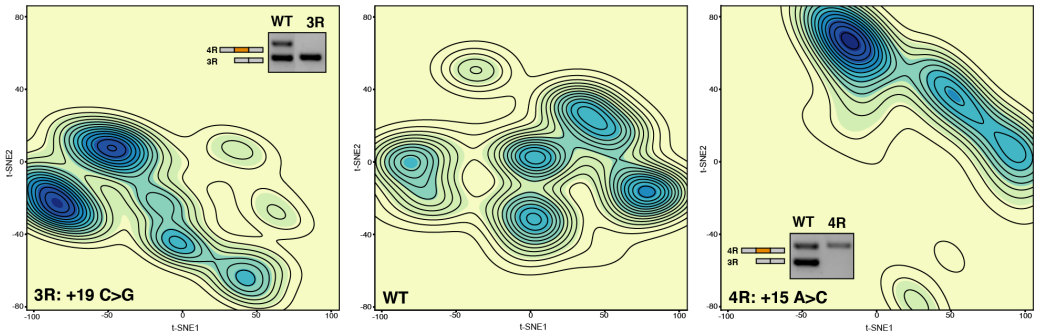

WT
3R
4R
3R: +19 C>G
WT
4R: +15 A>C

Figure 2: The 4R and 3R mutations shift DMS reactivity-guided structural ensemble of Exon 10-Intron 10 junction to more open and closed structures, respectively.

A) In vivo DMS-MaP structure probing data across Exon 10-Intron 10 junction of precursor MAPT transcript. T47D cells were treated with DMS. Structure probing data for junctions of interest were obtained using primers (Supplementary file 4) following RT of extracted RNA. DMS reactivity is plotted for each nucleotide. Each value is shown with its standard error and colored by reactivity based on the color scale. High DMS reactivities correspond to unstructured regions, whereas low DMS reactivities correspond to structured regions. Base pairs of predicted secondary structure guided by DMS reactivities are shown by arcs colored by pairing probabilities. Strongly predicted hairpin structure near exon-intron junction is highlighted by dotted box.

B) t-SNE Visualization of structural ensemble of wildtype (WT) and, +19C>G (3R) and +15A>C (4R) mutations. Structures were predicted using Boltzmann suboptimal sampling and guided by DMS reactivity data generated in A. Data were visualized using t-Distributed Stochastic Neighbor Embedding (t-SNE). Shown are 3000 structures corresponding to 1000 structures per sequence. Each dot represents a single structure and is colored by calculated unfolding free energy of splice site at exon-intron junction (3 exonic, 6 intronic bases). Data were clustered by k-means clustering and representative structures for three of the clusters are shown. The exon-intron junction is marked by EIJ on each structure. Positions of 3R and 4R mutations are marked by a red asterisk on their respective representative structures.

C) Density contour plots of structural ensemble of WT and, 3R and 4R mutations. Contour plots were derived from the distribution of points on the t-SNE plot in B. The darker the blue, the higher the density of structures. Contour lines additionally represent density of points. Color scales for the three plots are identical. Gel insets of RT-PCR products from splicing assays in HEK293 cells for 3R and 4R mutation are in their respective density plots.

Unfolding mRNA within the spliceosome B^{act} complex yields best prediction of Exon 10 splicing level

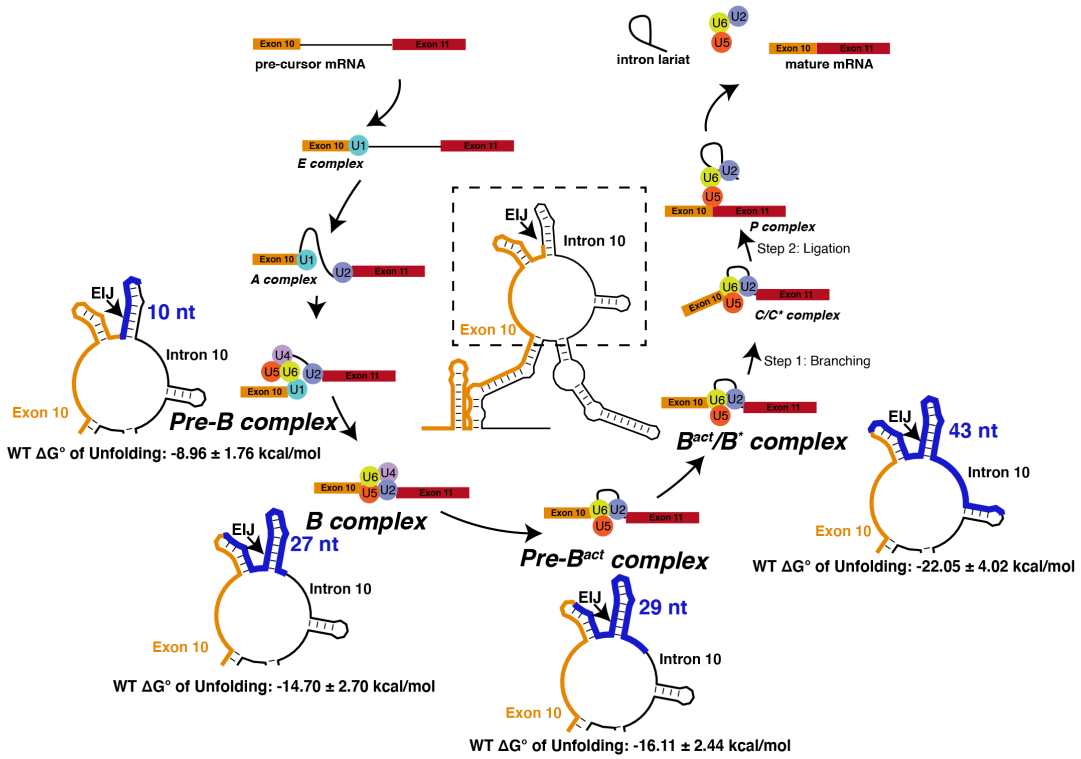
RNA structure controls alternative splicing by hindering or aiding accessibility of key regulatory regions to spliceosome components (McManus and Graveley 2011; Warf and Berglund 2010). The 5' splice site, defined as the last 3 nucleotides of the exon and first 6 nucleotides of the intron, is the minimum region of RNA that must be accessible for base pairing with the U1snRNA (Blanchette and Chabot 1997; Singh, Singh, and Androphy 2007). However, the splicing cycle, orchestrated by the spliceosome, traverses multiple stages to prepare the pre-mRNA and catalyze the two-step splicing reaction (Matera and Wang 2014) (Figure 3A). The RNA itself adopts many conformations as different components of the spliceosome bind to it (L. Zhang et al. 2019). In addition to the 5' splice site, a larger segment of RNA likely needs to unpair to accommodate the changing conformations induced by the spliceosome. We analyzed high resolution Cryo-EM structures of the human spliceosome Pre-B, B, Pre-B^{act}, and B^{act} complexes (Charenton, Wilkinson, and Nagai 2019; Bertram et al. 2017; Townsend et al. 2020; X. Zhang et al. 2018) to quantify the number of nucleotides around the 5' splice site for which sufficient density was observed in the cryo-EM structure and which were unpaired (Materials and methods). As can be seen in Figure 3A, the number of unpaired pre-mRNA nucleotides observed in each structure increased through the splicing cycle. Thus, it is likely that RNA structures outside the U1snRNA binding site have to be unfolded to accommodate splicing.

To evaluate the footprint of the spliceosome that best predicts splicing outcome, we initially focus on predicting 20 synonymous and intronic mutations as a training set. These mutations are most likely to have a structural component to their function (Sharma et al. 2019; Lin, Taggart, and Fairbrother 2016) (Figure 3-figure supplement 1A). The distribution of $\Delta G°$ of unfolding of the splice sites in the presence of these mutations was correlated with Exon 10 PSI (Figure 3-figure supplement 1B). We calculated the $\Delta G°$ of unfolding of the RNA near the 5' splice site in the four splicing stages' footprints. Features of the unfolding free energy distribution including mean and

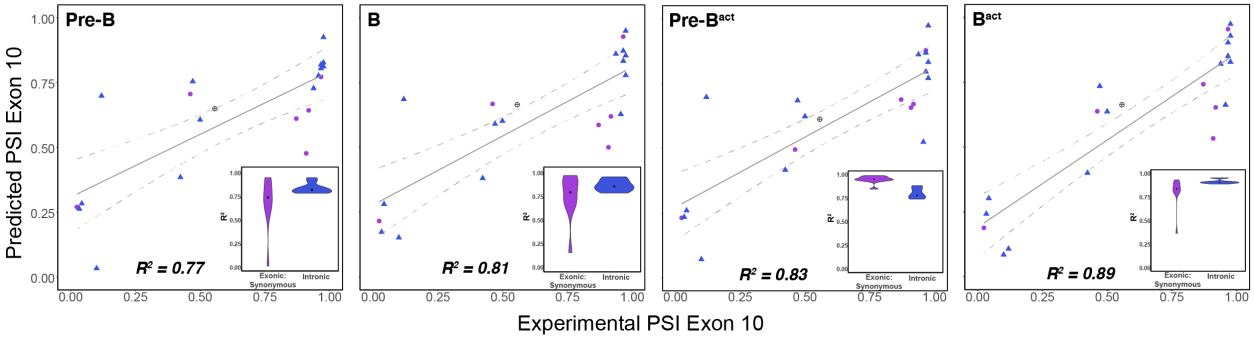
standard deviation were then used in a beta regression to predict Exon 10 PSI (Materials and methods; Eq. **1**). Unfolding larger regions of the exon-intron mRNA junction improved the predictive power of the model, and the B^{act} complex footprint yielded the best prediction accuracy (R^2 = 89%; Figure 3B). Crucially, we found that using features of the distribution of unfolding free energies in the structural ensemble produced greater predictive power than simply using the unfolding free energy of a single minimum free energy (MFE) structure (Figure 3-figure supplement 1C). We performed bootstrapping cross-validation and confirmed that unfolding the RNA within the B^{act} spliceosome complex yielded the best prediction (Figure 3C). We tested the structural ensemble-based model on 24 non-synonymous and compensatory mutations. Although the model performed well for compensatory mutations (median bootstrapped R^2=0.76), it yielded significantly less accurate predictions for non-synonymous mutations (median bootstrapped R^2=-0.21) (Figure 3-figure supplement 1D). One possible reason this structure-only model has limited performance is that it does not account for the effects of mutations on potential splicing regulatory elements (SREs) in the sequence.

Figure 3

A


pre-cursor mRNA
Exon 10
Exon 11
E complex
A complex
Pre-B complex
WT ΔG° of Unfolding: -8.96 ± 1.76 kcal/mol
EIJ
10 nt
Intron 10
B complex
27 nt
WT ΔG° of Unfolding: -14.70 ± 2.70 kcal/mol
Pre-Bact complex
29 nt
WT ΔG° of Unfolding: -16.11 ± 2.44 kcal/mol
Bact/B* complex
43 nt
WT ΔG° of Unfolding: -22.05 ± 4.02 kcal/mol
Step 1: Branching
C/C* complex
Step 2: Ligation
P complex
intron lariat
mature mRNA
U1
U2
U4
U5
U6

B


Pre-B
B
Pre-Bact
Bact
Predicted PSI Exon 10
Experimental PSI Exon 10
R² = 0.77
R² = 0.81
R² = 0.83
R² = 0.89
Exonic: Synonymous
Intronic

C

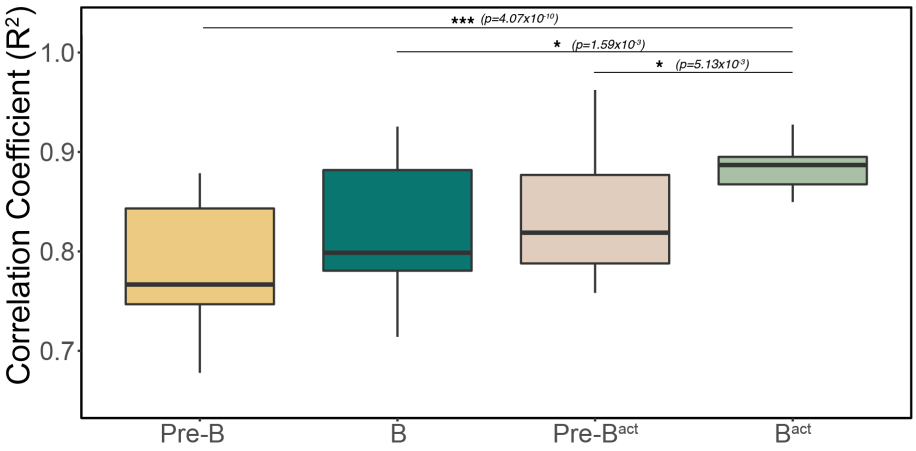

Correlation Coefficient (R²)
*** (p=4.07x10⁻¹⁰)
* (p=1.59x10⁻³)
* (p=5.13x10⁻³)
Pre-B
B
Pre-Bact
Bact

Figure 3: The best predictor of Exon 10 PSI for intronic and synonymous mutations was the unfolding free energy of pre-mRNA during the B^{act} stage of splicing

A) Spliceosome footprint on pre-mRNA during splicing cycle. Structure in the center of the cycle is the WT representative structure from Fig 2B. The dotted box indicates the zoomed-in region at each stage of interest. Cryo-EM structures of the human spliceosome complex at stages Pre-B (PDB ID: 6QX9), B (PDB ID: 5O9Z), Pre-B^{act} (PDB ID: 7ABF) and B^{act} (PDB ID: 5Z56) are available in the Protein Data Bank. The region around the 5' splice site of pre-mRNA within the spliceosome at each stage is highlighted in blue on the zoomed-in representative structure. The number of nucleotides for each stage is as follows: Pre-B (2 exonic, 8 intronic); B (10 exonic, 17 intronic); Pre-B^{act} (9 exonic, 20 intronic); B^{act} (12 exonic, 31 intronic). These values represent the minimum number of nucleotides required to be unfolded to be accessible to the spliceosome. The mean free energy and standard error to unfold RNA within the spliceosome at each stage is calculated for the WT structural ensemble and indicated under the zoomed-in structure.

B) Exon 10 PSIs of synonymous and intronic mutations predicted with the unfolding free energy of pre-mRNA within the spliceosome in B, Pre-B, Pre-B^{act}, B^{act} stages versus corresponding experimental PSIs measured in splicing assays. Exon 10 PSIs were predicted using Eq. 1. Grey line represents the best fit with dotted lines indicating the 95% confidence interval. Pearson correlation coefficients (R^2) of experimental to predicted PSIs were calculated for each stage. Violin plots (inset) show R^2s calculated for each mutation category by training and testing on subsets of all mutations by non-parametric bootstrapping; Synonymous (n=6), Intronic (n=14), Wildtype (n=1).

C) Overall Pearson correlation coefficients (R^2) calculated for experimental versus predicted Exon 10 PSIs by nonparametric bootstrapping of mutations. Subsets of mutations were randomly sampled 10 times, trained and tested using unfolding free energy of the exon-intron junction region of pre-mRNA within the spliceosome for the respective splicing stage. Pearson's R^2 was calculated by

comparing predicted PSIs to experimental PSIs. Level of significance: ***p-value $< 10^{-6}$, **p-value < 0.001, * p-value < 0.01

Effect of exonic non-synonymous mutations was best predicted by motif strength changes of splicing regulatory elements

Exon 10 splicing is highly regulated by differential binding of RBPs to *cis*-SREs within exon 10 and intron 10 (Qian and Liu 2014). While our structure-only model performs moderately well for 47 mutations (R^2=0.74) (see Supplementary file 2 for further details about mutations), *MAPT* Exon 10 PSIs of non-synonymous mutations were poorly predicted (median bootstrapped R^2 = -0.21, Figure 4-figure supplement 1B). Hence, we investigated whether these non-synonymous mutations are predicted better by incorporating changes to the strength of adjacent SREs. We identified SREs by similarity to reported general enhancer and silencer hexamer motifs (Fairbrother et al. 2002; Z. Wang et al. 2004; Yang Wang, Ma, et al. 2012; Yang Wang, Xiao, et al. 2012) (Materials and methods). We calculated the changes to splice site, enhancer, and silencer motif strengths in the presence of a mutation (Materials and methods) and visualized the motif strength changes in a heatmap (Figure 4A). We found that using splice site strength as the sole predictor yielded poor prediction of Exon 10 PSI in all mutation categories (Figure 4B; Eq. 3) because most mutations were outside the splice site. We quantified a weak positive correlation between PSI and enhancer strength and a significant negative correlation between PSI and silencer strength (Figure 4A; Figure 4-figure supplement 1C). We modeled Exon 10 PSI with the changes to the motif strength of all splicing regulatory elements (Eq. 4) and found an increase in prediction accuracy (R^2=0.51; Figure 4C) compared with using only splice site strength (R^2=0.29). Non-synonymous mutations were predicted more accurately using SRE strength with a median bootstrapped R^2 of 0.75.

Many RBPs have been identified that regulate *MAPT* Exon 10 splicing (Qian et al. 2011; Ian D'Souza and Schellenberg 2006; Kondo et al. 2004; J. Wang et al. 2004; L. Gao et al. 2007; S. Ding et al. 2012; Broderick, Wang, and Andreadis 2004; Yan Wang et al. 2010; Kar et al. 2006, 2011; P. Ray et al. 2011). To determine whether these proteins specific to Exon 10 splicing would improve the model's accuracy, we calculated changes to the strength of their RBP motifs obtained from high throughput sequencing

of bound RNAs (Dominguez et al. 2018; D. Ray et al. 2013) (Materials and methods). Unlike SRE motifs, there was no clear pattern or correlation between motif strength change and PSI (Figure 4-figure supplement 2A, B). Subsequently, the model's prediction accuracy was lower (R^2=0.08, Figure 4-figure supplement 2C), and changes to the strength of general SRE motifs were better predictors of Exon 10 PSI.

Figure 4

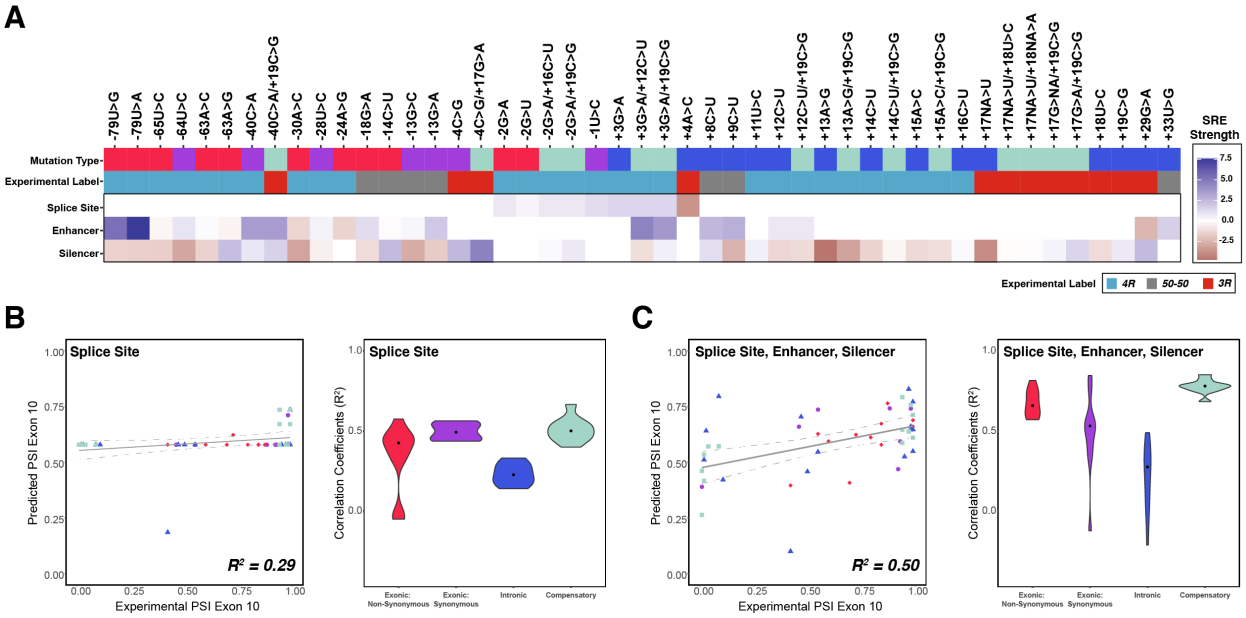


Figure 4: Combining the strength of all splicing regulatory elements improves prediction of Exon 10 PSI by 75% compared with using only splice site strength

A) Heatmap of splicing regulatory element (SRE) relative strength for 47 mutations compared with wildtype (WT). A value of 0 indicates mutation does not change WT SRE strength, positive values indicate SRE strength is greater than WT, and negative values indicate SRE strength is weaker than WT. Splice site strengths were calculated using MaxEntScan; a splice site was defined as the last 3 nucleotides of the exon and first 6 nucleotides of the intron. Enhancer and silencer strengths were calculated from position weight matrices of known motifs derived from cell-based screens (Materials and methods).

