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

Authors: Jayashree Kumar[a,b], Lela Lackey[a,c], Justin M. Waldern[a], Abhishek Dey[a], Anthony M. Mustoe[d], Kevin M. Weeks[e], David H. Mathews[f], 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] Verna and Marrs McClean Department of Biochemistry and Molecular Biology; Department of Molecular and Human Genetics; Therapeutic Innovation Center (THINC), Baylor College of Medicine, Houston, TX 77030

[e] Department of Chemistry, University of North Carolina, Chapel Hill, NC 27599-3290

[f] 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 highly regulated and is modulated by numerous factors. Quantitative predictions for how a mutation will affect precursor messenger RNA (mRNA) structure and downstream function is particularly challenging. Here we use a novel chemical probing strategy to visualize endogenous precursor and mature *MAPT* mRNA structures in cells. We used these data to estimate Boltzmann suboptimal structural ensembles, which were then analyzed to predict consequences of mutations on precursor mRNA structure. Further analysis of recent cryo-EM structures of the spliceosome at different stages of the splicing cycle revealed that the footprint of the B^{act} complex with precursor mRNA best predicted alternative splicing outcomes for exon 10 inclusion of the alternatively spliced *MAPT* gene, achieving 74% accuracy. We further developed a β-

regression weighting framework that incorporates splice site strength, RNA structure, and exonic/intronic splicing regulatory elements capable of predicting, with 90% accuracy, the effects of 47 known and six newly discovered mutations on inclusion of exon 10 of *MAPT*. This combined experimental and 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 a limited understanding of pre-mRNA structure-mediated splicing mechanisms (Taylor and Sobczak 2020). It has proven challenging to develop quantitative models capable of predicting splicing outcome, specifically the percent spliced in (PSI) for alternatively spliced exons. It is especially difficult to predict outcome alterations due to genetic variation at exon-intron junctions because mutations affect both the binding by RBPs and also pre-mRNA structure (Tazi, Bakkour, and Stamm 2009).

The consequences of mutations on pre-mRNA structure are difficult to predict. First, little is known about native pre-mRNA structure because pre-mRNAs are relatively short-lived in cells (Herzel et al. 2017). Only recently has high-resolution in-cell experimental characterization been applied to pre-mRNA structure determination (Mustoe, Busan, et al. 2018; Sun et al. 2019; Liu et al. 2021; Bubenik et al. 2020). Second, it is not clear which structures within a pre-mRNA modulate spliceosome assembly and activity. Finally, quantitative measures for the relative weighting of RBP affinity for individual motifs within a pre-mRNA relative to the importance of pre-mRNA structure are lacking.

In this study, we exploited several technical developments that address these issues to develop an integrated, RNA structure-based framework that accurately predicts splicing outcomes. We measured endogenous pre-mRNA structure in cells taking advantage of recent developments in mutational profiling (MaP) approaches for read-out of chemical probing data (Homan et al. 2014) with targeted amplification of specific exon-intron junctions. This novel approach enabled us to obtain single-nucleotide RNA structure probing data for endogenous pre- and mature mRNAs in the same cell. Our RNA structure modeling considers the equilibrium between multiple alternative structures (Dethoff et al. 2012; Lai et al. 2018) and employed data-guided Boltzmann suboptimal sampling (Spasic et al. 2018) to predict free energies of unfolding for structures in the ensemble. We additionally leveraged recent high-resolution structures of the spliceosome at various stages of the splicing cycle to deduce the effective spliceosomal footprint on pre-mRNA (L. Zhang et al. 2019), 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).

For validation of our framework, we studied the effects of 47 experimentally measured mutations near the exon 10 – intron 10 junction of the human *MAPT* gene, which encodes the Tau protein (Park, Ahn, and Gallo 2016; Catarina Silva and Haggarty 2020). Exons 9, 10, 11, and 12 encode the critical microtubule binding repeat domain in Tau. Exons 9, 11, and 12 are constitutively spliced, but exon 10 is alternatively spliced resulting in *MAPT* isoforms with either four microtubule binding repeats (4R) or three repeats (3R) when exon 10 is included or skipped, respectively. The normal ratio of 3R to 4R isoforms is approximately 1:1 (Hefti et al. 2018). Twenty-nine clinically validated disease-causing mutations have been identified in the region of the exon 10 – intron 10 junction (Stenson et al. 2003). These mutations result in impaired Tau function and are implicated in neurodegenerative disease (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 of the disease-associated mutations deregulate *MAPT* pre-mRNA splicing altering the ratio of 3R to 4R

(Hutton et al. 1998; D'Souza et al. 1999; Hasegawa et al. 1999; Jiang et al. 2000). The effect of an additional 27 mutations on exon 10 inclusion have been experimentally determined using cell-based splicing assays (D'Souza and Schellenberg 2000; Tan et al. 2019; Grover et al. 1999). The exon 10 junction is the best experimentally characterized junction of clinical importance in the human genome and is thus an excellent system for developing forward-predictive models of splicing. Our work provides a framework for integrating endogenous pre-mRNA structure probing data with a structure-based understanding of spliceosome assembly and *trans*-acting RBPs to qualitatively predict the effect of mutations at exon-intron junctions on splicing.