B) Exon 10 PSIs of 47 mutations predicted from change in splice site strength and plotted against experimental PSIs measured in splicing assays. Exon 10 PSIs predicted using Eq. 3. Each point on the scatterplot represents a mutation and is colored by mutation category. Grey line represents the best fit with dotted lines indicating the 95% confidence interval. Pearson correlation coefficient (R^2) calculated of experimental to predicted PSIs. Violin plot shows R^2s calculated for each category by training and testing on subsets of all mutations by non-

parametric bootstrapping; Exonic non-synonymous (n=11), Exonic synonymous (n=7), Intronic (n=15), Compensatory (n=14), Wildtype (n=1).

C) Exon 10 PSIs of 47 mutations predicted by combining change in splice site, enhancer, and silencer strength and plotted against experimental PSIs measured in splicing assays. Exon 10 PSIs predicted using Eq. 4.

Model with both structural and SRE motif changes yields best prediction of Exon 10 PSI

Our quantitative models showed that, although SRE motif changes accurately predicted the effects of non-synonymous mutations, structural changes were a better predictor of splicing outcomes of intronic and synonymous mutations. Combining all features (Eq. 6) yielded the highest prediction accuracy (R^2 = 0.89) (Figure 5A). This combined interactive model consistently produced more accurate predictions of Exon 10 PSI compared with a structure-only model and an SRE-only model for all mutation categories (Figure 5B). An additive model (Eq. 7) had lower prediction accuracy (R^2= 0.80) (Figure 5-figure supplement 1A), and this lower accuracy resulted primarily from less accurate PSI predictions of non-synonymous mutation effects (Figure 5-figure supplement 1B).

To determine whether structure or SRE changes were responsible for the splicing changes from each mutation, we hierarchically clustered the four primary features for the 47 experimentally validated mutations (Materials and methods). Six categories emerged from the clustering of features (Figure 5C) where approximately 80% of mutations modified both structure and silencer strength (Figure 5-figure supplement 1C). Further, we found that for more than 50% of mutations both structure and SRE motif strength were altered in the same direction and accordingly promoted Exon 10 splicing in that direction (Figure 5D). For the remaining mutations in which structure and SRE strength changed in opposite directions, structure dominated the direction of splicing for 18% of mutations, and SRE strength was dominant for 20% (Figure 5D). Overall, these results supported our conclusion that both structure and SREs have equally important effects in regulating splicing at this exon-intron junction.

Figure 5

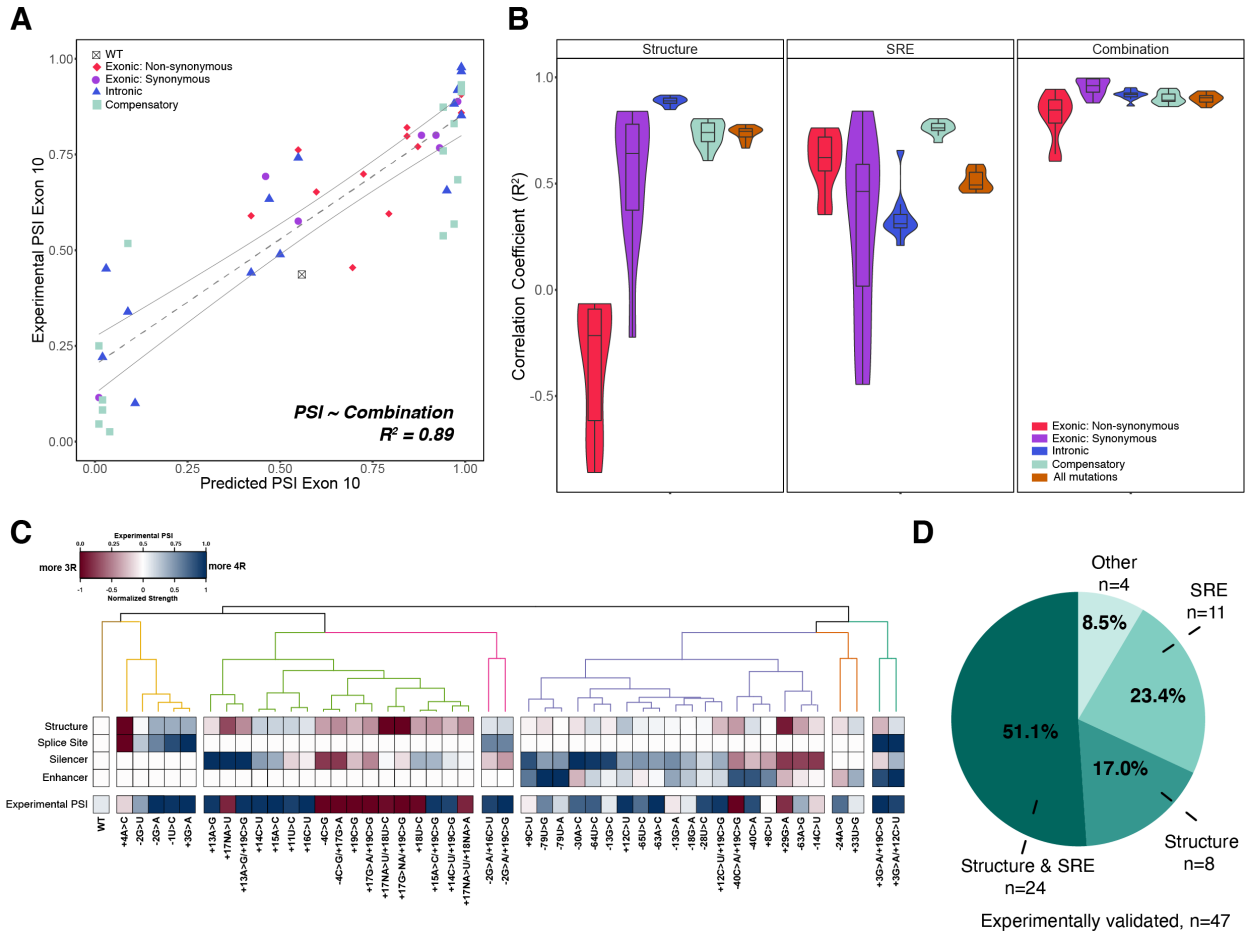


Figure 5: Combining structure and SRE strength into a unified model is the best predictor of Exon 10 PSI

A) Exon 10 PSIs of 47 mutations predicted from combined model using structure and SRE strength and fit to experimental PSIs measured in splicing assays. Exon 10 PSIs predicted using Eq. 6. Each point on scatterplot represents a mutation and is colored by mutation category. Grey line represents the best fit with dotted lines indicating the 95% confidence interval. Pearson correlation coefficient (R^2) calculated of experimental to predicted PSIs.

B) Violin plots of correlation coefficients for each mutation category for structure model, SRE model, and combined model. R^2s calculated for each mutation category by training and testing on subsets of all mutations by non-parametric bootstrapping 10 times. Structure model uses unfolding free energy of pre-mRNA

within spliceosome at B^{act} stage as predictor. SRE strength model uses relative change in SRE strength as predictor. Combined model using both structure and SRE strength and weighs the features based on if mutation is intronic/synonymous or non-synonymous (Materials and methods).

C) Heatmap of the normalized changes in structure and SRE strength for each mutation clustered by affected features. Features were normalized such that a value of 1 implied that change in the feature should result in Exon 10 being spliced in (4R isoform, blue), whereas a value of 0 implies Exon 10 should be spliced out (3R isoform, red). Mutations were clustered using hierarchal clustering on normalized features (Materials and methods). Experimental PSIs are plotted for each mutation with a PSI of 1 colored as blue, PSI of 0.5 colored as white and PSI of 0 colored as red.

D) Pie chart showing the features that regulate Exon 10 splicing for the 47 experimentally validated mutations. The pie chart was generated based on the heatmap in C. Exon 10 splicing for 51.1% of mutations is supported by changes in both structure and SRE, which implies that structure, at least one SRE, and PSI are either all blue or all red. Exon 10 splicing for 23.4% of mutations is supported by changes in SRE wherein one of the SREs is the same color as PSI. For 17.0% of mutations, structural changes support Exon 10 splicing wherein structure and PSI are the same color. For 4 mutations (8.5%), the colors of none of the features match the color of PSI.

Mutations around the MAPT Exon 10-Intron 10 junction skew to Exon 10 inclusion

Having established that our quantitative models accurately predicted Exon 10 PSIs for experimentally validated mutations, we interrogated the model by performing a systematic mutagenic analysis spanning a 100-nucleotide window of the exon-intron junction (Figure 6A). Our model predicts that more mutations result in the inclusion of Exon 10 (4R isoform). This is consistent with the observation that a majority (75%) of known disease associated mutations (Figure 6B) are also 4R; this result is also consistent when categorized by all substitution types (Figure 6-figure supplement 1A). We found that a significantly greater proportion of disease mutations (76.4%) resulted in changes to both structure and SRE compared with non-disease mutations (36.0%) (Figure 6C) suggesting that mutations which affect both structure and SREs have a greater likelihood of causing disease compared with mutations that alter only one of the two factors. Intriguingly, mutations overall caused a slight skew towards a structured exon-intron junction, which would result in decreased inclusion of Exon 10 (Figure 6A, Figure 6-figure supplement 1B). However, changes to SRE strength skewed towards increased inclusion of Exon 10 (Figure 6-figure supplement 1C), which suggested that SREs were acting to counter the effect of structural changes. Our model reveals how a complex balance of structure and SRE RBP binding sites effectively results in the observed 50:50 ratio of the 3R and 4R isoforms.

To assess the general applicability of our model beyond our mutation training set, we predicted Exon 10 PSIs for 55 variants of unknown significance (VUSs) found in dbSNP (see Supplementary file 3 for further details of VUSs). These are mutations observed in the human population but are not currently associated with disease. The mean Exon 10 PSI for VUSs was 0.66, and 70% were within a standard deviation of the mean (Figure 6D). We observed that only a few mutations were predicted to have a PSI of zero (3R) (Figure 6D red bar). We therefore experimentally verified with splicing assays (Materials and methods) 6 VUSs: 3 VUSs predicted to be 3R, 1 VUS predicted to be 4R and 2 VUSs predicted to maintain the WT splicing ratio (Figure 6D). We found these 6 predictions were correct (Figure 6E). The three 3R VUSs made the region around the

exon-intron junction more structured while the 4R VUS made the region less structured compared to WT (Figure 6-figure supplement 1D) matching the direction of Exon 10 splicing change. Though we see changes to SRE strength match up to Exon 10 splicing direction for +30U>C and -6G>A, this was not the case for +25C>G and +23U>C (Figure 6-figure supplement 1E). For +23U>C and +26G>A, we observed changes in structured-ness around the exon-intron junction and silencer strengths in diverging directions (Figure 6-figure supplement 1D, E) suggesting that these opposing changes would preserve the WT 3R/4R ratio.

Figure 6

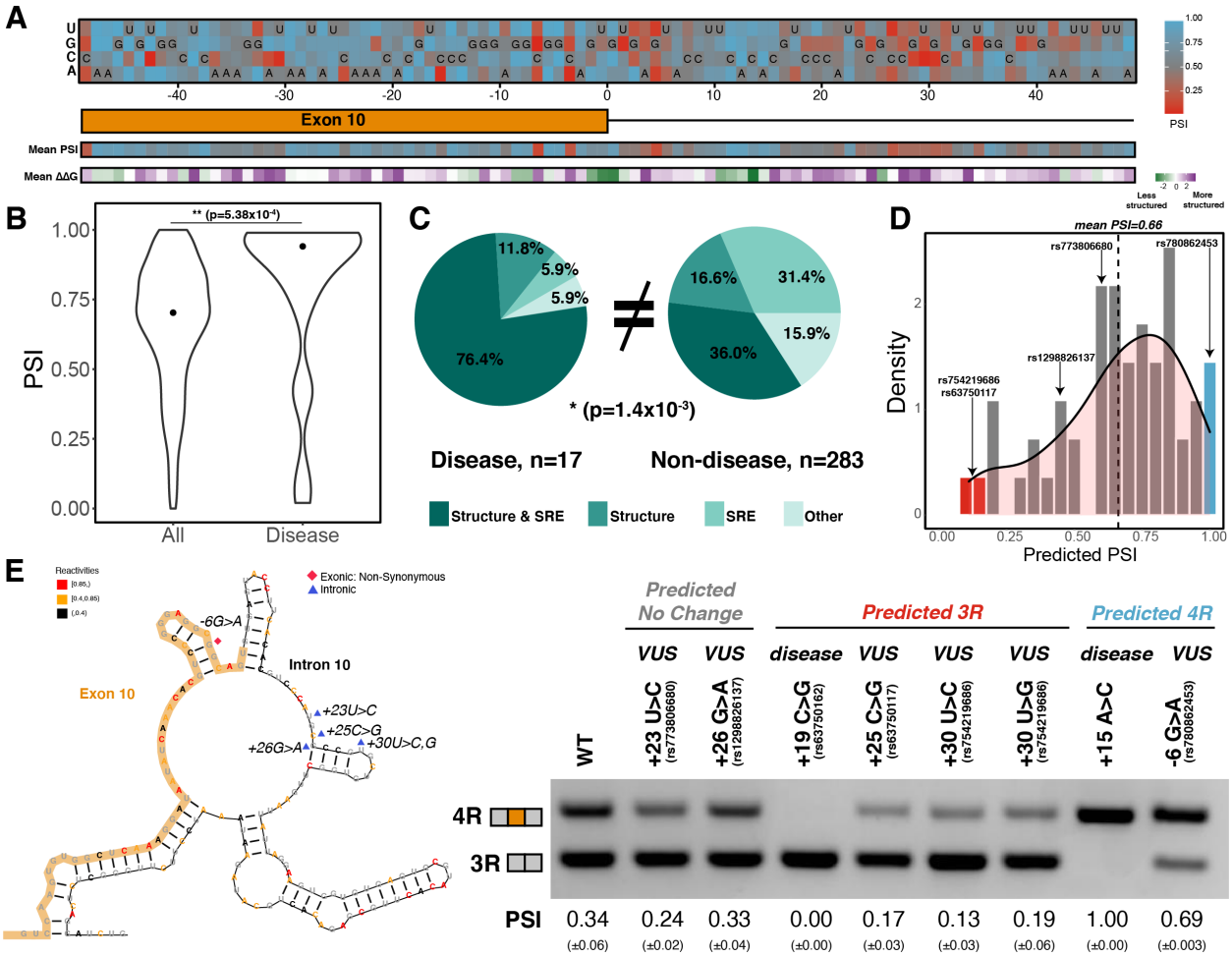


Figure 6: Mutations around Exon 10-Intron 10 junction skew towards inclusion of Exon 10

A) Heatmap of predicted Exon 10 PSIs for every possible mutation around 100 nucleotide window of Exon 10-Intron 10 junction. Combined model was trained using 47 mutations with experimental PSIs measured from splicing assays and used to predict PSIs for all mutation combinations for 100 nucleotides around the junction. Tiles with sequence indicate the wild type nucleotide at the position. Heatmap of mean PSI per position and mean relative change in unfolding free energy of pre-mRNA within spliceosome at B^{act} stage compared with wild type is shown below the gene diagram.

B) Violin plot of predicted PSIs for all possible mutations around Exon 10-Intron 10 junction and only disease mutations. All possible mutations (n=300), disease mutations (n=17)

C) Pie chart showing features that drive Exon 10 splicing for disease and non-disease mutations. The pie chart was generated by quantifying the number of mutations for which the direction of predicted Exon 10 PSI matched the direction of structure or SRE change. Exon 10 splicing for 76.4% of disease mutations is supported by changes to both structure and SRE compared with only 36.0% of non-disease mutations. The difference in proportions was tested with Fisher's exact test.

D) Histogram displaying the distribution of predicted PSIs using the combined model for 55 variants of unknown significance (VUSs) found in dbSNP. Density curve was overlaid on top of histogram showing that predicted PSIs skew away from 3R. Dotted line shows mean predicted PSI of 0.66. VUSs tested in splicing assays are indicated by their dbSNP RS IDs.

E) Representative gel of RT-PCR data for splicing assay in the presence of VUSs. Splicing reporter was transfected into HEK293 cells. The mean Exon 10 PSI displayed for each variant was calculated from three replicates and standard error is shown in brackets below. Structure diagram on left displays the location of the VUSs tested.

Discussion

In vivo DMS chemical probing of endogenous MAPT Exon 10-Intron 10 junction

Splicing specificity is complex (Baralle and Giudice 2017). The spliceosome does not rely on sequence alone to correctly identify 5' and 3' splice sites; other cues ensure correct binding to appropriate locations. In addition, the 5' splice site must be accessible to permit base pairing with the U1snRNA to initiate splicing (Roca et al. 2012). The *MAPT* Exon 10-Intron 10 junction is a well-studied example of the effect of 5' splice site secondary structure in splicing regulation. A hairpin was hypothesized initially because disease mutations close to the exon-intron junction (Hutton et al. 1998; Grover et al. 1999) shifted the isoform balance to either completely exclude or include Exon 10. Although NMR, in vitro chemical probing, and computation confirmed the presence of the hairpin (Varani et al. 1999; Chen et al. 2019; Lisowiec et al. 2015), recent studies showed that most RNAs were less structured in vivo and in the nucleus compared with in vitro conditions (Sun et al. 2019; Rouskin et al. 2014). However, our results revealed that this is not the case for the Exon10-Intron10 junction: in vivo chemical probing of the endogenous junction showcased strong evidence of structure.

In this study we observed that, in vivo, endogenous exons are less structured than introns, as found by Sun et al (Sun et al. 2019). Mature *MAPT* 3R and 4R exon-exon junctions are less structured compared with the pre-mRNA Exon 10-Intron 10 junction. The high correlation of structure we observed between the same exons found in different *MAPT* isoforms corroborates results observed with yeast ribosomal protein genes (Zubradt et al. 2016), which suggests that RNA folding in both pre- and post-spliced human exons is highly local and modular in exons.

Changes to structural ensemble around the 5' splice site are strong predictors of Exon 10 splicing

We showed that structural ensembles have an important function at the Exon 10-Intron 10 junction. If the 5' SS was always paired, only one isoform lacking Exon 10 would

result. However, the simultaneous presence of 3R and 4R isoforms implies that the junction is accessible in a subset of the structures. Unlike transfer RNAs and ribosomal RNAs that have single structures (Petrov et al. 2014), most RNAs are dynamic, unfolding and refolding within a landscape (Cruz and Westhof 2009; Giegé et al. 2012). We found disease mutations produced distinct shifts in the ensemble of the *MAPT* Exon 10-Intron 10 junction; the shifts corresponded to changes in the 3R:4R isoform ratio and confirmed that ensembles are essential at this junction. The activity of ensembles was corroborated by our quantitative model; including free energy features of the structural ensemble produced 1.5 times more accurate prediction of Exon 10 PSI compared with the unfolding free energy of the minimum free energy (MFE) structure.

Considering a larger spliceosome footprint on pre-mRNA produced more accurate prediction of Exon 10 PSI

The U1snRNA base pairs with the nine nucleotide sequence around the exon-intron junction (Roca et al. 2012). However, our analysis of the Cryo-EM structures of the human spliceosomal assembly cycle revealed that a larger region of the pre-mRNA interacts with the spliceosome during the splicing cycle and is therefore unfolded. Like the other main cellular ribonucleoprotein complex, the ribosome (Ingolia 2016), there is likely a spliceosomal footprint on the pre-mRNA and a minimum span around the splicing signals (5' splice site, 3' splice site and branch point) must be single-stranded for splicing to occur. Accordingly, our structural model performed most accurately when we used the unfolding free energy of 43 nucleotides around the 5' exon-intron junction that exists within the spliceosome B^{act} complex. This suggests that structures distal to the exon-intron junction regulate Exon 10 splicing, a finding that corroborates evidence that RNA structure near this exon-intron junction is more extended than previously determined (Tan et al. 2019). This result, combined with our use of Boltzmann suboptimal sampling demonstrates the key role of pre-cursor mRNA structure in splicing outcome.

RNA structure and SREs have complementary functions in MAPT Exon 10 regulation

Considerable evidence supports a function for either splicing regulatory elements and their corresponding RBPs or RNA structure in alternative splicing of *MAPT* Exon 10 at the 5' splice site (Andreadis 2012). However, there was no consensus as to which of the two factors is dominant. The regression model we developed established the relative importance of RNA structure vs. SREs at the exon-intron junction. We discovered a cooperative relationship between SREs and RNA structure whereby exonic non-synonymous mutations promoted splicing changes primarily by SRE motifs and exonic synonymous and intronic mutations by RNA structure around the exon-intron junction. A combined model that accounted for both structure and SREs was the most accurate predictor of Exon 10 PSI, and most experimentally validated mutations altered RNA structure and SRE motif strength around the Exon 10-Intron 10 junction in the same direction (Figure 5D). The model further suggested that the overall region favored increased Exon 10 inclusion (Figure 6 A,B), which confirmed previous experimental findings that inclusion is Exon 10's typical splicing mode (Q. S. Gao et al. 2000). This preference was proposed to be due to a weak 5' splice site (Ian D'Souza and Schellenberg 2005), and, indeed, we found that almost all experimentally validated mutations strengthened the splice site to increase inclusion of Exon 10 (Figure 4A). However, interestingly, our model revealed that structural changes caused by the mutations resulted in a more structured exon-intron junction, which would imply decreased Exon 10 inclusion. However, SRE strength alterations overall skewed more towards increased Exon 10 inclusion, which suggest that SREs and the RBPs that bind them buffer the effects of RNA structure to maintain the 1:1 isoform ratio at this junction. Our work revealed that structure and splicing regulatory elements most often have opposite effects on splicing outcomes. However, disease variants were the exception to this rule and resulted in a synergistic effect on splicing outcome (Fig. 6E), leading to a greater disruption of splicing, and therefore increased pathogenicity. The combined model was finally validated by accurate prediction of the effects of six previously untested VUSs on Exon 10 splicing (Figure 6E). As was the case with the complete mutagenesis, there were few VUSs predicted to completely alter the ratio of isoforms to entirely 3R: only 5 VUSs had PSIs less than 0.25. However, our model accurately

predicted the effect of the three 3R VUSs tested. Interestingly, the systematic computational mutagenesis revealed a hotspot of 3R mutations around 25-30 nucleotides downstream of the exon-intron junction (Figure 6A) and indeed the 3R VUSs experimentally validated were located in this region.

Quantitative modeling of splicing regulation at exon-intron junctions

Predictive models can measure the contribution of individual factors to an outcome. Structure around the 5' splice site and SRE motifs were excellent predictors of Exon 10 splicing in cells. The use of general SRE motifs enables this splicing framework to extend to other exon-intron junctions. By using a common dependent variable of Exon 10 PSI, we could use experimentally validated mutation data from disparate sources. Although our model provided an exact PSI prediction for each mutation, its principal utility was in predicting the direction in which the 3R:4R isoform ratio shifted from the wild type balance. On the basis of RNA-sequencing of brain tissue from healthy individuals, we find a range of Exon 10 PSIs between individuals and between tissues within an individual (Figure 1A). Even in individuals with progressive supranuclear palsy, a tauopathy in which *MAPT* variants are implicated, there is variability in Exon 10 PSIs between different brain tissues (Majounie et al. 2013). Ultimately, although it is likely that what is considered the correct ratio for normal brain function varies between tissues, our model provides a means to determine the baseline change of Exon 10 splicing simply based on sequence features. Many neurodegenerative diseases are caused by mutations around the *MAPT* Exon 10-Intron 10 junction, and there are no approved therapeutics that target this junction. Our work suggests that it is crucial to consider the larger structural context of the Exon 10-Intron 10 junction and the interplay between structure and SREs when considering the consequences of mutations on splicing regulation and the design of potential therapeutics to alter this ratio.

Materials and methods

MAPT Exon 10 PSIs for GTEx tissue types

Aligned BAM files of individual samples from the Genotype-Tissue Expression (GTEx) v8 project, for tissue types with *MAPT* TPM greater than 10, were accessed in the AnVIL/Terra environment (Kumar 2020a). Reads aligning to *MAPT* were extracted in Terra (Kumar 2020b) and downloaded. Exon 10 PSIs were quantified per BAM file with Outrigger (Song et al. 2017) using *MAPT* transcriptome reference from Ensembl GRCh38. Median values for each tissue type were calculated and then visualized on the brain diagram with R package, CerebroViz (Bahl, Koomar, and Michaelson 2017). Source file for Figure 1 provides Exon 10 PSI values for the 2,962 samples. An ANOVA test was run in R to test significance in variation between individuals versus within an individual (for individuals with MAPT expression in more than 7 tissues) (Supplementary file 1).

Culture of T47D and SH-SY5Y cells

Mammary gland carcinoma cells (T47D) were cultured in RPMI 1640 medium, supplemented with 10% Fetal Bovine Serum (FBS) and 0.2 units/mL of human insulin at 37°C and 5% CO_2. Bone marrow neuroblastoma SH-SY5Y cells were cultured in 1:1 mixture of 1X Minimum Essential Medium (MEM) and 1X F12 medium, supplemented with 10% FBS at 37 °C and 5% CO_2.

In vivo DMS-MaP probing for *MAPT* RNA

Approximately 20 million T47D cells and 30 million SHSY-5Y cells were harvested by centrifugation and resuspended in bicine buffered medium (300 mM Bicine pH 8.3, 150 mM NaCl, 5 mM $MgCl_2$) followed by treatment with DMS (1:10 ethanol diluted) for 5 min at 37°C. For the negative control (unmodified RNA), instead of DMS, an equivalent amount of ethanol was added to the same number of T47D and SH-SY5Y cells. After incubation, the reactions were neutralized by addition of 200 μl of 20% by volume β-mercaptoethanol. Total RNA was extracted by Trizol (ThermoFisher Scientific), treated

with TURBODNase (ThermoFisher Scientific), purified using Purelink RNA mini kit (ThermoFisher Scientific) and quantified with NanoDrop™ spectrophotometer.

DMS-MaP cDNA synthesis, library construction and sequencing for *MAPT* RNA

Purified RNA (9 μg) was reverse transcribed using Random Primer 9 (NEB) and SuperScript II reverse transcriptase under error prone conditions as described in Smola et al., 2015. The resultant cDNA was purified using G50 column (GE healthcare) and subjected to second strand synthesis (NEBNext Second Strand Synthesis Module). Supplementary file 4 lists PCR primers used for library generation. The cDNA was amplified by the NEB Q5 HotStart polymerase (NEB). Secondary PCR was performed to introduce TrueSeq barcodes (Smola et al. 2015). All samples were purified using the Ampure XP (Beckman Coulter) beads and quantification of the libraries was performed with Qubit dsDNA HS Assay kit (ThermoFisher Scientific). Final libraries were run on Agilent Bioanalyzer for quality check. TrueSeq libraries were then sequenced as necessary for their desired length as paired end 2×151 and 2×301 read multiplex runs on MiSeq platform (Illumina) for pre-cursor and mature MAPT isoforms respectively. Sequenced reads have been uploaded to the NCBI SRA database under BioSample accession numbers SAMN21372031 and SAMN21372032.

In vivo DMS-MaP probing for *SSU ribosome*

For in vivo ribosomal structure data, we used approximately 10 million T47D cells in 10 cm plates for each condition. We removed the growth media, added 1.8 mL of bicine buffered growth medium (200 mM Bicine pH 8.3) followed by treatment at 37°C with 200 uL of DMS diluted in ethanol (1.25% final DMS) for 5 min. For the negative control (unmodified RNA), instead of DMS, an equivalent amount of ethanol was added to the same number of T47D cells. After incubation, all reactions were neutralized by addition of ice cold 20% by volume β-mercaptoethanol and kept on ice for 5 minutes. Total RNA was extracted by Trizol (ThermoFisher Scientific), chloroform and isoamyl alcohol using phase lock heavy tubes (5PRIME Phase Lock Gel). RNA was purified

using Purelink RNA mini kit (ThermoFisher Scientific), treated with TURBODNase (ThermoFisher Scientific) and quantified with NanoDrop™ spectrophotometer.

DMS-MaP cDNA synthesis, library construction and sequencing for *SSU ribosome*

Purified RNA (5 ug) was reverse transcribed using Random Primer 9 (NEB) and SuperScript II reverse transcriptase under error prone conditions as described Smola et al., 2015. The resultant cDNA was purified using G50 column (GE healthcare) and subjected to second strand synthesis (NEBNext Second Strand Synthesis Module). Standard Nextera DNA library protocol (Illumina) was used to fragment the cDNA and add sequencing barcodes. All samples were purified using Ampure XP (Beckman Coulter) beads and quantification of the libraries was performed with Qubit dsDNA HS Assay kit (ThermoFisher Scientific). Final libraries were run on Agilent Bioanalyzer for quality check. Libraries were sequenced as paired end 2×151 read multiplex runs on MiSeq platform (Illumina). Sequenced reads have been uploaded to the NCBI SRA database under BioSample accession numbers SAMN21372031 and SAMN21372032.

DMS-MaP analysis

Sequenced reads were analyzed using the ShapeMapper pipeline(Busan and Weeks 2018), version (v2.1.4) which calculates the DMS reactivity of each nucleotide *i* as follows:

$$R_i = mutr_S - mutr_U$$

where $mutr_S$ is the mutation rate in the sample treated with DMS, $mutr_U$ is the mutation rate in the untreated control. DMS reactivities were normalized within a sample and per nucleotide type (A, C, U, G) using the normalization method described in Busan and Weeks, 2018. DMS probing data were collected for the Exon 9-Exon 11 and Exon 9-Exon 10-Exon 11 junctions using a single pair of primers listed in Supplementary file 4. The ShapeMapper pipeline ran for the two junctions in a single run with reference sequences for both junctions entered in one FASTA file. For the SSU, sequenced reads were first aligned to the SSU ribosome sequence using Bowtie2 parameters from Busan

and Weeks, 2018. Using samtools, alignments with MAPQ score greater than 10 were kept, sorted, and converted back into FASTQ files after which the ShapeMapper pipeline was executed.

Updating DMS parameters for RNAstructure using SSU ribosome data from T47D cells

To use DMS data to guide secondary structure prediction by the Rsample (Spasic et al. 2018) component of RNAstructure (Reuter and Mathews 2010), we calibrated the expected DMS reactivities per nucleotide. Using the SSU ribosome mapping data and the known secondary structure (Petrov et al. 2014), we determined histograms for DMS reactivities for unpaired nucleotides, nucleotides paired at helix ends, and nucleotides paired in base pairs stacked between two other base pairs. These DMS histograms can be invoked by Rsample with the "--DMS" command line switch as part of RNAstructure 6.4 or later. The histograms had long tails to relatively high reactivities. We empirically found that limiting reactivities in the histograms and in the input data to a reactivity of 5 (where higher values are set to 5) gave the best performance at improving SSU rRNA secondary structure. The "--max 5" parameter is used with Rsample to apply this limitation. Pre-release of Rsample code including in vivo DMS parameters is included as a zip file for review, and will be included in RNAstructure 6.4.

Base-pairing probabilities for SSU

The partition function for the SSU was generated using Rsample, using either the sequence or using the sequence and the DMS reactivities. All possible *i-j* base pairing probabilities were summed for each nucleotide *i* to generate a base pairing probability per nucleotide *i*.

ROC curves for predicting SSU base pairs

Using the known secondary structure of the SSU, we assigned a nucleotide as either 0 or 1 if it was paired or unpaired. DMS reactivities were used to predict whether a nucleotide was paired; the higher the DMS reactivity, the more likely a nucleotide is

unpaired. Base pairing probabilities were subtracted from 1 to obtain the probability that a base was unpaired, with 0 implying base was paired and 1 implying that base was unpaired. ROC curves and AUC values were generated using the plotROC (Sachs 2017) R package.

Arc plots

Arc plots were generated using Superfold (Siegfried et al. 2014) modified to process DMS reactivity data.

Generating structural ensemble of Exon 10-Intron 10 *MAPT* junction

The partition function of the Exon 10-Intron 10 MAPT junction for wild type (WT) and mutations was calculated with DMS reactivities as restraints using Rsample (Spasic et al. 2018). The DMS reactivities, which were collected for the WT sequence, were also used for the mutations to restrain the structural space with the reactivity made NA at the nucleotide where the mutation occurred. The program stochastic (Reuter and Mathews 2010) was used to sample 1000 structures (in CT format) from the Boltzmann distribution wherein the likelihood a structure is sampled was proportional to the probability that it occurred in the distribution (Y. Ding and Lawrence 2003).

t-SNE visualization of structural ensembles of WT, 3R and 4R mutations of Exon 10-Intron 10 *MAPT* junction

Structural ensembles were generated as described above for WT, 3R, and 4R mutations. For each sequence, the 1000 structures in CT format were converted to dot-bracket (db) format with ct2dot (Reuter and Mathews 2010), after which the db structure was transformed into the element format using rnaConvert in the Forgi package (Kerpedjiev, Höner Zu Siederdissen, and Hofacker 2015). In the element format, every base is represented by the subtype of RNA structure in which it is found: stem (s), hairpin (h), loop(m), 5'end(f), and 3'end(t). Hence, each db structure is a string of characters. These characters were digitized (f, t:0, s:1, h:2, m:3) to create a numerical matrix with 1000 rows and 234 columns, the length of the exon-intron junction.

Combining the matrices for the three sequences resulted in a 3000×234 matrix. This matrix was entered into the tSNE function from the scikit-learn python package (Pedregosa et al. 2011) and dimensionality was reduced to a 3000×2 matrix which was then plotted with ggplot2 (Wickham 2016) in R. The $\Delta G°$ of unfolding of the splice site was calculated for each of the 3000 structures as described below. Source file for Figure 3B lists t-SNE reduced data with corresponding free energies.

Determining representative structures for clusters in t-SNE plot

The 3000×2 matrix, the result of t-SNE dimensionality reduction, was clustered using k-means clustering with the k-means function from the scikit-learn python package (Pedregosa et al. 2011). The value of k was set to 5 as determined visually. A custom python script was used to deduce the representative structure for each cluster by first calculating the most common RNA structure subtype at each position. The actual structure in the ensemble, most similar to the RNA structure with the most common subtypes at each position, was then determined and deemed to be the representative structure of that cluster.

Visualizing density of structures in t-SNE plot

A custom python script was written. For the WT and, 3R, and 4R mutant sequences, a meshgrid was created for the three matrices using a 1000-point interpolation and NumPy (Harris et al. 2020) meshgrid function which returns two two-dimensional arrays that represent all the possible x-y coordinates for the three matrices. A gaussian kernel was next fit and evaluated for each 1000×2 matrix with SciPy gaussian_kde function (Virtanen et al. 2020) to smoothen over the meshgrid. Contour lines were generated for the smoothed data with Matplotlib contour function (Hunter 2007) and contourf was used to plot the data.

Quantifying nucleotides around the 5' splice site in cryo-EM structure

The Protein Databank (PDB) files for Pre-B (PDB ID: 6QX9), B (PDB ID: 5O9Z), Pre-B^{act} (PDB ID: 7ABF) and B^{act} (PDB ID: 5Z56) complexes were downloaded from the

PDB website. A custom python script was used to extract pre-mRNA from each PDB file. The number of nucleotides were counted for mRNA found near the 5' splice site. The result was visually confirmed by visualizing the PDB on PyMol.

Calculating ΔG° of unfolding of a region of interest

The ΔG° of a structure was calculated using the efn2 program in RNAstructure (Reuter and Mathews 2010). The base pairs within a region of interest were removed using a custom python script. The ΔG° of the "unfolded" structure was next re-calculated with efn2. The ΔG° of unfolding of a region was the subtraction of the ΔG° of the original structure from the ΔG° of the unfolded structure. For example, for determining the ΔG° of unfolding of the splice site, we removed all base pairs within the last 3 nucleotides of the exon and the first 6 nucleotides of the intron.

Calculating the change in strength of SRE motifs

Splice Site: Strength of the WT splice site was calculated with MaxEntScan (Yeo and Burge 2004). Strength was recalculated in the presence of splice site mutations either in the last 3 bases of Exon 10 or first 6 bases of Intron 10. WT strength was subtracted from the mutant strength: a 0 implied no change in splice site strength, positive values implied that a mutation made splice site stronger, resulting in increased inclusion of Exon 10, and negative values implied that a mutation made splice site weaker and decreased inclusion of Exon 10.

Enhancers and Silencers: Overrepresented hexamers in cell-based screens of general exonic and intronic splicing enhancers (ESEs, ISEs) and silencers (ESSs, ISSs) were obtained from Fairbrother et al., 2002, Wang et al., 2004, Wang, Ma et al., 2012 and Wang, Xiao et al., 2012. Position weight matrices (PWMs) of hexamers for each category were re-calculated as described in Fairbrother et al., 2002 and collated in Supplementary file 5. There were 8 clusters of ESE motifs, 7 of ESS motifs, 7 of ISE motifs, and 8 clusters of ISS motifs; each cluster had an associated PWM. For each PWM, a threshold strength was found by taking the 95th percentile value of strength of all possible k-mers of PWM length. This threshold was used to determine whether there

was a valid SRE motif at a particular position. The strength of the PWM motif was calculated across the exon-intron junction using a sliding window for both WT sequence and per mutation. The only windows that differed were around the location of the mutation, and only windows with strength above the threshold were considered. The WT strength was subtracted from the mutation strength for each window, and all windows were then summed to yield a Δstrength for every PWM per mutation. The average of the non-zero Δstrengths was calculated for ESE, ESS, ISE and ISS categories. The ESE and ISE Δstrengths were summed to obtain an enhancer strength, and the ESS and ISS Δstrengths were summed to obtain a silencer strength. Supplementary file 6 presents all SRE Δstrengths for the 47 mutations and 55 VUSs.

Calculating the change in strength of RBP motifs

Position Frequency Matrices (PFMs) were available from Ray et al. 2013 for the following RBPs: SRSF1, SRSF2, SRSF7, SRSF9, SRSF10, PCBP2, RBM4 and SFPQ. PFMs were converted into PWMs by normalizing frequencies to 0.25 (Prior probability for nucleotide frequency) and calculating the log2 value. Overrepresented hexamers were available from Dominguez et al., 2018 for the following RBPs: SRSF11, SRSF4, SRSF5 and SRSF8. PFMs for those RBPs were calculated as described in Fairbrother et al., 2002. Δstrength in RBP motifs were calculated the same way as SRE motifs. The average of non-zero values of RBPs implicated in either the inclusion or exclusion of Exon 10 was computed separately. All RBP Δstrengths for the 47 mutations are found in Supplementary file 6.