Results

MAPT 3R and 4R mRNA isoforms are expressed at a consistent 1:1 ratio across tissues

To confirm that MAPT pre-mRNA splicing results in a 1:1 ratio of alternatively spliced isoforms (Goedert et al. 1989; Andreadis 2005) in a large population, we analyzed RNA-sequencing data from the Genotype-Tissue Expression (GTEx) database (Lonsdale et al. 2013). We analyzed data from tissue types with median *MAPT* transcripts per million greater than 10 (Figure 1–figure supplement 1A) and calculated the PSI value for exon 10 for each sample (Figure 1A-source data 1; Materials and methods). We examined data from 2,315 tissue samples from 375 individuals of median age 61 (Figure 1A and Figure 1–figure supplement 1B). A PSI of 0 indicates that none of the *MAPT* transcripts in a sample included exon 10 (3R), whereas a PSI of 1 corresponds to exon 10 inclusion in all transcripts in a sample (4R).

PSI for exon 10 varied across tissue types and within and between individuals. However, 75% of samples were within a standard deviation of the median PSI of 0.54, demonstrating that the 3R to 4R isoform ratio was close to 1:1 among individuals and across tissues. Within the brain, the pituitary gland demonstrated the largest variation in PSI and the cerebellum the least variation. The pituitary gland also had the lowest median PSI (0.38). However, the median PSI differed by no more than 0.25 across all

brain tissues. Interestingly, although *MAPT* function in breast tissue is not understood compared with its function in the brain, there was greater variation in PSI in breast tissue, and the median PSI in breast tissue was lower than in the pituitary gland (Figure 1–figure supplement 1B). There was also a large amount of variation within tissues of an individual (Figure 1–figure supplement 1C), although there was significantly greater variation between than within individuals (see Supplementary file 1 for ANOVA table). Furthermore, exon 10 inclusion variability (0.2) was between the variability for a *MAPT* constitutively spliced exon (0.1) and another *MAPT* alternatively spliced exon (0.3) (Figure 1–figure supplement 1D). As levels of RBP expression varied considerably across individuals and tissues (Figure 1–figure supplement 1E), sequence and structural features of the *MAPT* pre-mRNA likely regulate inclusion of exon 10.

Structures of 3R and 4R *MAPT* mature mRNA isoforms are similar and mostly unstructured

The structures of the mature 3R and 4R isoforms and *MAPT* pre-mRNA have not been assessed in their endogenous context in cells. Here, we used dimethyl sulfate probing read out by mutational profiling (DMS-MaP) as described previously (Mustoe et al. 2019; Homan et al. 2014) to asses *MAPT* pre-mRNA and mature mRNA structures in T47D cells, a breast cancer line, and in neuronal SH-SY5Y cells. We used region-specific primers (Smola et al. 2015) to selectively amplify mature 3R and 4R transcripts during library preparation (Supplementary file 4; Materials and methods). This approach leverages the read-through capability of MaP technology to probe the structure of distinct alternatively spliced isoforms in the same cells. High DMS reactivities correspond to less structured regions, whereas low DMS reactivities correspond to more structured regions. DMS reactivities for replicates and cell lines were highly correlated (Figure 1–figure supplement 2A; Figure 1–figure supplement 2B; Figure 1–figure supplement 2D; Figure 1–figure supplement 2E).

As an internal control for our probing experiments, we also collected DMS-MaP data for the small subunit ribosomal RNA (SSU), which has a well-defined secondary structure (Petrov et al. 2014). As expected, the DMS reactivities of unpaired nucleotides were

significantly higher than for paired nucleotides both for RNA probed in cells and for RNA isolated from cells prior to probing (Figure 1–figure supplement 3A and B). This experiment confirmed that our DMS probing recapitulates native RNA secondary structure regardless of the presence of proteins, consistent with previous studies (Woods et al. 2017; Lackey et al. 2018). We used the SSU in-cell reactivity data to calibrate the estimation of equilibrium ensembles (Materials and methods), and we confirmed that structure modeling guided by experimental DMS reactivities yielded a more accurate estimation of the SSU structure than the model not informed by chemical probing data (Figure 1–figure supplement 3C).