Models and bootstrapping

Exon 10 PSI was limited to values between 0 and 1 with 0 signifying that no transcripts had Exon 10 and 1 implying that all transcripts had Exon 10. Hence, standard linear regression was no longer appropriate and features were fit with a beta regression model to Exon 10 PSI. Regression parameters were determined using the betareg package (Cribari-Neto and Zeileis 2010) in R. Bootstrapping was performed by sampling without replacement 70% of the mutations to train and test the model and calculating the

Pearson correlation coefficient (R^2) between true values and predictions of the sample. This bootstrapping was executed 10 times resulting in a range of R^2s, ensuring that no subset of mutations skewed model performance. Since there were only 4 mutations that maintained the wildtype 3R to 4R ratio in our training set, we added 3 additional variants of unknown significance (VUSs) from the Single Nucleotide Polymorphism database (dbSNP) which we experimentally verified preserved the wildtype splicing pattern (see Supplementary file 7 for gel). WT VUSs tested and added to the training set were assigned a PSI of 0.5 to indicate equivalence to the WT sequence. Eq. **1**, the structure ensemble model, uses four characteristics describing ***X***, the $\Delta G°$ of unfolding of the region of interest around the exon-intron junction for 1000 structures in the ensemble. Eq. **2**, the minimum free energy (MFE) model, uses just ***Y***, the $\Delta G°$ of unfolding of the exon-intron junction found within the spliceosome at the B^{act} stage for the single MFE structure. Eq. **3**, the splice site model, uses the difference in splice site strength between WT sequence and a mutation where *SS* represents splice site. Eq. **4**, the combined SRE model, uses the difference in SRE strength between WT sequence and a mutation where *SS* represents splice site, *E* represents enhancer, and *S* represents silencer. Eq. **5**, the RBP model, uses the difference in RBP motif strength between WT sequence and a mutation where *Ex* represents RBPs involved in the exclusion of Exon 10 and *In* represents RBPs involved in the inclusion of Exon 10. Eq. **6** is the interactive model between structure and SRE, and Eq.**7** is the additive model. *isNonSynonymous*, *isSynonymous* and *isIntronic* represent the category of mutation and is either 0 or 1. Supplementary file 6 summarizes the performance of the models and features utilized.

$$PSI \sim Mean(X) + SD(X) + Skew(X) + Kurtosis(X) \qquad [\mathbf{1}]$$

$$PSI \sim Y \qquad [\mathbf{2}]$$

$$PSI \sim \Delta SS \qquad [\mathbf{3}]$$

$$PSI \sim \Delta E + \Delta S + \Delta SS \qquad [\mathbf{4}]$$

$$PSI \sim \Delta Ex + \Delta In \qquad [\mathbf{5}]$$

$$PSI \sim [Mean(X) + SD(X) + Skew(X) + Kurtosis(X)] * [isSynonymous + isIntronic] + [\Delta E + \Delta S + \Delta SS] * [isNonSynonymous] \qquad [\mathbf{6}]$$

$$PSI \sim [Mean(X) + SD(X) + Skew(X) + Kurtosis(X)] + [\Delta E + \Delta S + \Delta SS] \qquad [\mathbf{7}]$$

Clustering changes in structural and SRE features

For each feature, non-zero values greater than the 95th percentile value were set to the 95th percentile or, if less than the 5th percentile value, were set to the 5th percentile for visualization, after which all values were normalized to the maximum absolute value. Silencer Δstrength and mean ΔG° of unfolding of exon-intron junction of ensemble were inverted to follow the visualization such that values closer to 1 would result in greater Exon 10 inclusion and values closer to 0 would result in lower Exon 10 inclusion. Features were then assigned values 0 or 1 depending on whether the feature changed at all in the presence of the mutation. These digitized features were first clustered by hierarchal clustering resulting in 6 clusters. Each individual cluster was clustered again by hierarchal clustering but using the normalized feature values instead of 0s and 1s.

Splicing Assays

HEK-293 cells (ATCC CRL-1573) were grown at 37°C in 5% CO_2 in Dulbecco's Modified Eagle Medium (Gibco) supplemented with 10% FBS (Omega Scientific) and 0.5% Penicillin Streptomycin (Gibco). The wild type splicing reporter plasmid was generously provided by the Roca lab and is described in Tan et al., 2019. Single-nucleotide point mutations were generated using a Q5 site-directed mutagenesis kit (NEB) and confirmed by Sanger sequencing, or custom ordered directly from GenScript. Reporter plasmids (2 μg) were transfected into HEK-293 cells in 6-well plates at ~60-90% confluency using Lipofectamine 3000 (ThermoFisher Scientific). Cells were harvested after 1 day by aspirating the media and resuspending the cells in 1 mL Trizol

reagent (ThermoFisher Scientific). RNA was isolated using the PureLink RNA Isolation Kit (ThermoFisher Scientific) with on-column DNase treatment, following manufacturer's instructions. RNA (1 μg) was reverse transcribed to cDNA using Superscript VILO reverse transcriptase (ThermoFisher Scientific). Reverse transcriptions were performed by annealing (25°C 10 minutes), extension (50°C 10 minutes), and inactivation (85°C 10 min) steps. Heat-inactivated controls were prepared by heating the reaction without RNA at 85°C for 10 minutes prior to adding RNA, then following the described reaction conditions. The cDNA was PCR amplified with NEB Q5 HotStart polymerase (NEB) using splicing assay primers from IDT (AGACCCAAGCTGGCTAGCGTT forward, GAGGCTGATCAGCGGGTTTAAAC reverse) with 25 cycles. PCR product was purified and concentrated using the PureLink PCR micro clean up kit (ThermoFisher Scientific), following manufacturer's instructions. Splicing products were visualized by loading ~200 ng of DNA on a 2% agarose gel in 1X tris-acetate EDTA (TAE) buffer and staining with ethidium bromide. Gel images were quantified with ImageJ.

Supplementary files, figure source files, SNRNASMs and code are available at GitHub repository: https://git.io/JuSW8

Acknowledgements

This work was supported by the US National Institutes of Health R01 HL111527 and R35 GM 140844 to A.L. and R01 GM076485 to D.M. The authors wish to thank the Roca Lab for providing wildtype splicing reporter plasmids, Dr. Zefeng Wang for intronic splicing enhancer and silencer motifs, and Drs. Peter Castaldi, John Platig and Kevin Weeks for insightful discussions.

Competing Interests

The authors have declared that no competing interests exist.

References

Adivarahan, Srivathsan, Nathan Livingston, Beth Nicholson, Samir Rahman, Bin Wu, Olivia S. Rissland, and Daniel Zenklusen. 2018. "Spatial Organization of Single MRNPs at Different Stages of the Gene Expression Pathway." *Molecular Cell* 72 (4): 727-738.e5. https://doi.org/10.1016/J.MOLCEL.2018.10.010.

Andreadis, Athena. 2005. "Tau Gene Alternative Splicing: Expression Patterns, Regulation and Modulation of Function in Normal Brain and Neurodegenerative Diseases." *Biochimica et Biophysica Acta - Molecular Basis of Disease* 1739 (2): 91–103. https://doi.org/10.1016/j.bbadis.2004.08.010.

———. 2012. "Tau Splicing and the Intricacies of Dementia." *Journal of Cellular Physiology* 227 (3): 1220–25. https://doi.org/10.1002/jcp.22842.

Bahl, Ethan, Tanner Koomar, and Jacob J Michaelson. 2017. "CerebroViz: An R Package for Anatomical Visualization of Spatiotemporal Brain Data." *Bioinformatics (Oxford, England)* 33 (5): 762–63. https://doi.org/10.1093/bioinformatics/btw726.

Baralle, Francisco E, and Jimena Giudice. 2017. "Alternative Splicing as a Regulator of Development and Tissue Identity." *Nature Reviews Molecular Cell Biology* 18 (7): 437–51. https://doi.org/10.1038/nrm.2017.27.

Bertram, Karl, Dmitry E Agafonov, Olexandr Dybkov, Berthold Kastner, Reinhard Lü, and Holger Stark Correspondence. 2017. "Cryo-EM Structure of a Pre-Catalytic Human Spliceosome Primed for Activation." *Cell* 170: 701-706.e11. https://doi.org/10.1016/j.cell.2017.07.011.

Blanchette, Marco, and Benoit Chabot. 1997. "A Highly Stable Duplex Structure Sequesters the 5' Splice Site Region of HnRNP A1 Alternative Exon 7B." *RNA*, no. 3: 405–19.

Broderick, Jennifer, Junning Wang, and Athena Andreadis. 2004. "Heterogeneous Nuclear Ribonucleoprotein E2 Binds to Tau Exon 10 and Moderately Activates Its Splicing." *Gene* 331 (1–2): 107–14. https://doi.org/10.1016/j.gene.2004.02.005.

Buratti, Emanuele, and Francisco E Baralle. 2004. "Influence of RNA Secondary Structure on the Pre-MRNA Splicing Process." *MOLECULAR AND CELLULAR BIOLOGY* 24 (24): 10505–14. https://doi.org/10.1128/MCB.24.24.10505–10514.2004.

Busan, Steven, and Kevin M Weeks. 2018. "Accurate Detection of Chemical Modifications in RNA by Mutational Profiling (MaP) with ShapeMapper 2." *RNA* 24 (2): 143–48. https://doi.org/10.1261/rna.061945.117.

Catarina Silva, M., and Stephen J. Haggarty. 2020. "Tauopathies: Deciphering Disease Mechanisms to Develop Effective Therapies." *International Journal of Molecular Sciences* 21 (23): 1–49. https://doi.org/10.3390/ijms21238948.

Charenton, Clément, Max E. Wilkinson, and Kiyoshi Nagai. 2019. "Mechanism of 5' Splice Site Transfer for Human Spliceosome Activation." *Science* 364 (6438): 362–67.

https://doi.org/10.1126/science.aax3289.

Chen, Jonathan L., Walter N. Moss, Adam Spencer, Peiyuan Zhang, Jessica L. Childs-Disney, and Matthew D. Disney. 2019. “The RNA Encoding the Microtubule-Associated Protein Tau Has Extensive Structure That Affects Its Biology.” *PLOS ONE* 14 (7): e0219210. https://doi.org/10.1371/journal.pone.0219210.

Clark, Lorraine N, Parvoneh Poorkaj, Zbigniew Wszolek, Daniel H Geschwind, Ziad S Nasreddine, Bruce Miller, Diane Li, et al. 1998. “Pathogenic Implications of Mutations in the Tau Gene in Pallido-Ponto-Nigral Degeneration and Related Neurodegenerative Disorders Linked to Chromosome 17.” *Proceedings of the National Academy of Sciences of the United States of America* 95 (22): 13103–7. https://doi.org/10.1073/pnas.95.22.13103.

Cribari-Neto, Francisco, and Achim Zeileis. 2010. “Beta Regression in {R}.” *Journal of Statistical Software* 34 (2): 1–24. https://doi.org/10.18637/jss.v034.i02.

Cruz, José Almeida, and Eric Westhof. 2009. “The Dynamic Landscapes of RNA Architecture.” *Cell* 136 (4): 604–9. https://doi.org/10.1016/j.cell.2009.02.003.

D’Souza, Ian, P Poorkaj, M Hong, D Nochlin, V. M.- Y. Lee, T D Bird, and G D Schellenberg. 1999. “Missense and Silent Tau Gene Mutations Cause Frontotemporal Dementia with Parkinsonism-Chromosome 17 Type, by Affecting Multiple Alternative RNA Splicing Regulatory Elements.” *Proceedings of the National Academy of Sciences* 96 (10): 5598–5603. https://doi.org/10.1073/pnas.96.10.5598.

D’Souza, Ian, and Gerard D. Schellenberg. 2005. “Regulation of Tau Isoform Expression and Dementia.” *Biochimica et Biophysica Acta - Molecular Basis of Disease*, January 3, 2005. https://doi.org/10.1016/j.bbadis.2004.08.009.

———. 2006. “Arginine/Serine-Rich Protein Interaction Domain-Dependent Modulation of a Tau Exon 10 Splicing Enhancer: Altered Interactions and Mechanisms for Functionally Antagonistic FTDP-17 Mutations Δ280K and N279K.” *Journal of Biological Chemistry* 281 (5): 2460–69. https://doi.org/10.1074/jbc.M505809200.

D’Souza, Ian, and Gerard David Schellenberg. 2000. “Determinants of 4-Repeat Tau Expression. Coordination between Enhancing and Inhibitory Splicing Sequences for Exon 10 Inclusion.” *Journal of Biological Chemistry* 275 (23): 17700–709. https://doi.org/10.1074/jbc.M909470199.

Dethoff, Elizabeth A., Jeetender Chugh, Anthony M. Mustoe, and Hashim M. Al-Hashimi. 2012. “Functional Complexity and Regulation through RNA Dynamics.” *Nature*. Nature Publishing Group. https://doi.org/10.1038/nature10885.

Ding, Shaohong, Jianhua Shi, Wei Qian, Khalid Iqbal, Inge Grundke-Iqbal, Cheng Xin Gong, and Fei Liu. 2012. “Regulation of Alternative Splicing of Tau Exon 10 by 9G8 and Dyrk1A.” *Neurobiology of Aging* 33 (7): 1389–99. https://doi.org/10.1016/j.neurobiolaging.2010.11.021.

Ding, Ye, and Charles E Lawrence. 2003. “A Statistical Sampling Algorithm for RNA Secondary Structure

Prediction." *Nucleic Acids Research* 31 (24): 7280–7301. https://doi.org/10.1093/nar/gkg938.

Dominguez, Daniel, Peter Freese, Maria S. Alexis, Amanda Su, Myles Hochman, Tsultrim Palden, Cassandra Bazile, et al. 2018. "Sequence, Structure, and Context Preferences of Human RNA Binding Proteins." *Molecular Cell* 70(5): 854–867.e9. https://doi.org/10.1016/j.molcel.2018.05.001.

Donahue, Christine P, Christina Muratore, Jane Y Wu, Kenneth S Kosik, and Michael S Wolfe. 2006. "Stabilization of the Tau Exon 10 Stem Loop Alters Pre-MRNA Splicing." *Journal of Biological Chemistry* 281 (33): 23302–6. https://doi.org/10.1074/jbc.C600143200.

Fairbrother, William G., Ru-Fang Yeh, Phillip A. Sharp, and Christopher B. Burge. 2002. "Predictive Identification of Exonic Splicing Enhancers in Human Genes." *Science* 297: 1007–13. http://dx.doi.org/10.1126/science.1073774.

Ferrari, Silvia L.P., and Francisco Cribari-Neto. 2004. "Beta Regression for Modelling Rates and Proportions." *Journal of Applied Statistics* 31 (7): 799–815. https://doi.org/10.1080/0266476042000214501.

Gao, Lei, Junning Wang, Yingzi Wang, and Athena Andreadis. 2007. "SR Protein 9G8 Modulates Splicing of Tau Exon 10 via Its Proximal Downstream Intron, a Clustering Region for Frontotemporal Dementia Mutations." *Molecular and Cellular Neuroscience* 34 (1): 48–58. https://doi.org/10.1016/j.mcn.2006.10.004.

Gao, Qing Sheng, John Memmott, Robert Lafyatis, Stefan Stamm, Gavin Screaton, and Athena Andreadis. 2000. "Complex Regulation of Tau Exon 10, Whose Missplicing Causes Frontotemporal Dementia." *Journal of Neurochemistry* 74 (2): 490–500. https://doi.org/10.1046/j.1471-4159.2000.740490.x.

Giegé, Richard, Frank Jühling, Joern Pütz, Peter Stadler, Claude Sauter, and Catherine Florentz. 2012. "Structure of Transfer RNAs: Similarity and Variability." *Wiley Interdisciplinary Reviews: RNA* 3 (1): 37–61. https://doi.org/10.1002/WRNA.103.

Goedert, M., M. G. Spillantini, R. A. Crowther, S. G. Chen, P. Parchi, M. Tabaton, D. J. Lanska, et al. 1999. "Tau Gene Mutation in Familial Progressive Subcortical Gliosis." *Nature Medicine* 5 (4): 454–57. https://doi.org/10.1038/7454.

Goedert, M., M G. Spillantini, R. Jakes, D Rutherford, and R A Crowther. 1989. "Multiple Isoforms of Human Microtubule-Associated Protein Tau: Sequences and Localization in Neurofibrillary Tangles of Alzheimer's Disease." *Neuron* 3 (4): 519–26. https://doi.org/10.1016/0896-6273(89)90210-9.

Grover, Andrew, Henry Houlden, Matt Baker, Jennifer Adamson, Jada Lewis, Guy Prihar, Stuart Pickering-Brown, Karen Duff, and Mike Hutton. 1999. "5' Splice Site Mutations in Tau Associated with the Inherited Dementia FTDP-17 Affect a Stem-Loop Structure That Regulates Alternative Splicing of Exon 10*." *Journal of Biological Chemistry* 274 (21): 15134–43. http://www.jbc.org/content/274/21/15134.full.pdf.

Halvorsen, Matthew, Joshua S Martin, Sam Broadaway, and Alain Laederach. 2010. "Disease-Associated

Mutations That Alter the RNA Structural Ensemble." *PLoS Genetics* 6 (8). https://doi.org/10.1371/journal.pgen.1001074.

Harris, Charles R, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. "Array Programming with {NumPy}." *Nature* 585 (7825): 357–62. https://doi.org/10.1038/s41586-020-2649-2.

Hasegawa, Masato, Michael J Smith, Masaaki Iijima, Takeshi Tabira, and Michel Goedert. 1999. "FTDP-17 Mutations N279K and S305N in Tau Produce Increased Splicing of Exon 10." *FEBS Letters* 443 (2): 93–96. https://doi.org/10.1016/S0014-5793(98)01696-2.

Hefti, Marco M, Kurt Farrell, Soong Ho Kim, Kathryn R Bowles, Mary E Fowkes, Towfique Raj, and John F Crary. 2018. "High-Resolution Temporal and Regional Mapping of MAPT Expression and Splicing in Human Brain Development." *PLoS ONE* 13 (4). https://doi.org/10.1371/journal.pone.0195771.

Herzel, Lydia, Diana S. M. Ottoz, Tara Alpert, and Karla M. Neugebauer. 2017. "Splicing and Transcription Touch Base: Co-Transcriptional Spliceosome Assembly and Function." *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm.2017.63.

Homan, Philip J, Oleg V Favorov, Christopher A Lavender, Olcay Kursun, Xiyuan Ge, Steven Busan, Nikolay V Dokholyan, and Kevin M Weeks. 2014. "Single-Molecule Correlated Chemical Probing of RNA." *Proceedings of the National Academy of Sciences* 111 (38): 13858–63. https://doi.org/10.1073/pnas.1407306111.

Hunter, J D. 2007. "Matplotlib: A 2D Graphics Environment." *Computing in Science \& Engineering* 9 (3): 90–95. https://doi.org/10.1109/MCSE.2007.55.

Hutton, M., C. L. Lendon, P. Rizzu, M. Baker, Susanne Froelich, H. H. Houlden, S. Pickering-Brown, et al. 1998. "Association of Missense and 5'-Splice-Site Mutations in Tau with the Inherited Dementia FTDP-17." *Nature* 393 (6686): 702–4. https://doi.org/10.1038/31508.

Ingolia, Nicholas T. 2016. "Ribosome Footprint Profiling of Translation throughout the Genome." *Cell* 165 (1): 22–33. https://doi.org/10.1016/j.cell.2016.02.066.

Iseki, Eizo, Takehiko Matsumura, Wami Marui, Hiroaki Hino, Toshinari Odawara, Naoya Sugiyama, Kyoko Suzuki, Hajime Sawada, Tetsuaki Arai, and Kenji Kosaka. 2001. "Familial Frontotemporal Dementia and Parkinsonism with a Novel N296H Mutation in Exon 10 of the Tau Gene and a Widespread Tau Accumulation in the Glial Cells." *Acta Neuropathologica* 102 (3): 285–92. http://dx.doi.org/10.1007/s004010000333.