The median in-cell DMS reactivity of the mature *MAPT* isoforms was 0.22, significantly greater than the median in-cell DMS reactivity of the SSU, which was 0.008 (Figure 1–figure supplement 3D). This difference was recapitulated in cell-free samples (Figure 1–figure supplement 3D). These results suggested that the nucleotides of the mature *MAPT* isoforms were more accessible and less paired overall as compared with the highly structured SSU. Reactivities of exon 9 and exon 11 were highly correlated between the 3R and 4R isoforms (Figure 1B and C; Figure 1–figure supplement 2C). In the 4R isoform, approximately 89% of base pairs were contained within the exon units; only 11% of base pairs were between residues from exon 10 with those of exon 11 (Figure 1–figure supplement 2F). This result suggests that the mature exons fold as independent structural units.

Figure 1

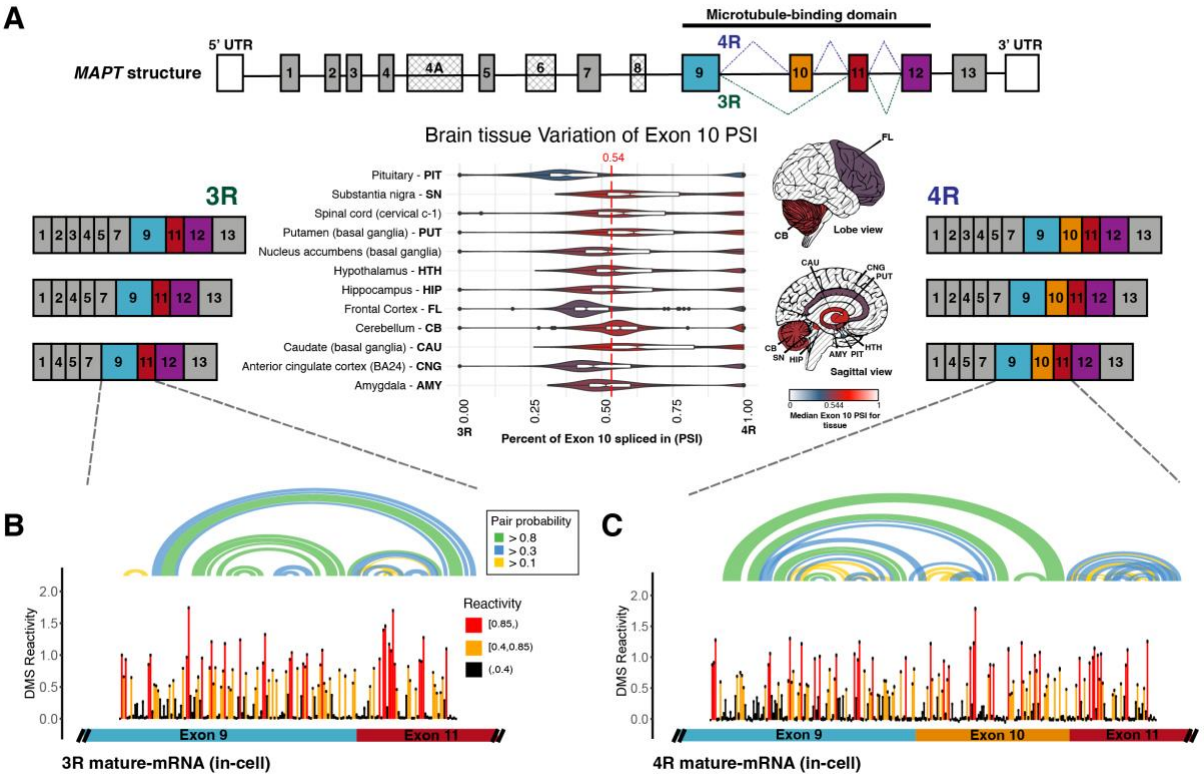

Figure 1: 3R and 4R mature *MAPT* transcripts are expressed in a 1:1 ratio in samples from human subjects and mature exons appear to fold as independent structural units.

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. This is highlighted by the alternate lines from the 5' splice site of Exon 9 to either the 3' splice site of Exon 10 (4R) or the 3' splice site of Exon 11 (3R). 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-cell DMS-MaP structure probing data across Exon 9 – Exon 11 junction of 3R mature *MAPT* transcript. T47D cells were treated with DMS. Structure probing data for junctions of interest were obtained using amplicon sequencing with region-specific 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 less structured regions, whereas low DMS reactivities correspond to more structured regions. The base pairs of the predicted secondary structure guided by DMS reactivities (using A/C nucleotides only) are shown in the arcs colored by pairing probabilities.