Jiang, Zhihong, Jocelyn Cote, Jennifer M Kwon, Alison M Goate, and Jane Y Wu. 2000. "Aberrant Splicing of Tau Pre-MRNA Caused by Intronic Mutations Associated with the Inherited Dementia Frontotemporal Dementia with Parkinsonism Linked to Chromosome 17." *Molecular and Cellular Biology* 20 (11): 4036–48. https://doi.org/10.1128/MCB.20.14.5360-5360.2000.

Kar, Amar, Kazuo Fushimi, Xiaohong Zhou, Payal Ray, Chen Shi, X. Chen, Zhiren Liu, She Chen, and Jane Y Wu. 2011. "RNA Helicase P68 (DDX5) Regulates Tau Exon 10 Splicing by Modulating a

Stem-Loop Structure at the 5' Splice Site." *Molecular and Cellular Biology* 31 (9): 1812–21. https://doi.org/10.1128/MCB.01149-10.

Kar, Amar, Necat Havlioglu, Woan Yuh Tarn, and Jane Y Wu. 2006. "RBM4 Interacts with an Intronic Element and Stimulates Tau Exon 10 Inclusion." *Journal of Biological Chemistry* 281 (34): 24479–88. https://doi.org/10.1074/jbc.M603971200.

Kerpedjiev, Peter, Christian Höner Zu Siederdissen, and Ivo L Hofacker. 2015. "Predicting RNA 3D Structure Using a Coarse-Grain Helix-Centered Model." *RNA* 21 (6): 1110–21. https://doi.org/10.1261/rna.047522.114.

Kondo, Shinichi, Noriaki Yamamoto, Tomohiko Murakami, Masayo Okumura, Akila Mayeda, and Kazunori Imaizumi. 2004. "Tra2β, SF2/ASF and SRp30c Modulate the Function of an Exonic Splicing Enhancer in Exon 10 of Tau Pre-MRNA." *Genes to Cells* 9 (2): 121–30. https://doi.org/10.1111/j.1356-9597.2004.00709.x.

Kumar, Jayashree. 2020a. "AnVIL_GTEx_V8_hg38_JKversion." Terra. 2020. https://app.terra.bio/#workspaces/fccredits-iridium-pear-1617/AnVIL_GTEx_V8_hg38_JKversion.

———. 2020b. "ExtractReadsFromBamFileForAGivenRegion_SortAndIndex." Terra. 2020. https://app.terra.bio/#workspaces/jk-billing-terra87/AnVIL_GTEx_V8_hg38_JKversion_NewWS/workflows/JayashreeKumar_UNCChapelHill/https://app.terra.bio/#workspaces/fccredits-iridium-pear-1617/AnVIL_GTEx_V8_hg38_JKversion/workflows/JayashreeKumar_UNCChapelHil.

Lai, Wan Jung C., Mohammad Kayedkhordeh, Erica V. Cornell, Elie Farah, Stanislav Bellaousov, Robert Rietmeijer, Enea Salsi, David H. Mathews, and Dmitri N. Ermolenko. 2018. "MRNAs and LncRNAs Intrinsically Form Secondary Structures with Short End-to-End Distances." *Nature Communications* 9 (1). https://doi.org/10.1038/s41467-018-06792-z.

Lin, Chien Ling, Allison J Taggart, and William G Fairbrother. 2016. "RNA Structure in Splicing: An Evolutionary Perspective." *RNA Biology* 13 (9): 766–71. https://doi.org/10.1080/15476286.2016.1208893.

Lisowiec, Jolanta, Dorota Magner, Elzbieta Kierzek, Elzbieta Lenartowicz, and Ryszard Kierzek. 2015. "Structural Determinants for Alternative Splicing Regulation of the MAPT Pre-MRNA." *RNA Biology* 12 (3): 330–42. https://doi.org/10.1080/15476286.2015.1017214.

Lonsdale, John, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, et al. 2013. "The Genotype-Tissue Expression (GTEx) Project." *Nature Genetics* 45 (6): 580–85. https://doi.org/10.1038/ng.2653.

Maaten, Laurens Van Der, and Geoffrey Hinton. 2008. "Visualizing Data Using T-SNE." *Journal of Machine Learning Research* 9: 2579–2605. https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf.

Majounie, Elisa, William Cross, Victoria Newsway, Allissa Dillman, Jana Vandrovcova, Christopher M.

Morris, Michael A. Nalls, et al. 2013. “Variation in Tau Isoform Expression in Different Brain Regions and Disease States.” *Neurobiology of Aging* 34 (7): 1922.e7-1922.e12. https://doi.org/10.1016/j.neurobiolaging.2013.01.017.

Matera, A. Gregory, and Zefeng Wang. 2014. “A Day in the Life of the Spliceosome.” *Nature Reviews Molecular Cell Biology* 15 (2): 108–21. https://doi.org/10.1038/nrm3742.

McManus, C. Joel, and Brenton R. Graveley. 2011. “RNA Structure and the Mechanisms of Alternative Splicing.” *Current Opinion in Genetics and Development*. https://doi.org/10.1016/j.gde.2011.04.001.

Mirra, Suzanne S ;, Jill R ; Murrell, Marla ; Gearing, and Maria G Spillantini. 1999. “Tau Pathology in a Family with Dementia and a P301L Mutation in Tau.” *Journal of Neuropathology and Experimental Neurology* 58 (4): 335. https://search.proquest.com/docview/229718949/fulltextPDF/AA5FABCB40F44E3FPQ/1?accountid=14244.

Mustoe, Anthony M., Steven Busan, Greggory M. Rice, Christine E. Hajdin, Brant K. Peterson, Vera M. Ruda, Neil Kubica, Razvan Nutiu, Jeremy L. Baryza, and Kevin M. Weeks. 2018. “Pervasive Regulatory Functions of MRNA Structure Revealed by High-Resolution SHAPE Probing.” *Cell* 173 (1): 181-195.e18. https://doi.org/10.1016/j.cell.2018.02.034.

Park, Sun Ah, Sang Il Ahn, and Jean Marc Gallo. 2016. “Tau Mis-Splicing in the Pathogenesis of Neurodegenerative Disorders.” *BMB Reports* 49 (8): 405–13. https://doi.org/10.5483/BMBRep.2016.49.8.084.

Pedregosa, F, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, et al. 2011. “Scikit-Learn: Machine Learning in Python.” *Journal of Machine Learning Research* 12: 2825–30.

Petrov, Anton S., Chad R. Bernier, Burak Gulen, Chris C. Waterbury, Eli Hershkovits, Chiaolong Hsiao, Stephen C. Harvey, et al. 2014. “Secondary Structures of RRNAs from All Three Domains of Life.” *PLoS ONE* 9 (2): e88222. https://doi.org/10.1371/journal.pone.0088222.

Qian, Wei, Khalid Iqbal, Inge Grundke-Iqbal, Cheng Xin Gong, and Fei Liu. 2011. “Splicing Factor SC35 Promotes Tau Expression through Stabilization of Its MRNA.” *FEBS Letters* 585 (6): 875–80. https://doi.org/10.1016/j.febslet.2011.02.017.

Qian, Wei, and Fei Liu. 2014. “Regulation of Alternative Splicing of Tau Exon 10.” *Neuroscience Bulletin* 30 (2): 367–77. https://doi.org/10.1007/s12264-013-1411-2.

Ray, Debashish, Hilal Kazan, Kate B Cook, Matthew T Weirauch, Hamed S Najafabadi, Xiao Li, Serge Gueroussov, et al. 2013. “A Compendium of RNA-Binding Motifs for Decoding Gene Regulation.” *Nature* 499 (7457): 172–77. https://doi.org/10.1038/nature12311.

Ray, Payal, Amar Kar, Kazuo Fushimi, Necat Havlioglu, Xiaoping Chen, and Jane Y. Wu. 2011. “PSF Suppresses Tau Exon 10 Inclusion by Interacting with a Stem-Loop Structure Downstream of Exon 10.” *Journal of Molecular Neuroscience* 45 (3): 453–66. https://doi.org/10.1007/s12031-011-9634-z.

Reuter, Jessica S, and David H Mathews. 2010. “RNAstructure: Software for RNA Secondary Structure

Prediction and Analysis." *BMC Bioinformatics 2010 11:1* 11 (1): 1–9. https://doi.org/10.1186/1471-2105-11-129.

Rizzu, Patrizia, John C. Van Swieten, Marijke Joosse, Masato Hasegawa, Martijn Stevens, Aad Tibben, Martinus F. Niermeijer, et al. 1999. "High Prevalence of Mutations in the Microtubule-Associated Protein Tau in a Population Study of Frontotemporal Dementia in the Netherlands." *American Journal of Human Genetics* 64 (2): 414–21. https://doi.org/10.1086/302256.

Roca, Xavier, Martin Akerman, Hans Gaus, Andrés Berdeja, C. Frank Bennett, and Adrian R. Krainer. 2012. "Widespread Recognition of 5′ Splice Sites by Noncanonical Base-Pairing to U1 SnRNA Involving Bulged Nucleotides." *Genes and Development* 26 (10): 1098–1109. https://doi.org/10.1101/gad.190173.112.

Rouskin, Silvi, Meghan Zubradt, Stefan Washietl, Manolis Kellis, and Jonathan S Weissman. 2014. "Genome-Wide Probing of RNA Structure Reveals Active Unfolding of MRNA Structures in Vivo. TL - 505." *Nature* 505 VN- (7485): 701–5. https://doi.org/10.1038/nature12894.

Sachs, Michael C. 2017. "{plotROC}: A Tool for Plotting ROC Curves." *Journal of Statistical Software, Code Snippets* 79 (2): 1–19. https://doi.org/10.18637/jss.v079.c02.

Sharma, Yogita, Milad Miladi, Sandeep Dukare, Karine Boulay, Maiwen Caudron-Herger, Matthias Groß, Rolf Backofen, and Sven Diederichs. 2019. "A Pan-Cancer Analysis of Synonymous Mutations." *Nature Communications* 10 (1): 1–14. https://doi.org/10.1038/s41467-019-10489-2.

Siegfried, Nathan A, Steven Busan, Greggory M Rice, Julie A E Nelson, and Kevin M Weeks. 2014. "RNA Motif Discovery by SHAPE and Mutational Profiling (SHAPE-MaP)." *Nature Methods* 11 (9): 959–65. https://doi.org/10.1038/nmeth.3029.

Singh, Natalia N, Ravindra N Singh, and Elliot J Androphy. 2007. "Modulating Role of RNA Structure in Alternative Splicing of a Critical Exon in the Spinal Muscular Atrophy Genes." *Nucleic Acids Research* 35 (2): 371–89. https://doi.org/10.1093/nar/gkl1050.

Smola, Matthew J, Greggory M Rice, Steven Busan, Nathan A Siegfried, and Kevin M Weeks. 2015. "Selective 2′-Hydroxyl Acylation Analyzed by Primer Extension and Mutational Profiling (SHAPE-MaP) for Direct, Versatile and Accurate RNA Structure Analysis." *Nature Protocols* 10 (11): 1643–69. https://doi.org/10.1038/nprot.2015.103.

Song, Yan, Olga B Botvinnik, Michael T Lovci, Boyko Kakaradov, Patrick Liu, Jia L Xu, and Gene W Yeo. 2017. "Single-Cell Alternative Splicing Analysis with Expedition Reveals Splicing Dynamics during Neuron Differentiation." *Molecular Cell* 67 (1): 148-161.e5. https://doi.org/10.1016/j.molcel.2017.06.003.

Spasic, Aleksandar, Sarah M Assmann, Philip C Bevilacqua, and David H Mathews. 2018. "Modeling RNA Secondary Structure Folding Ensembles Using SHAPE Mapping Data." *Nucleic Acids Research* 46 (1): 314–23. https://doi.org/10.1093/nar/gkx1057.

Spillantini, M. G., Jill R Murrell, Michel Goedert, Martin R Farlow, Aaron Klug, and Bernardino Ghetti.

1998. “Mutation in the Tau Gene in Familial Multiple System Tauopathy with Presenile Dementia.” *Proceedings of the National Academy of Sciences* 95 (13): 7737–41. https://doi.org/10.1073/pnas.95.13.7737.

Stenson, Peter D, Edward V Ball, Matthew Mort, Andrew D Phillips, Jacqueline A Shiel, Nick S T Thomas, Shaun Abeysinghe, Michael Krawczak, and David N Cooper. 2003. “Human Gene Mutation Database (HGMD): 2003 Update.” *Human Mutation* 21 (6): 577–81. https://doi.org/10.1002/humu.10212.

Sun, Lei, Furqan M. Fazal, Pan Li, James P. Broughton, Byron Lee, Lei Tang, Wenze Huang, Eric T. Kool, Howard Y. Chang, and Qiangfeng Cliff Zhang. 2019. “RNA Structure Maps across Mammalian Cellular Compartments.” *Nature Structural and Molecular Biology* 26 (4): 322–30. https://doi.org/10.1038/s41594-019-0200-7.

Tan, Jiazi, Lixia Yang, Alan Ann Lerk Ong, Jiahao Shi, Zhensheng Zhong, Mun Leng Lye, Shiyi Liu, et al. 2019. “A Disease-Causing Intronic Point Mutation C19G Alters Tau Exon 10 Splicing via RNA Secondary Structure Rearrangement.” *Biochemistry*. https://doi.org/10.1021/acs.biochem.9b00001.

Taylor, Katarzyna, and Krzysztof Sobczak. 2020. “Intrinsic Regulatory Role of RNA Structural Arrangement in Alternative Splicing Control.” *International Journal of Molecular Sciences* 21 (14): 1–35. https://doi.org/10.3390/ijms21145161.

Tazi, Jamal, Nadia Bakkour, and Stefan Stamm. 2009. “Alternative Splicing and Disease.” *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1792 (1): 14–26. https://doi.org/10.1016/J.BBADIS.2008.09.017.

Townsend, Cole, Majety N. Leelaram, Dmitry E. Agafonov, Olexandr Dybkov, Cindy L. Will, Karl Bertram, Henning Urlaub, Berthold Kastner, Holger Stark, and Reinhard Lührmann. 2020. “Mechanism of Protein-Guided Folding of the Active Site U2/U6 RNA during Spliceosome Activation.” *Science* 370 (6523). https://doi.org/10.1126/science.abc3753.

Varani, Luca, Masato Hasegawa, Maria Grazia Spillantini, Michael J Smith, Jill R Murrell, Bernardino Ghetti, Aaron Klug, Michel Goedert, and Gabriele Varani. 1999. “Structure of Tau Exon 10 Splicing Regulatory Element RNA and Destabilization by Mutations of Frontotemporal Dementia and Parkinsonism Linked to Chromosome 17 (Alternative MRNA Splicing Intronic Mutations Stem-Loop RNA Structure).” *Neurobiology* 96: 8229–34. https://doi.org/10.1073/pnas.96.14.8229.

Virtanen, Pauli, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, et al. 2020. “{SciPy} 1.0: Fundamental Algorithms for Scientific Computing in Python.” *Nature Methods* 17: 261–72. https://doi.org/10.1038/s41592-019-0686-2.

Wang, Junning, Qing Sheng Gao, Yingzi Wang, Robert Lafyatis, Stefan Stamm, and Athena Andreadis. 2004. “Tau Exon 10, Whose Missplicing Causes Frontotemporal Dementia, Is Regulated by an Intricate Interplay of Cis Elements and Trans Factors.” *Journal of Neurochemistry* 88 (5): 1078–90. https://doi.org/10.1046/j.1471-4159.2003.02232.x.

Wang, Yan, Lei Gao, Sze Wah Tse, and Athena Andreadis. 2010. “Heterogeneous Nuclear Ribonucleoprotein E3 Modestly Activates Splicing of Tau Exon 10 via Its Proximal Downstream Intron, a Hotspot for Frontotemporal Dementia Mutations.” *Gene* 451 (1–2): 23–31. https://doi.org/10.1016/j.gene.2009.11.006.

Wang, Yang, Meng Ma, Xinshu Xiao, and Zefeng Wang. 2012. “Intronic Splicing Enhancers, Cognate Splicing Factors and Context-Dependent Regulation Rules.” *Nature Structural & Molecular Biology* 19 (10): 1044–52. https://doi.org/10.1038/nsmb.2377.

Wang, Yang, Xinshu Xiao, Jianming Zhang, Rajarshi Choudhury, Alex Robertson, Kai Li, Meng Ma, Christopher B Burge, and Zefeng Wang. 2012. “A Complex Network of Factors with Overlapping Affinities Represses Splicing through Intronic Elements.” *Nature Structural & Molecular Biology* 20 (1): 36–45. https://doi.org/10.1038/nsmb.2459.

Wang, Zefeng, and Christopher B Burge. 2008. “Splicing Regulation: From a Parts List of Regulatory Elements to an Integrated Splicing Code.” *RNA* 14: 802–13. https://doi.org/10.1261/rna.876308.

Wang, Zefeng, Michael E. Rolish, Gene Yeo, Vivian Tung, Matthew Mawson, and Christopher B. Burge. 2004. “Systematic Identification and Analysis of Exonic Splicing Silencers.” *Cell* 119 (6): 831–45. https://doi.org/10.1016/j.cell.2004.11.010.

Warf, M Bryan, and J Andrew Berglund. 2010. “Role of RNA Structure in Regulating Pre-MRNA Splicing.” *Trends in Biochemical Sciences*. https://doi.org/10.1016/j.tibs.2009.10.004.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Yeo, Gene, and Christopher B Burge. 2004. “Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals.” *Journal of Computational Biology* 11 (2–3): 377–94. https://doi.org/10.1089/1066527041410418.

Zhang, Lingdi, Anne Vielle, Sara Espinosa, and Rui Zhao. 2019. “RNAs in the Spliceosome: Insight from CryoEM Structures.” *Wiley Interdisciplinary Reviews: RNA* 10 (3): 1–11. https://doi.org/10.1002/wrna.1523.

Zhang, Xiaofeng, Chuangye Yan, Xiechao Zhan, Lijia Li, Jianlin Lei, and Yigong Shi. 2018. “Structure of the Human Activated Spliceosome in Three Conformational States.” *Cell Research* 28 (3): 307–22. https://doi.org/10.1038/cr.2018.14.

Zubradt, Meghan, Paromita Gupta, Sitara Persad, Alan M Lambowitz, Jonathan S Weissman, and Silvi Rouskin. 2016. “DMS-MaPseq for Genome-Wide or Targeted RNA Structure Probing in Vivo.” *Nature Methods* 14 (1): 75–82. https://doi.org/10.1038/nmeth.4057.

Supplementary Figures

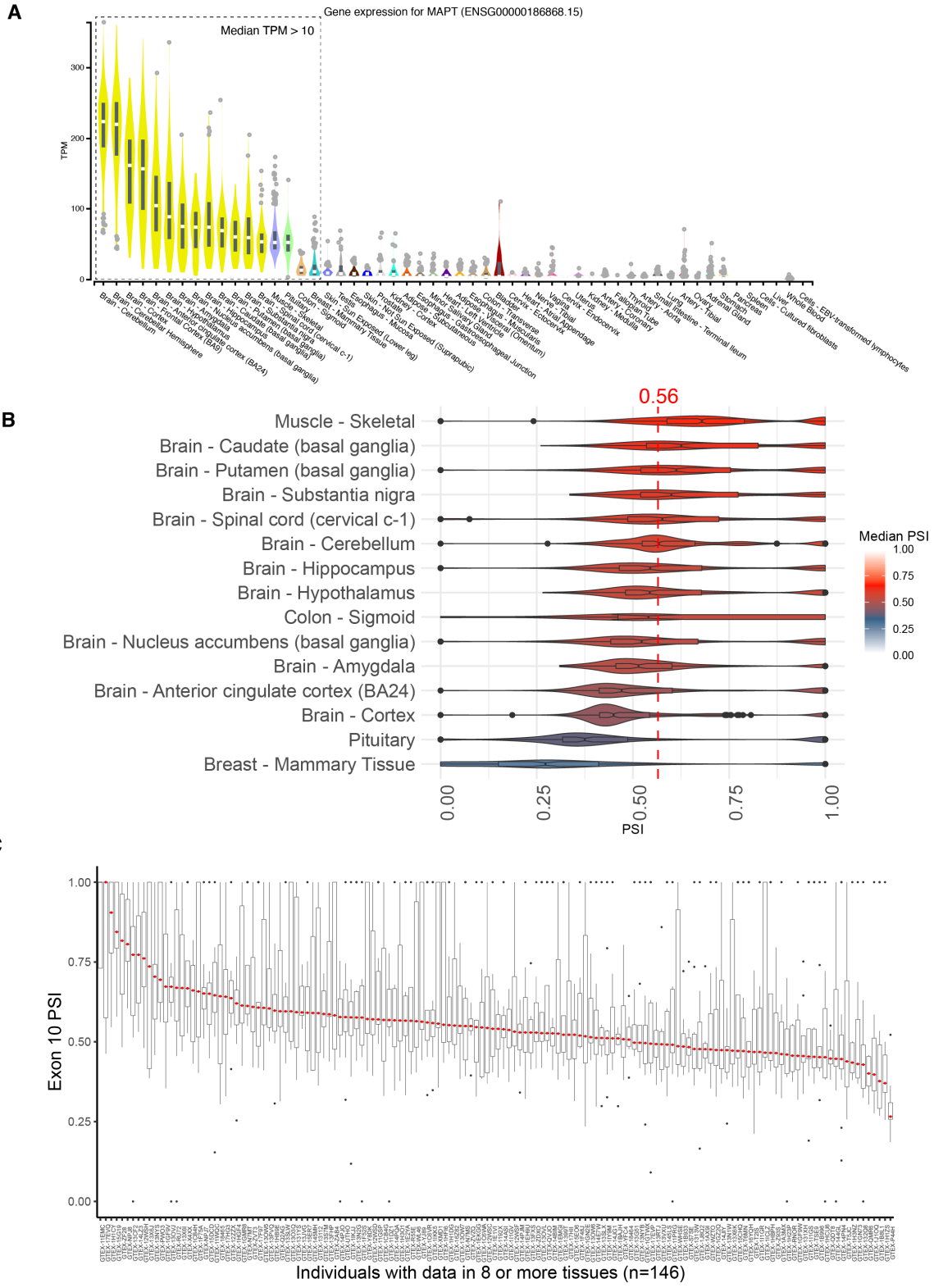


Figure 1-figure supplement 1: Distribution of Exon 10 PSIs calculated for RNA-seq data from GTEx database.

A) Distribution of TPM values of *MAPT* gene expression for tissues in the GTEx database sorted by median TPM. Dotted box indicates tissues with median TPM greater than 10. MAPT is highly expressed in the brain, and there is little expression in other tissues. Figure was downloaded from the GTEx website.
B) Distribution of Exon 10 PSI for 12 central nervous system, muscle-skeletal, colon-sigmoid, and breast-mammary tissue types. Percent Spliced In (PSI) of Exon 10 was calculated from RNA-seq data for 2,962 tissue samples among 15 tissue types collected from 818 individuals in GTEx v8 database. The violin plot for each tissue type and the corresponding region on the brain diagram is colored by the median PSI for all samples of that type. The median PSI of 0.56 for all tissue samples is indicated by the red dotted line.
C) Distribution of Exon 10 PSI for tissues per individual. Only individuals with *MAPT* expression data in 8 or more tissues were plotted. Median PSI for each individual is labelled by red dot on box plot.

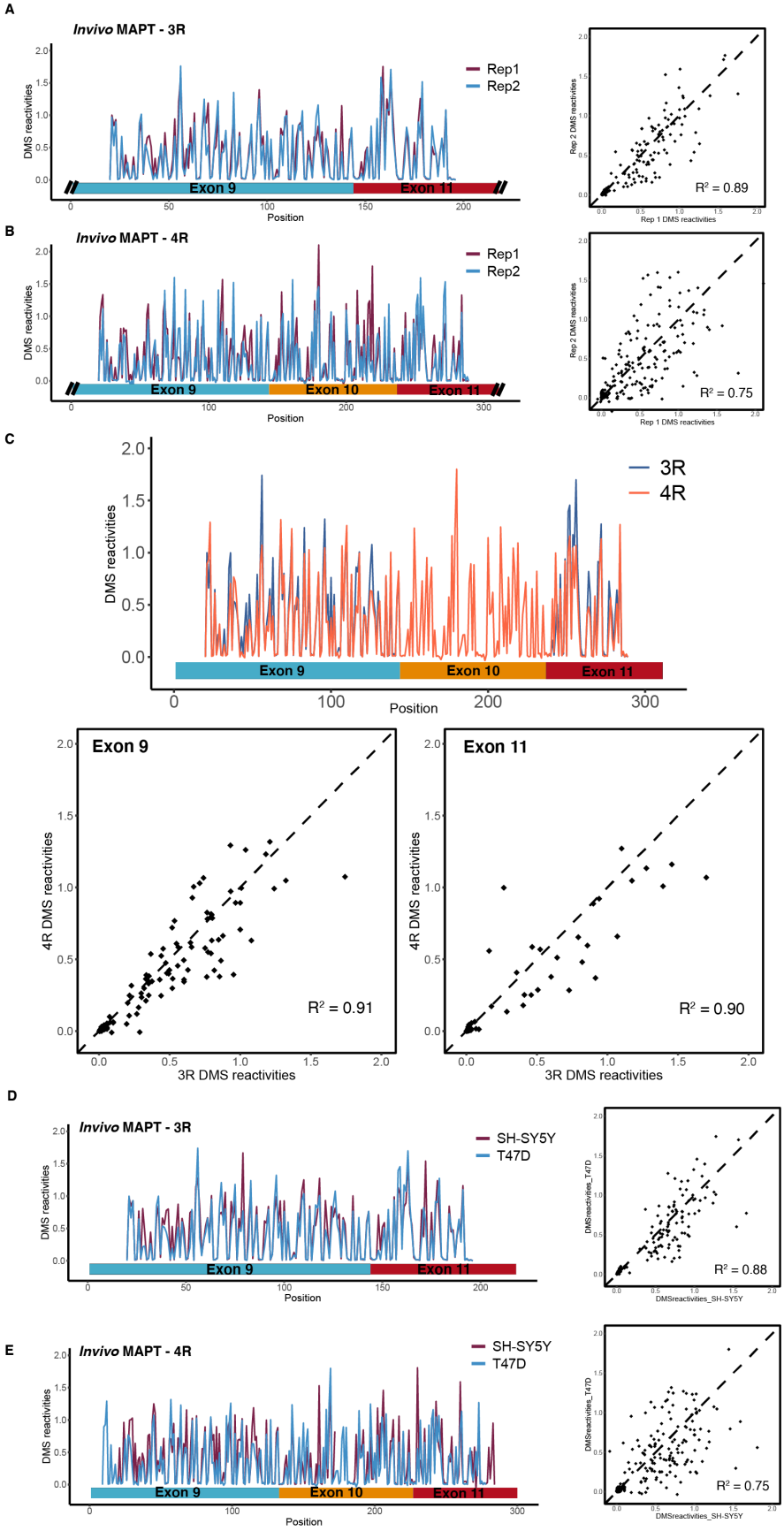

A
Invivo MAPT - 3R
Rep1
Rep2
DMS reactivities
Exon 9
Exon 11
Position
Rep 2 DMS reactivities
Rep 1 DMS reactivities
R² = 0.89
B
Invivo MAPT - 4R
Rep1
Rep2
DMS reactivities
Exon 9
Exon 10
Exon 11
Position
Rep 2 DMS reactivities
Rep 1 DMS reactivities
R² = 0.75
C
3R
4R
DMS reactivities
Exon 9
Exon 10
Exon 11
Position
Exon 9
4R DMS reactivities
3R DMS reactivities
R² = 0.91
Exon 11
4R DMS reactivities
3R DMS reactivities
R² = 0.90
D
Invivo MAPT - 3R
SH-SY5Y
T47D
DMS reactivities
Exon 9
Exon 11
Position
DMSreactivities_T47D
DMSreactivities_SH-SY5Y
R² = 0.88
E
Invivo MAPT - 4R
SH-SY5Y
T47D
DMS reactivities
Exon 9
Exon 10
Exon 11
Position
DMSreactivities_T47D
DMSreactivities_SH-SY5Y
R² = 0.75

Figure 1-figure supplement 2: DMS structure probing data for mature *MAPT* 3R and 4R isoforms

A) DMS reactivity data from T47D cells for two replicates for Exon 9-Exon 11 junction (3R isoform). Structure probing data for junctions of interest were obtained using primers (Supplementary file 4) following RT of extracted RNA. DMS reactivity is plotted for each nucleotide across spliced junctions for both replicates overlaid in plot on the left. For scatter plot on the right, DMS reactivity for Rep 1 vs Rep 2 is plotted per nucleotide with Pearson's correlation coefficient displayed.
B) DMS reactivity data from T47D cells for two replicates for Exon 9-Exon 10-Exon 11 junction (4R isoform).
C) Comparison of DMS reactivity data for 3R vs 4R isoforms. Replicates 1 and 2 were pooled for each isoform. Top plot shows DMS reactivity plotted for each nucleotide with isoforms overlaid. No data were collected for Exon 10 for the 3R isoform because Exon 10 is spliced out. Bottom left scatter plot shows DMS reactivities for Exon 9 in the 3R vs 4R context, whereas bottom right scatter plot shows DMS reactivities for Exon 11 in the 3R vs 4R context. Pearson's correlation coefficient is shown for each comparison.
D) DMS reactivity data from T47D and SH-SY5Y cells for Exon 9-Exon 11 junction (3R isoform). Replicates from T47D cells were pooled. DMS reactivity is plotted for each nucleotide across spliced junctions for both replicates overlaid in plot on the left. For scatter plot on the right, DMS reactivity for T47D vs SH-SY5Y is plotted per nucleotide with Pearson's correlation coefficient displayed.
E) DMS reactivity data from T47D and SH-SY5Y cells for Exon 9-Exon 10-Exon 11 junction (4R isoform).

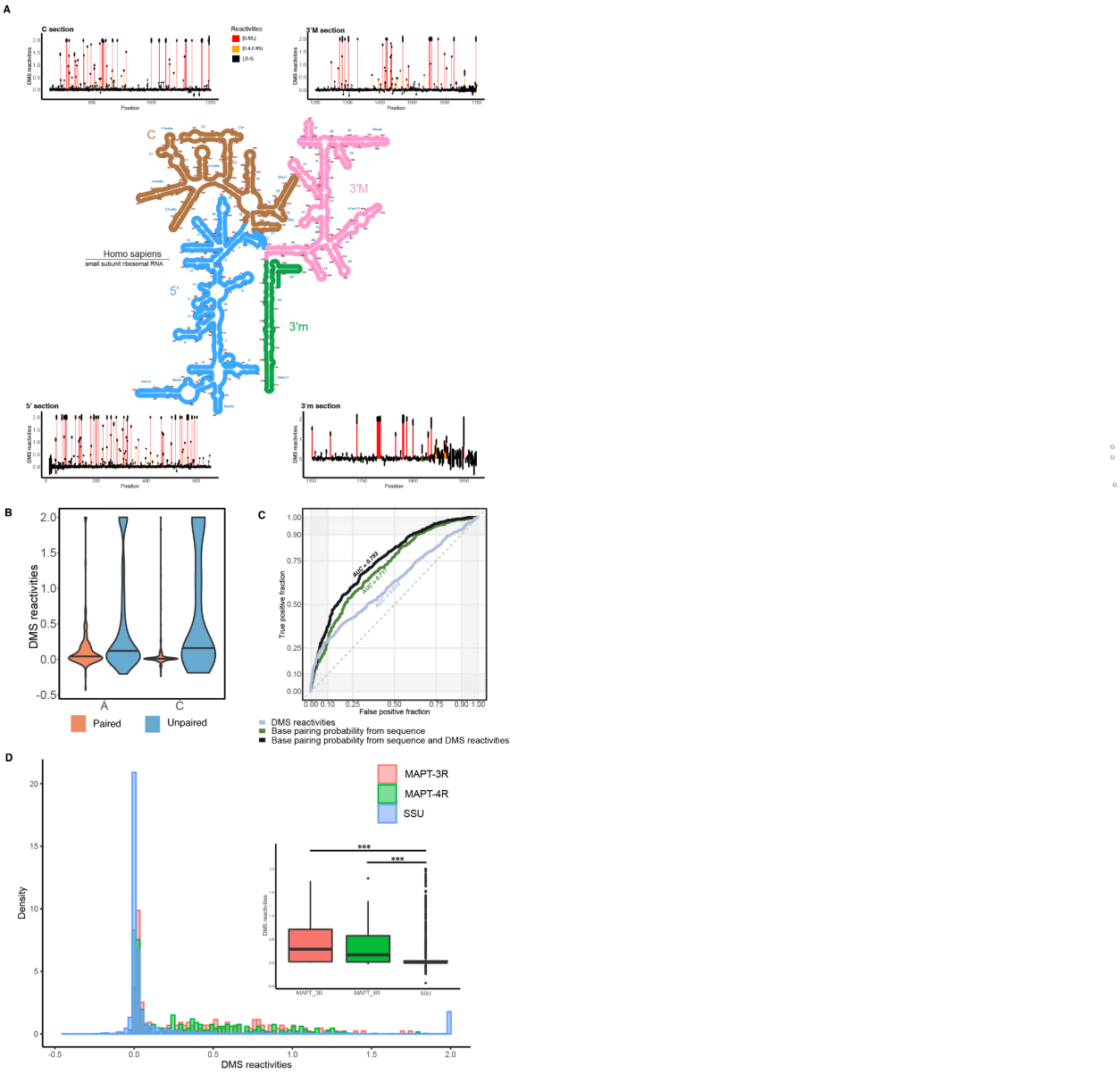


Figure 1-figure supplement 3: DMS structure probing data for human small subunit ribosome RNA (SSU) collected from T47D cells.

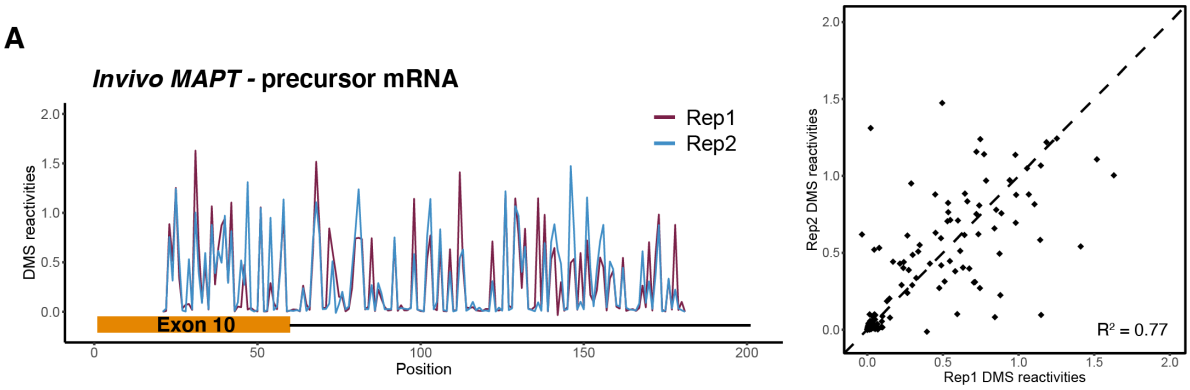
A) T47D cells were treated with DMS. Data for SSU were extracted by aligning reverse transcribed RNA-seq data against the SSU sequence, after which reactivities were calculated. DMS reactivities are plotted for each of the four sub-domains of the SSU. Each value is shown with its standard error and colored by reactivity based on color scale. Reactivities were cut off at 2. High DMS reactivities correspond to unstructured regions, whereas low DMS reactivities correspond to structured regions. The secondary structure of the SSU was downloaded from Loren William's

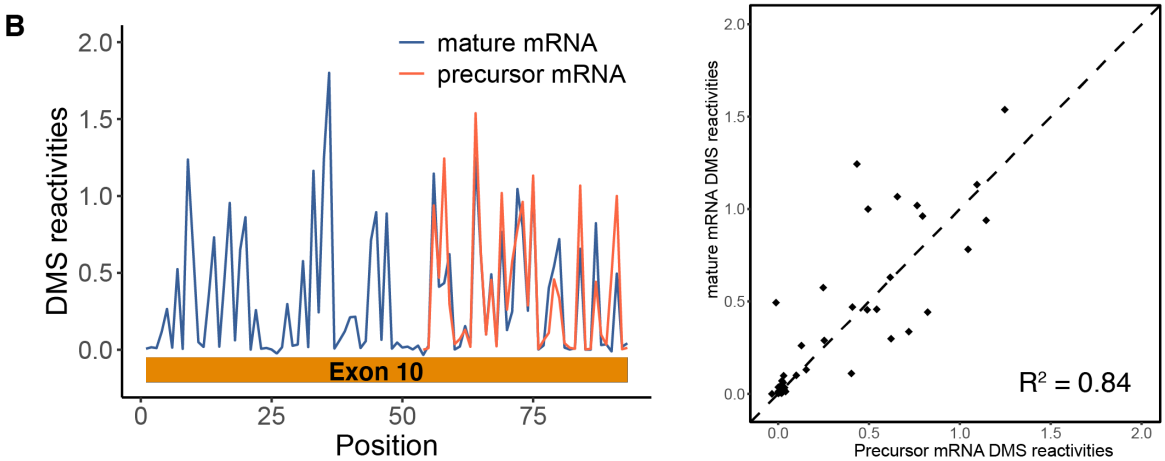
lab Ribosome Gallery website (http://apollo.chemistry.gatech.edu/RibosomeGallery/eukarya/H%20sapiens/SSU/index.html).

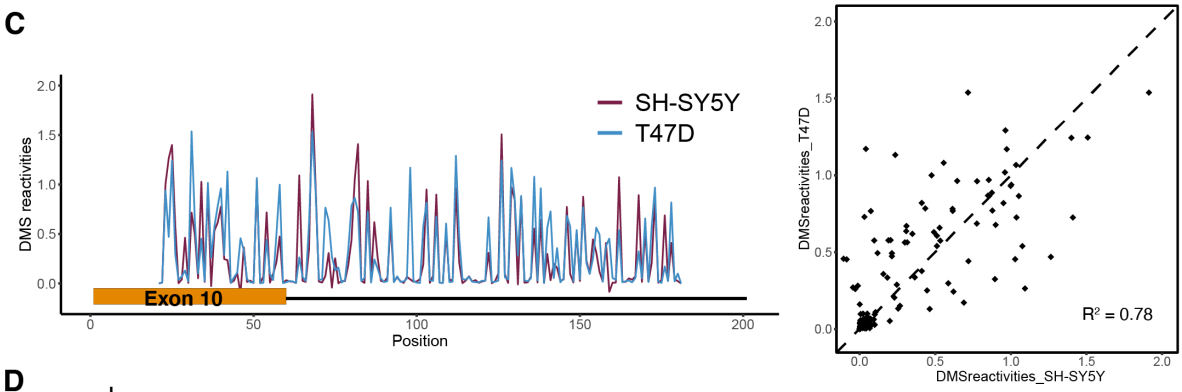
B) Violin plots showing distribution of DMS reactivities for adenines and cytosines partitioned by paired versus unpaired nucleotides. Pairing status of nucleotides was determined from the known secondary structure of the SSU. Median DMS reactivity is indicated by thick horizontal black line on violin plot.