C) In-cell DMS-MaP structure probing data across Exon 9 – Exon 10 – Exon 11 junction of 4R mature *MAPT* transcript

MAPT **pre-mRNA Exon 10-Intron 10 junction is more structured than the mature isoforms in cells**

RNA structure around exon-intron junctions has been shown to regulate alternative splicing (Warf and Berglund 2010; Buratti and Baralle 2004), and a hairpin structure at the exon 10 – intron 10 junction of *MAPT* pre-mRNA 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). The structure of the *MAPT* pre-mRNA in the exon 10 – intron 10 junction region has been studied using biophysical techniques and chemical probing of in vitro-transcribed fragments and using computational methods (Varani et al. 1999; Lisowiec et al. 2015; Tan et al. 2019; Chen et al. 2019), but the pre-mRNA structure had not previously been analyzed in cells. We obtained DMS-MaP data for this junction from endogenous pre-mRNA in T47D cells (Figure 2A). Replicates were highly correlated (Figure 2–figure supplement 1A). Similar reactivity data were also observed in SH-SY5Y cells (Figure 2–figure supplement 1C), despite the likely differences in RBP populations compared to T47D cells (Figure 1–figure supplement 1E).

The reactivities for exon 10 in the pre-mRNA and mature 4R isoform were highly correlated (Figure 2–figure supplement 1B). This high correlation was unexpected given that the pre-mRNA undergoes splicing during the 5-minute treatment of the cells with DMS. As we observed for the mature 4R isoform, exon 10 in the pre-mRNA mostly formed base pairs with other exon 10 nucleotides (Figure 2–figure supplement 1F). However, when we compared DMS reactivities for pre-mRNA and the mature 4R isoform, we found that DMS reactivity in exon 10 was significantly lower for the pre-mRNA (median in-cell DMS reactivity: 0.08) than for the 4R isoform (median in-cell DMS reactivity: 0.22) (Figure 2–figure supplement 1D, E). This was also the case for RNA probed under cell-free conditions. The pre-mRNA is thus apparently more structured than mature mRNA independent of protein protection.

Disease mutations change the *MAPT* pre-mRNA structural ensemble and splicing of exon 10

Many RNAs adopt an ensemble of structures instead of a single structure (Halvorsen et al. 2010; Adivarahan et al. 2018). We posited that a structural ensemble near the *MAPT* exon 10 – intron 10 junction regulates exon 10 splicing and that disease-associated

mutations alter the composition of the structural ensemble to disrupt splicing regulation. We used Boltzmann sampling of RNA structures supported by DMS reactivity data (Spasic et al. 2018) (Materials and methods) to sample 1000 structures each for the wild-type. We also generated ensembles for two RNAs that bear mutations in intron 10 that are known to alter *MAPT* splicing: (i) an A to C mutation at position +15 (+15A>C) that favors 3R isoform, and (ii) a C to G mutation at position +19 (+19C>G) that favors the 4R isoform (Tan et al. 2019). These mutant ensembles were generated using the same DMS reactivities as the wild-type RNA, with the exception of the mutation site (see Materials and methods), and thus provide a well-controlled prediction of the impact that each mutation will have on the ensemble.

We visualized the structural ensemble for the 3000 structures using t-Distributed stochastic neighbor embedding (t-SNE) (Van Der Maaten and Hinton 2008) and identified five clusters (Figure 2B; Materials and methods). Each dot in the t-SNE plot (Figure 2B) corresponds to a single structure and is colored by the $\Delta G^{\ddagger}$ of unfolding (Mustoe, Busan, et al. 2018) 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 $\Delta G^{\ddagger}$ is the cost of disrupting a given structure without allowing the RNA to refold (Mustoe, Busan, et al. 2018; Mustoe, Corley, et al. 2018). We quantified and visualized the density of structures from the t-SNE plot (Figure 2C) and calculated representative structures for each cluster (Figure 2D and Figure 2–figure supplement 2B; Materials and methods). The wild-type sequence forms structures distributed across the entire space with about 70% of structures found in Clusters 2, 3, and 4 (Figure 2–figure supplement 2B). By contrast, in the +19C>G mutant that strongly favors the 3R isoform (Tan et al. 2019), more than 55% of structures belong to Cluster 1, which is defined by a fully base-paired 5' splice site (Figure 2D). Conversely, greater than 50% of structures in the ensemble of the +15 A>C mutant (Cluster 5), which shifts the isoform balance entirely to 4R (Tan et al. 2019), were characterized by lower $\Delta G^{\ddagger}$ of unfolding for the splice site region (Figure 2B, C). Correspondingly, the 5' splice site for the Cluster 5 representative structure was less structured than that of Cluster 1 (Figure 2D). Based on these results, we concluded

285 that mutations shift the structural ensemble of the MAPT exon 10 – intron 10 junction,
286 and these structural shifts correspondingly change exon 10 splicing.

Figure 2

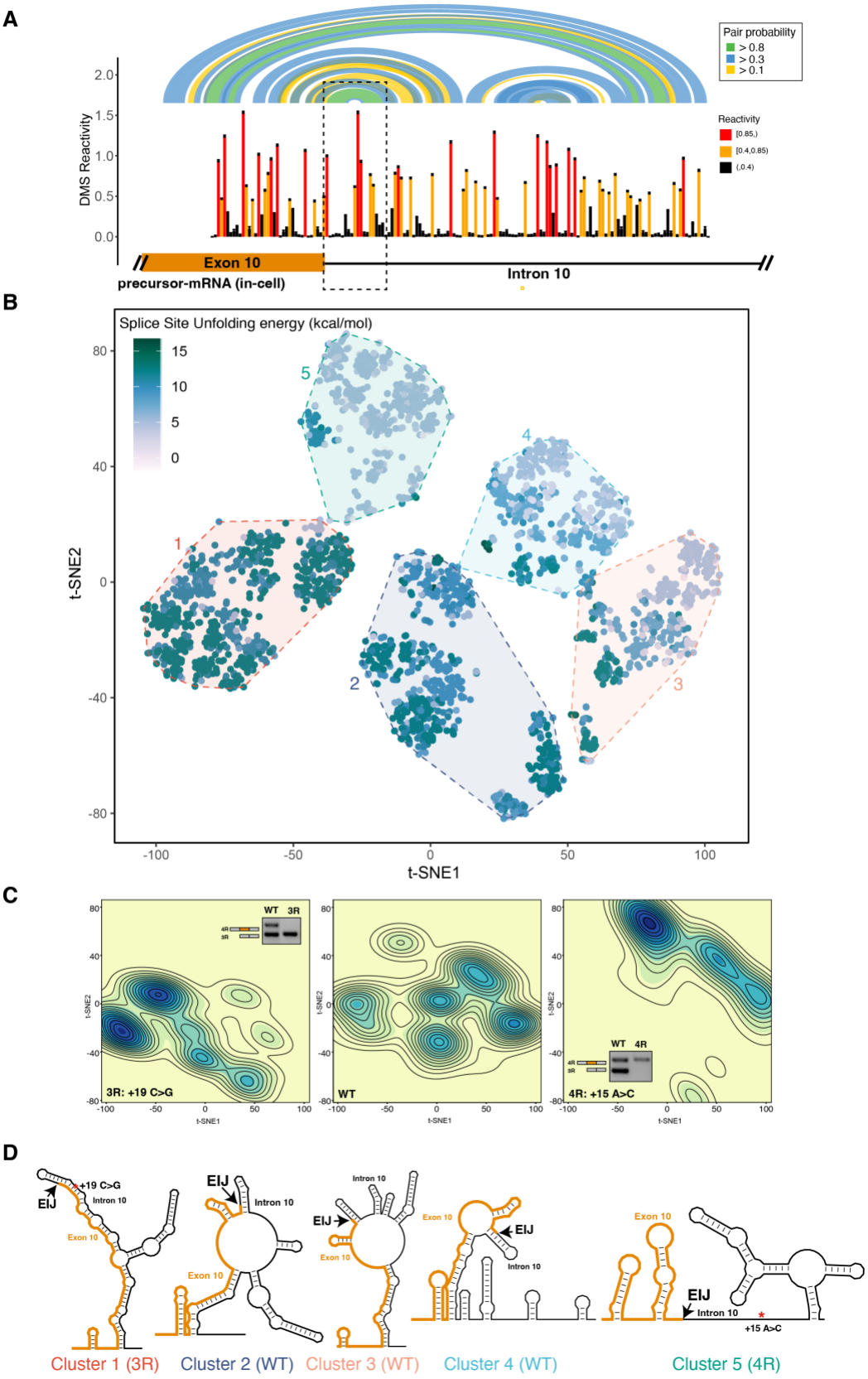
A
Pair probability
> 0.8
> 0.3
> 0.1
Reactivity
[0.85,)
[0.4,0.85)
(,0.4)
DMS Reactivity
Exon 10
Intron 10
precursor-mRNA (in-cell)
B
Splice Site Unfolding energy (kcal/mol)
t-SNE1
t-SNE2
C
WT 3R
3R: +19 C>G
WT
WT 4R
4R: +15 A>C
D
+19 C>G
EIJ
Intron 10
Exon 10
Cluster 1 (3R)
Cluster 2 (WT)
Cluster 3 (WT)
Cluster 4 (WT)
Cluster 5 (4R)
+15 A>C

Figure 2: The 4R and 3R mutations shift DMS reactivity-guided structural ensemble of Exon 10 – Intron 10 junction to be less structured and more structured, respectively.