C) ROC curves for predicting whether a nucleotide in the SSU is paired. Three different parameters were used for each of the three curves: DMS reactivities, base pairing probabilities predicted from SSU sequence, and base pairing probabilities for SSU sequence that were guided by DMS reactivities. The area under the curve (AUC) for each curve was calculated with AUCs closer to 1 corresponding to higher accuracy of predictions. Dotted line indicates AUC of 0.5 which corresponds to a model making random predictions.

D) Comparison of distribution of DMS reactivities between SSU, *MAPT* 3R and 4R isoforms. Larger plot shows a density histogram of the DMS reactivities for each RNA. Inset boxplots display distribution of DMS reactivities. Level of significance: ***p-value $< 10^{-6}$


A
Invivo MAPT - precursor mRNA
Rep1
Rep2
DMS reactivities
Exon 10
Position
Rep2 DMS reactivities
Rep1 DMS reactivities
R² = 0.77


B
mature mRNA
precursor mRNA
DMS reactivities
Exon 10
Position
mature mRNA DMS reactivities
Precursor mRNA DMS reactivities
R² = 0.84


C
SH-SY5Y
T47D
DMS reactivities
Exon 10
Position
DMSreactivities_T47D
DMSreactivities_SH-SY5Y
R² = 0.78

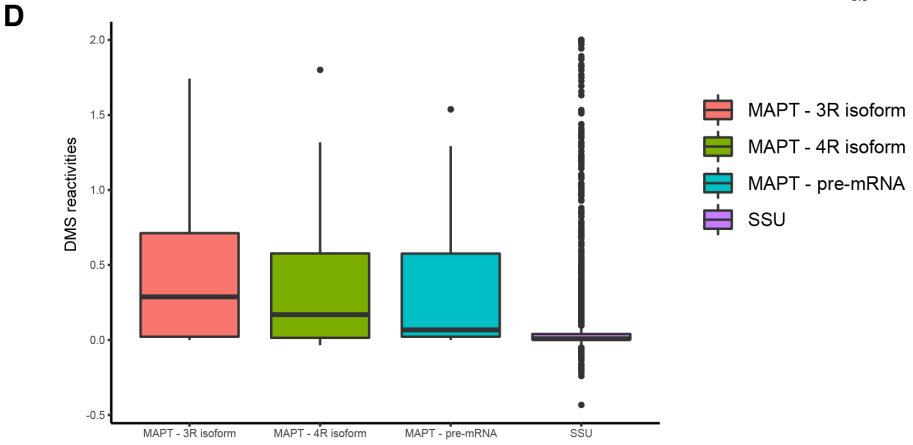
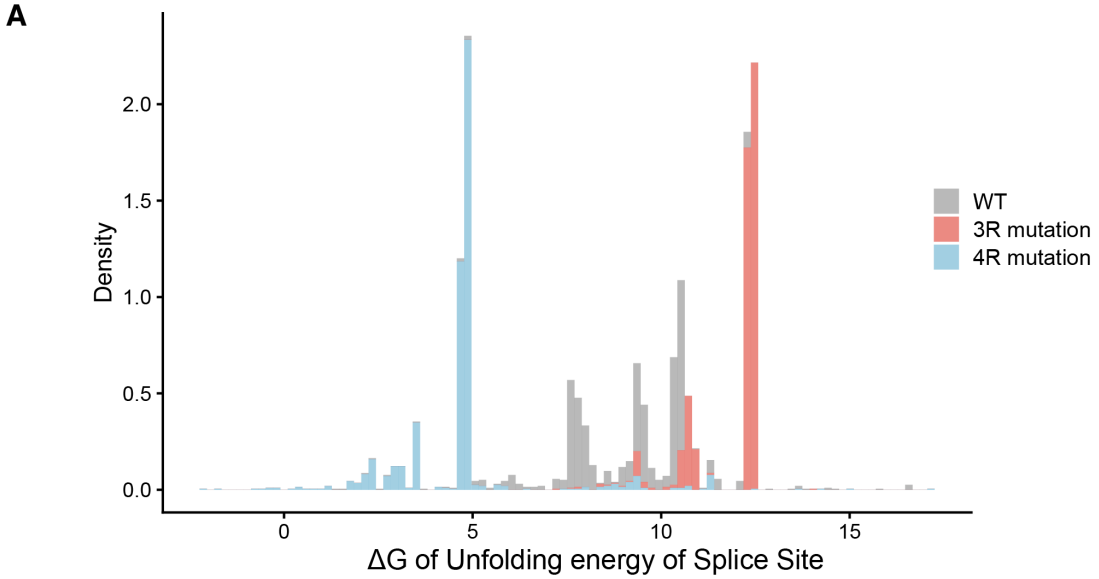

D
DMS reactivities
MAPT - 3R isoform
MAPT - 4R isoform
MAPT - pre-mRNA
SSU

Figure 2-figure supplement 1: DMS structure probing data for precursor *MAPT* Exon 10-Intron 10 junction

A) DMS reactivity data from T47D cells for two replicates for Exon 10-Intron 10 junction. Structure probing data for junctions of interest were obtained using primers following RT of extracted RNA. DMS reactivity is plotted for each nucleotide across spliced junctions for both replicates overlaid in plot on the left. For scatter plot on the right, DMS reactivity for Rep 1 vs Rep 2 is plotted per nucleotide with Pearson's correlation coefficient displayed.
B) DMS reactivity data comparing Exon 10 in precursor vs mature transcript. Replicates 1 and 2 were pooled for each transcript. Right plot shows DMS reactivity plotted for each nucleotide with mature and precursor RNAs overlaid. DMS data for all of Exon 10 could not be collected for the precursor RNA due to the position of primers chosen for amplification. Scatter plot on the left shows DMS reactivities for Exon 10 in the precursor vs mature mRNA context with Pearson's correlation coefficient shown for the comparison.
C) DMS reactivity data from T47D and SH-SY5Y cells for Exon 10-Intron 10 junction. Replicates from T47D cells were pooled. DMS reactivities are plotted for each nucleotide across exon-intron junctions for both cell types overlaid in plot on the left. For scatter plot on the right, DMS reactivity for T47D vs SH-SY5Y is plotted per nucleotide with Pearson's correlation coefficient displayed.
D) Boxplots of distribution of DMS reactivities between SSU, *MAPT* 3R isoform, 4R isoform and pre-cursor mRNA.


A
Density
2.0
1.5
1.0
0.5
0.0
0
5
10
15
ΔG of Unfolding energy of Splice Site
WT
3R mutation
4R mutation

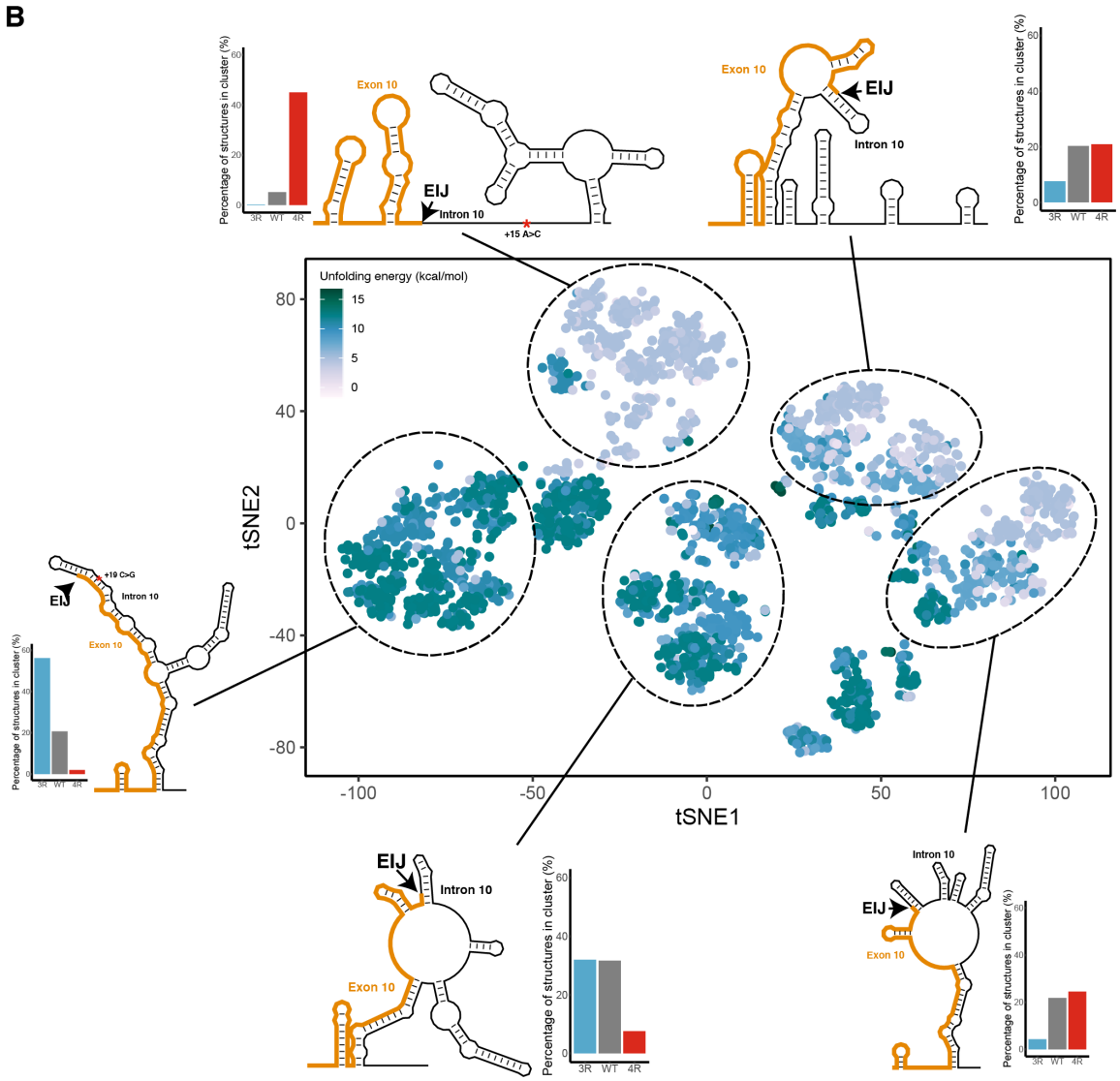

B
Percentage of structures in cluster (%)
3R
WT
4R
Exon 10
EIJ
Intron 10
+15 A>C
Unfolding energy (kcal/mol)
15
10
5
0
tSNE2
tSNE1
80
40
0
-40
-80
-100
-50
0
50
100
+19 C>G

Figure 2-figure supplement 2: The 3R and 4R mutations shift the WT structural ensemble towards less and more accessible exon-intron junctions, respectively.

A) Density histogram showing the distribution of unfolding free energies of the splice site (defined as last 3 nucleotides of exon, first 6 nucleotides of intron) for all structures in the ensemble for WT, 3R and 4R mutated sequence. Distributions for each sequence are colored according to the legend.

B) t-SNE Visualization of structural ensemble of wild type (WT) and, 3R (+19C>G) and 4R (+15A>C) mutations. Structures were predicted using Boltzmann suboptimal sampling and guided by DMS reactivity data (in Figure 2A). Data were visualized using t-Distributed Stochastic Neighbor Embedding (t-SNE). Shown are 3000 structures corresponding to 1000 structures per category. Each dot represents a single structure and was colored by calculated unfolding free energy of splice site at exon-intron junction (3 exonic bases, 6 intronic bases). Data were clustered by k-means clustering and representative structures for five clusters are shown. Bar plots next to the representative structure show the proportion of the cluster in WT, 3R and 4R. The exon-intron junction is marked by EIJ on each structure. Position of 3R and 4R mutations are marked by a red asterisk on their respective representative structures. There are two additional representative structures shared by WT and 4R sequence which have similar structural contexts around the EIJ as the representative WT structure in Figure 2B.

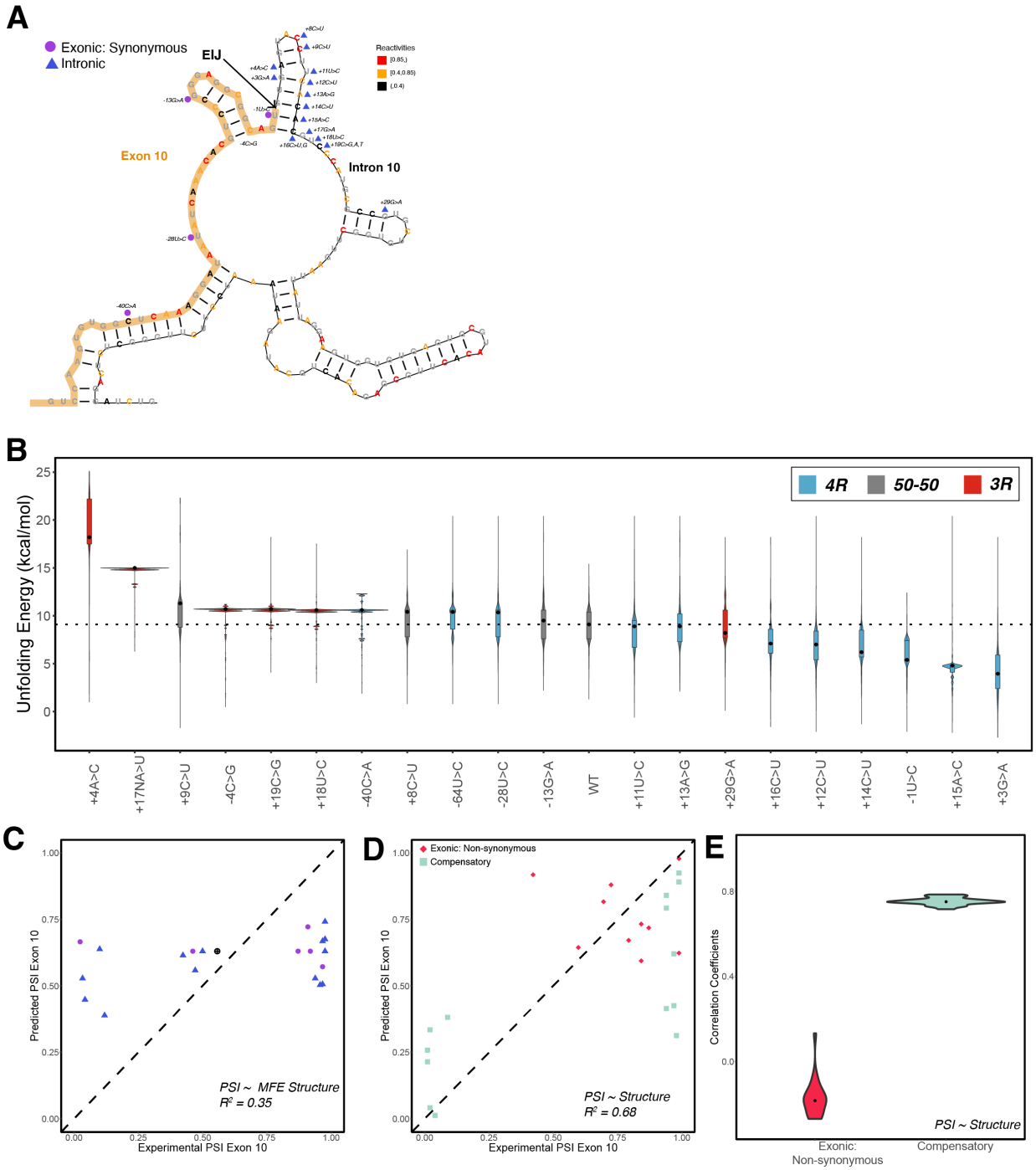


Figure 3-figure supplement 1: Structure is a poor predictor of Exon 10 PSI for exonic non-synonymous mutations

A) Positions of intronic and synonymous experimentally validated mutations, used in training the structure model, are shown on the wild type representative structure

from Figure 2B. Each nucleotide is colored by its corresponding reactivity value based on the color scale.

B) Violin plots showing the distribution of unfolding free energy of the exon-intron pre-mRNA in the spliceosome B^{act} complex for the 1000 structures in the ensembles of the 22 intronic and synonymous mutations. Each violin plot is colored by whether the mutation promotes the 3R or 4R isoform ratio or the ratio remains 50:50.

C) Exon 10 PSIs of 22 mutations predicted using unfolding free energy of the exon-intron pre-mRNA in B^{act} complex of the spliceosome for the single minimum free energy (MFE) structure and plotted against experimental PSIs measured in splicing assays. Exon 10 PSIs predicted using Eq. 2. Each point on the scatterplot represents a mutation and is colored by mutation category. Dotted diagonal line is the x=y line, and the closer the points are to the diagonal, the more accurate the prediction. Pearson correlation coefficient (R^2) of experimental to predicted PSIs was calculated.

D) Exon 10 PSIs of non-synonymous and compensatory mutations predicted using the unfolding free energy of pre-mRNA within the spliceosome B^{act} stage plotted against corresponding experimental PSIs measured in splicing assays. Exon 10 PSIs were predicted using Eq. 1.

E) Violin plots show R^2s calculated for each mutation category by training and testing on subsets of all mutations by non-parametric bootstrapping; Non-synonymous (n=10), Compensatory (n=14).

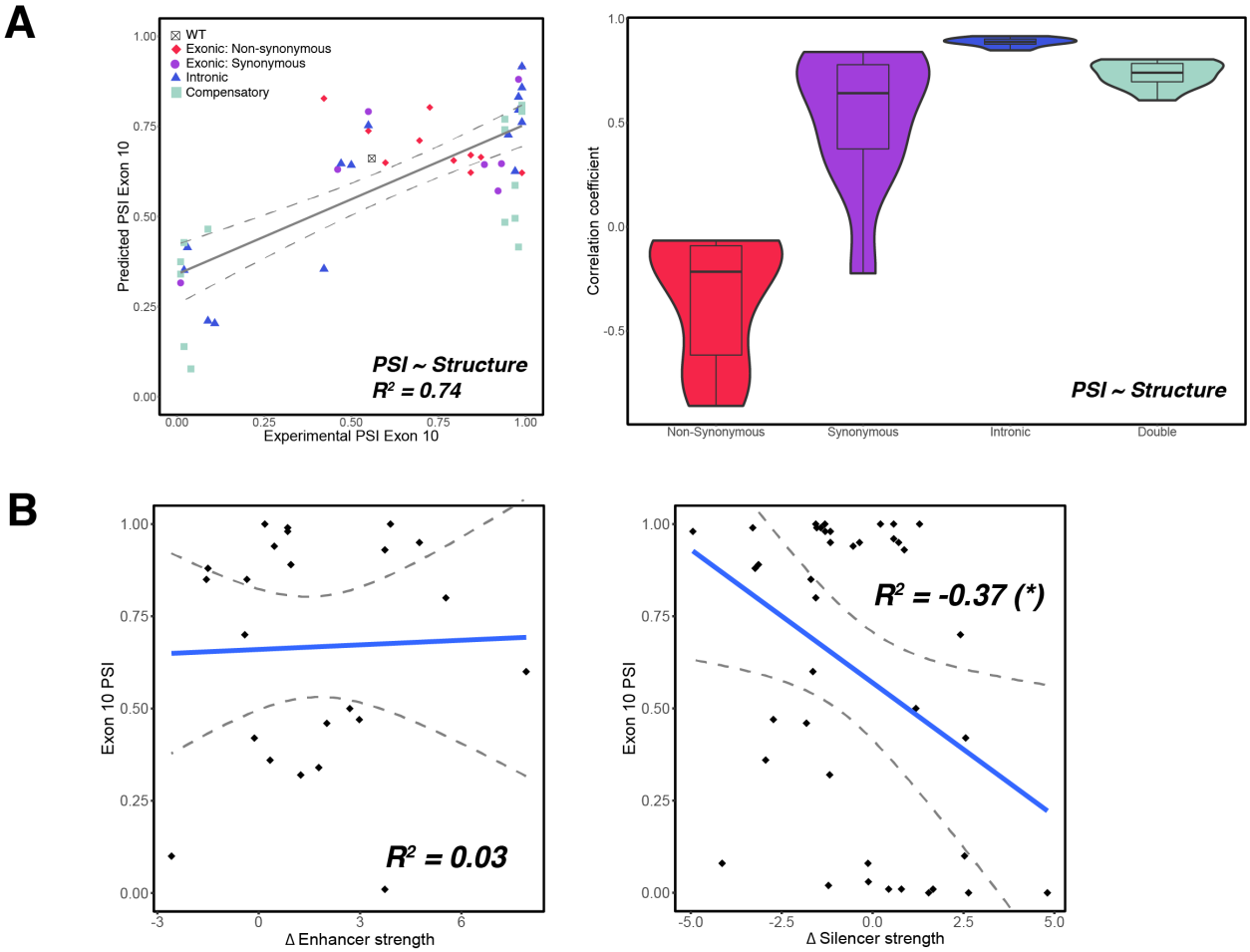


Figure 4-figure supplement 1: RBP binding motif strength is a poor predictor of Exon 10 PSI for all mutations

A) Exon 10 PSIs of 47 mutations predicted from structural change and plotted against experimental PSIs measured in splicing assays. Exon 10 PSIs predicted using Eq. 1. Each point on the scatterplot represents a mutation and is colored by mutation category. Grey line represents the best fit with dotted lines indicating the 95% confidence interval. Pearson correlation coefficient (R^2) of experimental to predicted PSIs. Violin plot shows R^2s calculated for each category by training and testing on subsets of all mutations by non-parametric bootstrapping; Exonic non-synonymous (n=11), Exonic synonymous (n=7), Intronic (n=15), Compensatory (n=14), Wildtype (n=1).