A) In-cell DMS-MaP structure probing data across Exon 10 – Intron 10 junction of precursor *MAPT* transcript in T47D cells. Structure probing data for junctions of interest were obtained using amplicon sequencing with 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 median DMS reactivities correspond to less structured regions, whereas low median DMS reactivities correspond to more structured regions. Base pairs of predicted secondary structure guided by A/C 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 for A/C nucleotides 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 the calculated unfolding free energy of the 5' splice site at exon-intron junction (3 exonic, 6 intronic bases). Clusters have been circled and enumerated using k-means clustering with k=5.

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. Inserts are gel images from representative of splicing assays using a reporter plasmid expressing either the wild-type sequence (WT), the +19C>G (3R) mutation or +15A>C (4R) mutation in HEK293 cells, where the RNA was extracted and reverse transcribed to measure the isoform ratio using specific PCR amplification (Materials and methods). In WT, both 3R (Exon 9 – Exon 11) and 4R (Exon 9 – Exon 10 – Exon 11) isoforms are expressed (two bands). In

the presence of the 3R mutation, only the 3R isoform is expressed (one band) whereas for the 4R mutation only the 4R isoform is expressed (one band). Gel insets for the 3R and 4R mutation are in their respective density plots.

D) Representative structures for the five clusters are shown. The cluster number is indicated below each structure. 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.

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

RNA structure has been shown to control alternative splicing by regulating accessibility of key regions to spliceosome components (McManus and Graveley 2011; Warf and Berglund 2010). The 5' splice site is the minimum region of RNA that must be accessible for base pairing with the U1 snRNA (Blanchette and Chabot 1997; Singh, Singh, and Androphy 2007). In our structural ensemble analysis of the *MAPT* exon 10 – intron 10 junction (Figure 2), we found that shifts in the unfolding energy of the 5' splice site in wild-type and mutant pre-mRNAs corresponded to changes in exon 10 inclusion levels. 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). Hence, we hypothesized that more than just the 5' splice site nucleotides might need to unpair to facilitate the splicing reaction. We analyzed high-resolution cryo-EM structures of the human spliceosome at Pre-B (PDB ID: 6QX9), B (PDB ID: 5O9Z), Pre-B^{act} (PDB ID: 7ABF), and B^{act} (PDB ID: 5Z56) stages (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 associated with the spliceosome (Materials and methods). The number of pre-mRNA nucleotides, as observed in each structure, increased through the splicing cycle (Figure 3A).

To identify the spliceosome complex footprint that best predicts splicing outcome, we examined the relationship between unfolding energy and splicing outcome for 20 synonymous or intronic mutations in exon 10 and intron 10 (Figure 3–figure supplement 1A). These mutations are more likely to affect splicing (Supek et al. 2014; H. Lin et al. 2019) and structure (Sharma et al. 2019; C. L. Lin, Taggart, and Fairbrother 2016) than mutations that alter the protein sequence. The distribution of $\Delta G^{\ddagger}$ of unfolding of the 5' splice site in the presence of these mutations was correlated with exon 10 PSI (Figure 3–figure supplement 1B). We then calculated the $\Delta G^{\ddagger}$ of unfolding of the RNA for regions overlapping the 5' splice site that correspond to the footprints of each of the four spliceosome intermediates. Features of the unfolding $\Delta G^{\ddagger}$ distribution, including mean and standard deviation, were then used in a β-regression to predict exon 10 PSI

(Materials and methods; Eq. 2). Unfolding larger regions around the 5' splice site improved the predictive power of the model, and the B^{act} complex footprint yielded the best prediction accuracy (R^2 = 0.89; Figure 3B). Crucially, we found that using features of the distribution of unfolding $\Delta G^{\ddagger}$ in the structural ensemble produced greater predictive power than simply using the unfolding $\Delta G^{\ddagger}$ of a single minimum free energy structure, supporting the importance of RNA ensemble behavior to splicing outcome (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). Synonymous mutations that alter exon 10 inclusion lie a mean distance of 54 nucleotides from the exon-intron junction, whereas those in the intron are a mean of 14 nucleotides from the junction. The variation in bootstrapped correlation coefficients decreased as a larger region around the exon-intron junction was unfolded, suggesting that the synonymous mutations affect distal structures.

We then tested the structural ensemble-based model on an additional 24 non-synonymous and compensatory mutations found in exon 10 and intron 10. Compensatory mutations are double mutations that were designed to rescue changes in exon 10 splicing caused by a single mutation (Grover et al. 1999). 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 clear limitation of this structure-only model is that it does not account for the effects of mutations on potential splicing regulatory elements (SREs) in the sequence, which are also known to control alternative splicing (Z. Wang and Burge 2008).

Figure 3

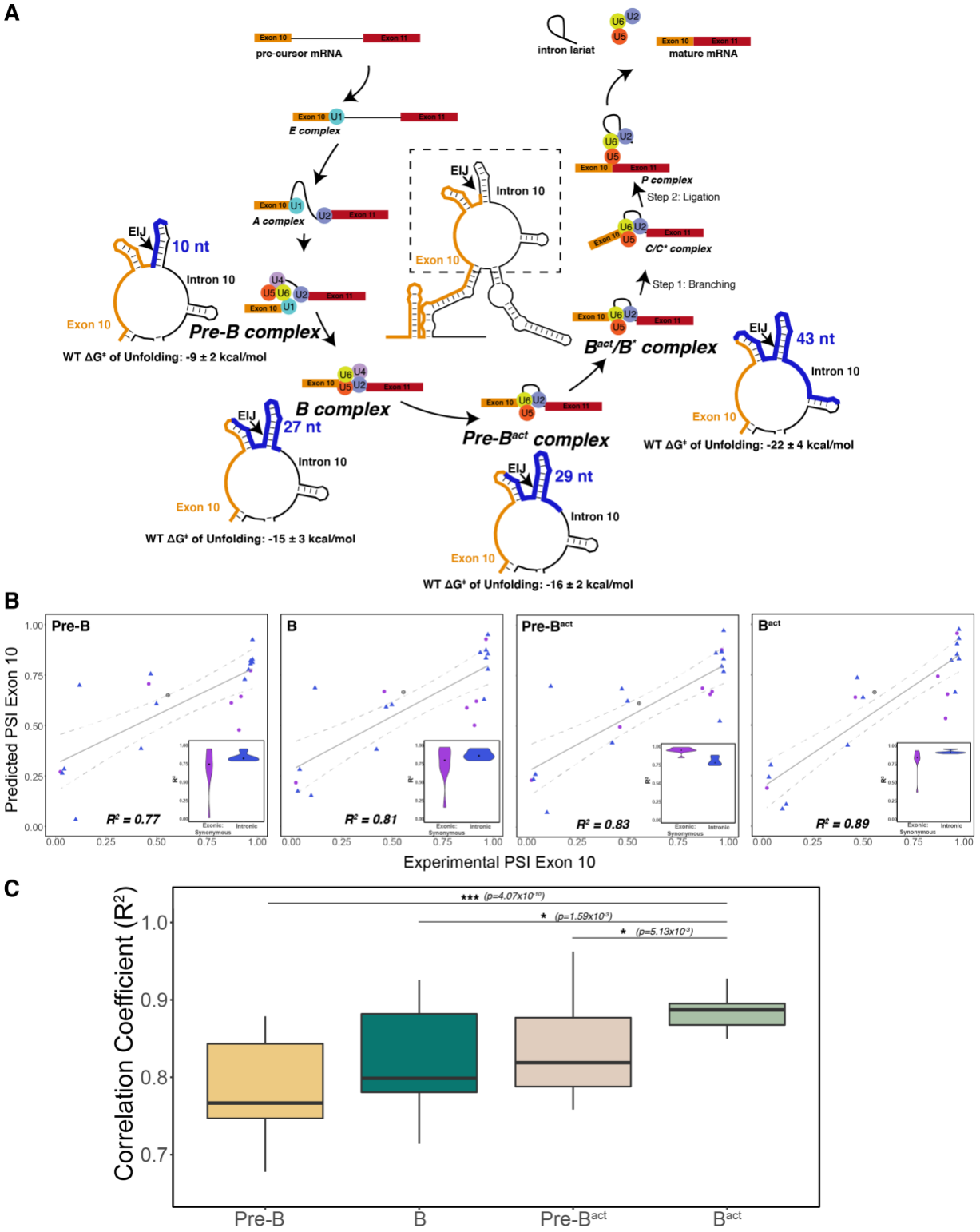

Figure 3: The best predictor of Exon 10 PSI for intronic and synonymous mutations was the unfolding free energy of pre-mRNA during the B^{act} stage of splicing

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. 2. 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. A two-tailed Wilcoxon Rank Sum test was used to determine significance between B^{act} complex and the other

three complexes. Level of significance: ***p-value < 10^{-6}, **p-value < 0.001, * p-value < 0.01

Consideration of motif strengths of splicing regulatory elements improves prediction of Exon 10 PSI for non-synonymous mutations