B) Scatter plot of change in enhancer or silencer strength versus Exon 10 PSI. Each point represents a mutation. Blue line represents the line of best fit with dotted

lines indicating the 95% confidence interval. Pearson correlation coefficient (R^2) is shown. The negative correlation between silencer strength and Exon 10 PSI is statistically significant with a p-value of 0.004.

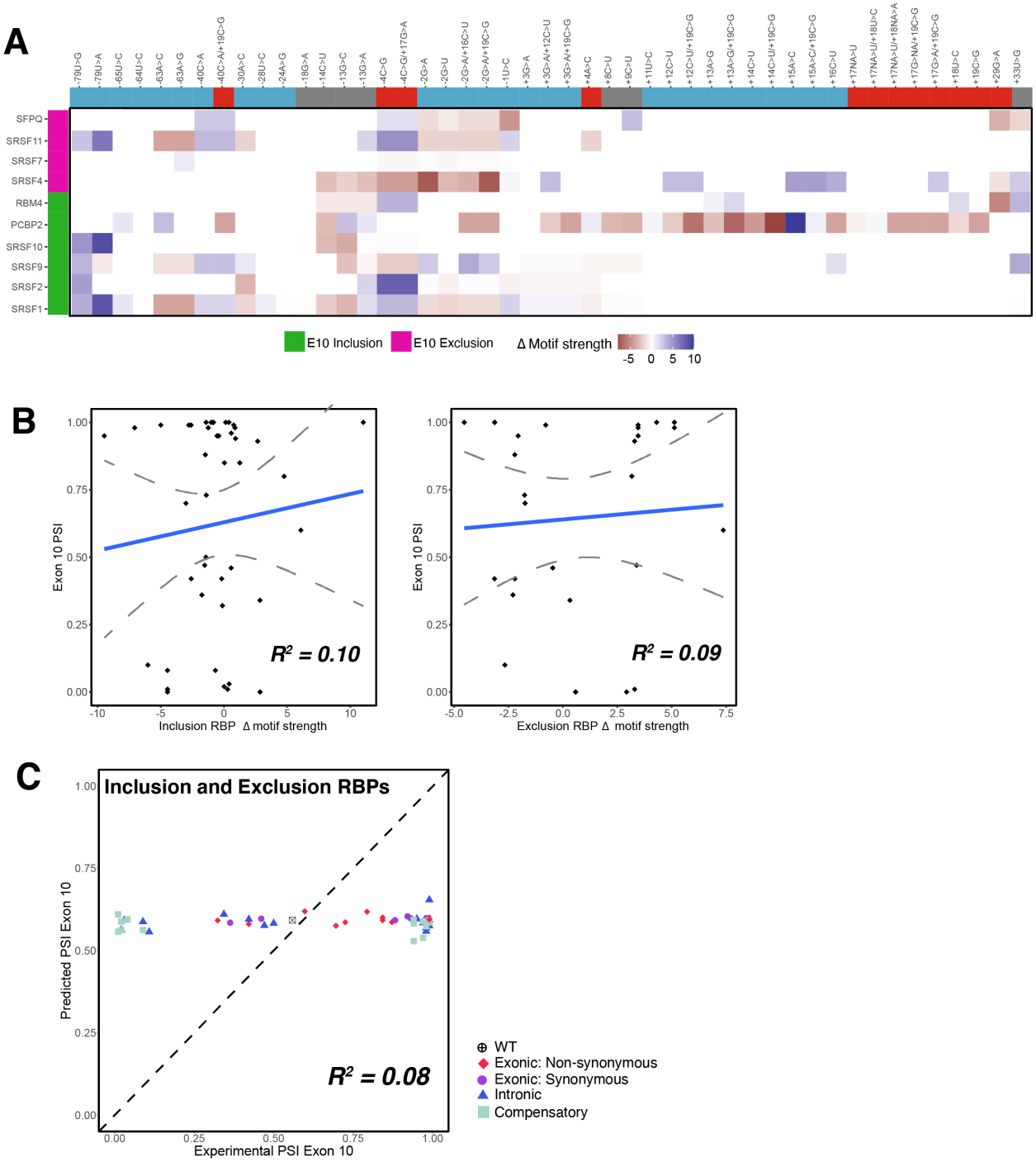


Figure 4-figure supplement 2: RBP binding motif strength is a poor predictor of Exon 10 PSI for all mutations

A) Heatmap of relative RBP binding motif strengths compared to wild type for 44 mutations. A value of 0 indicates that the mutation does not change RBP binding motif strength, a positive value indicates increase in RBP binding motif strength, and a negative value indicates weaker strength. RBPs implicated in the regulation of Exon 10 splicing were collected from Qian & Liu, 2014 and the RBP binding motifs were from Dominguez et al., 2018 and Ray et al., 2013. RBPs on the left, implicated in the splicing inclusion of MAPT Exon 10, are highlighted in pink, and RBPs involved in the exclusion of Exon 10 are highlighted in green. Mutations are marked based on whether they promote the 3R or 4R isoform ratio or the ratio remains 50:50.
B) Scatter plot displaying change in RBP motif strength versus Exon 10 PSI, categorized based on whether the RBP is implicated in exclusion or inclusion of Exon 10. Neither correlation coefficient is statistically significant.
C) Exon 10 PSIs of 44 mutations and wild type predicted using change in RBP motif strength and plotted against experimental PSIs measured in splicing assays. Exon 10 PSIs predicted using Eq. 5. Each point represents a mutation and is colored by category of mutation. Dotted diagonal line is the x=y line, and the closer the points are to the diagonal, the more accurate the prediction. Pearson correlation coefficient (R^2) of experimental to predicted PSIs was calculated.

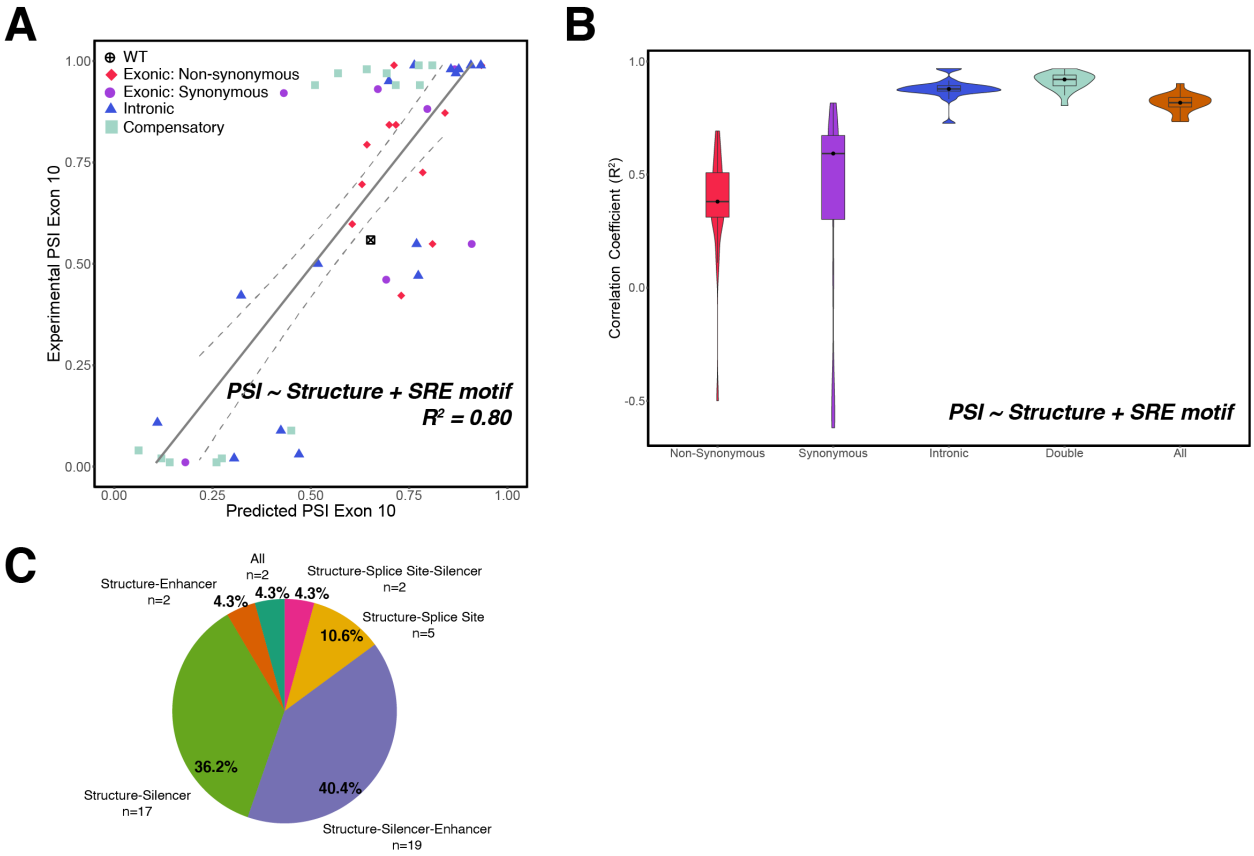


Figure 5-figure supplement 1: Additive model of structure and SRE has poorer predictive performance compared with an interactive model specifically for synonymous and non-synonymous mutations

A) Exon 10 PSIs of 47 mutations and wild type predicted using addition between structure and SRE strength and fit to experimental PSIs measured in splicing assays. Exon 10 PSIs are predicted using Eq. 7. Each point on scatterplot represents a mutation and is colored by category of mutation. Grey line represents the best fit with dotted lines indicating the 95% confidence interval. Pearson correlation coefficient (R^2) of experimental to predicted PSIs was calculated.

B) Violin plots showing correlation coefficients for each mutation category for structure and SRE additive model. R^2s calculated for each mutation category by training and testing on subsets of all mutations by non-parametric bootstrapping 10 times.

C) Pie chart of number and proportion of experimentally validated mutations in each cluster for heatmap in Fig 5B. Color of segment of pie chart matches up to the color of dendrogram branch in Fig 5B.

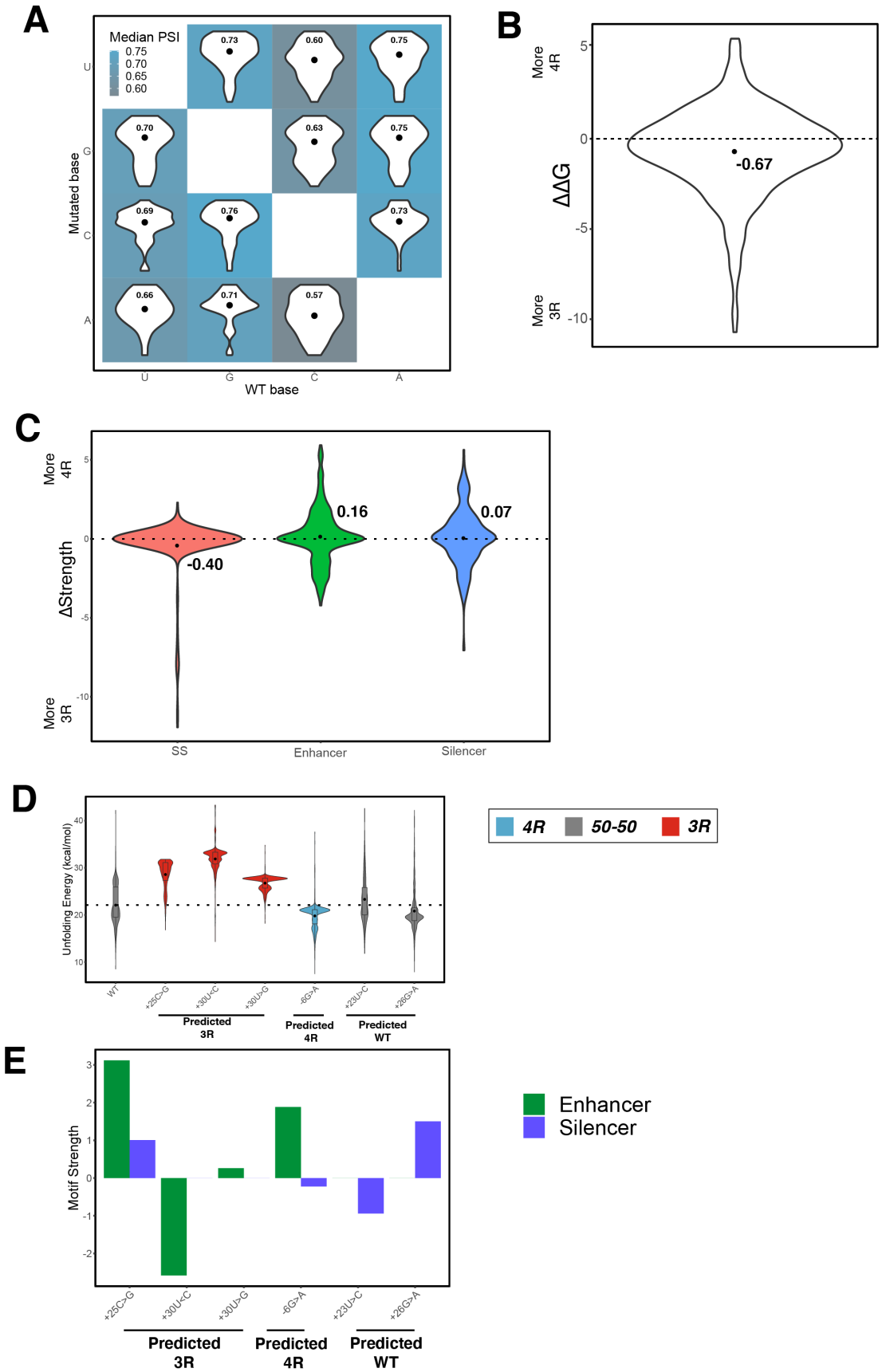

A
Median PSI
0.75
0.70
0.65
0.60
0.73
0.60
0.75
0.70
0.63
0.75
0.69
0.76
0.73
0.66
0.71
0.57
Mutated base
U
G
C
A
WT base
B
More 4R
More 3R
ΔΔG
-0.67
C
More 4R
More 3R
ΔStrength
-0.40
0.16
0.07
SS
Enhancer
Silencer
D
Unfolding Energy (kcal/mol)
WT
+25C>G
+30U<C
+30U>G
-6G>A
+23U>C
+26G>A
Predicted 3R
Predicted 4R
Predicted WT
4R
50-50
3R
E
Motif Strength
+25C>G
+30U<C
+30U>G
-6G>A
+23U>C
+26G>A
Predicted 3R
Predicted 4R
Predicted WT
Enhancer
Silencer

Figure 6-figure supplement 1: Complete mutagenesis of 100-nucleotide window spanning Exon 10-Intron 10 junction

A) Heatmap of mean predicted Exon 10 PSIs grouped by wild type and mutant nucleotide. Mutations were grouped by wild type and mutant nucleotide, and mean predicted PSIs were calculated by group and colored according to color scale. Violin plots of the distribution of PSI per group are shown in tile corresponding to group. On each tile, mean PSI is indicated by dot and labeled within violin plot.

B) Violin plot of the distribution of normalized change in unfolding free energy of the exon-intron pre-mRNA in the spliceosome B^{act} complex from WT for all mutations around a 100-nucleotide window of exon-intron junction. Mean of -0.67 is indicated by dot. Dotted line represents the 0 value where there is no difference between WT and mutant unfolding free energy. Positive values imply region becomes less structured and has increased inclusion of Exon 10 (4R isoform); negative values are interpreted as more structured and decreased inclusion of Exon 10 (3R isoform).

C) Violin plots showing the distribution of normalized change in splice site, enhancer, and silencer strength compared with WT for all mutations spanning a 100-nucleotide window of exon-intron junction. Mean is indicated by large black dots on violin plots. Dotted lines represent the 0 value where there is no difference from WT strength for mutation. Positive values suggest increased inclusion of Exon 10 (4R isoform), whereas negative values are interpreted as decreased inclusion of Exon 10 (3R isoform).

D) Violin plot shows the distribution of unfolding free energy of the exon-intron pre-mRNA in the spliceosome B^{act} complex for the 1000 structures in the ensembles of wild type and the 6 VUSs experimentally tested. Each violin plot is colored by whether the mutation promotes the 3R or 4R isoform ratio or the ratio remains 50:50. The dotted line indicates the median unfolding free energy of the WT ensemble.

E) Bar plots display the change in enhancer and silencer strength of the 6 VUSs compared with WT.

Supplementary Files

Supplementary file 1: ANOVA table for between individuals and within individuals Exon 10 PSI comparison

Supplementary file 2: Details on 47 experimentally tested *MAPT* mutations used in training model

Supplementary file 3: Details on 55 variants of unknown significance (VUSs) in *MAPT* from dbSNP

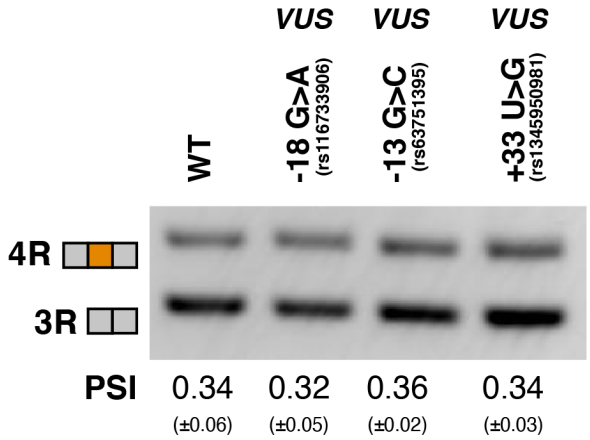
Supplementary file 4: Primers used for amplification of exon-exon or exon-intron junctions

Supplementary file 5: Re-calculated Position Weight Matrices for ESEs, ESSs, ISEs, ISSs

Supplementary file 6: Details on beta regression model results and features used for each training and test set

Supplementary file 7: Gel of RT-PCR data for splicing assay for new WT VUSs

Supplementary file 7



Representative gel of RT-PCR data for splicing assay in the presence of WT VUSs. Splicing reporter was transfected into HEK293 cells. The mean Exon 10 PSI displayed for each variant was calculated from three replicates and standard error is shown in brackets below.