Exon 10 splicing is highly regulated by differential binding of RBPs to *cis*-SREs within exon 10 and intron 10 (Qian and Liu 2014). The expression patterns of RBPs known to

bind *MAPT* pre-mRNA vary across tissues and individuals (Figure 1–figure supplement 1E) and are not predictive of exon 10 PSI. Additionally, while our structure-only model performs moderately well for 47 mutations (R^2=0.74) (see Supplementary file 2 for further details about mutations), the structure only model performs particularly poorly for non-synonymous mutations (median bootstrapped R^2 = -0.21, Figure 4-figure supplement 1B). Hence, we hypothesized that consideration of mutation-induced changes in binding of SREs might improve our model. 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), and calculated changes to splice site, enhancer, and silencer motif strengths due to mutations (Figure 4A; Materials and methods). We found that using splice site strength as the sole predictor yielded poor prediction of exon 10 PSI for all mutation categories (Figure 4B; Eq. 4). There was a weak positive correlation between PSI and enhancer strength and a significant negative correlation between PSI and silencer strength (Figure 4A and Figure 4–figure supplement 1B). When exon 10 PSI was modeled with the changes to the motif strength of all splicing regulatory elements, prediction accuracy increased (R^2=0.51; Figure 4C) compared with that obtained when only splice site strength was considered (R^2=0.29); for non-synonymous mutations accuracy was even higher (R^2=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; 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 focusing on binding motifs for these proteins would improve model accuracy, we identified RBP sites based on previous data 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 changes due to *MAPT* mutations and exon 10 PSI (Figure 4–figure supplement 2A, B). Model prediction accuracy was lower (R^2=0.08, Figure 4–figure supplement 2C) than when predictions considered general SRE motifs. Thus, going forward we chose to use SRE motifs for our combined models.

499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523

Figure 4

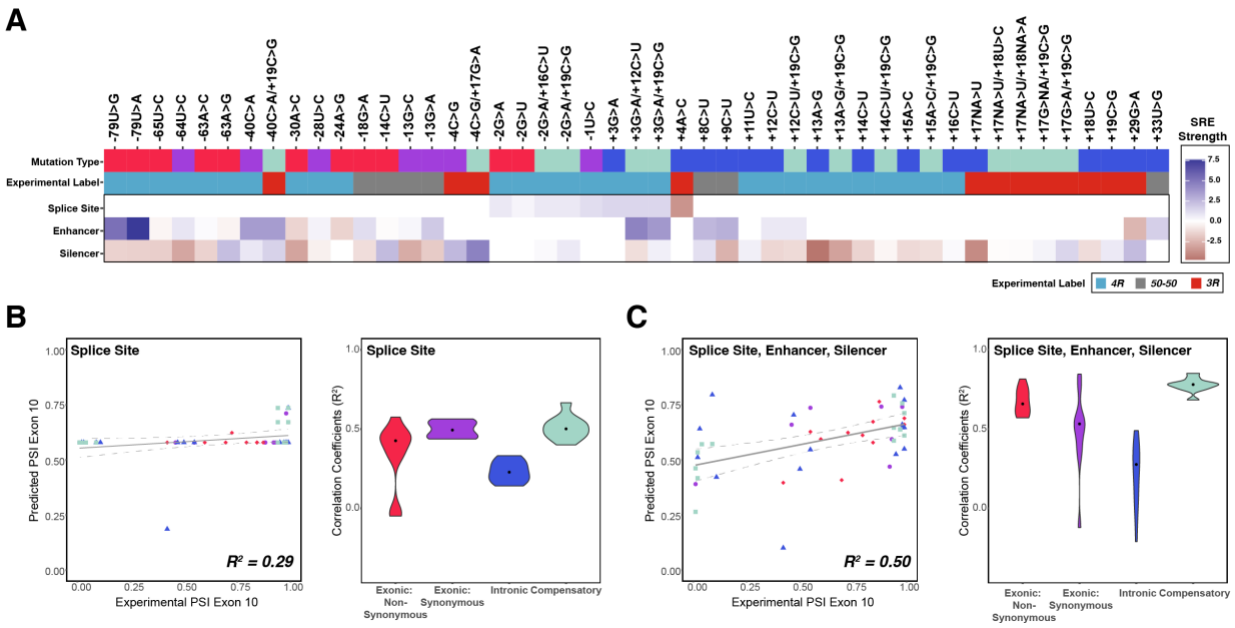

Figure 4: Combining the strength of all splicing regulatory elements significantly improves prediction of Exon 10 PSI compared to using only splice site strength

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). Mutation Type refers to whether the mutation is exonic non-synonymous, exonic synonymous, intronic or compensatory. Experimental Label is the label given by the original study that experimentally validated each mutation using a splicing assay.

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. 4. 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. 5.

Model with both RNA structure and SRE motif changes yields best prediction of exon 10 PSI

We next set out to determine if combining both structural and SRE features further improved prediction. Indeed, a 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 (R^2 = 0.89; Figure 5A, B). An alternative additive model had lower prediction accuracy (R^2= 0.80) (Figure 5-figure supplement 1A), particularly for non-synonymous mutations (Figure 5-figure supplement 1B. This suggests that considering the category of mutation is critical in accurately modeling the effects on splicing.

To determine whether structure or SRE changes were responsible for the splicing phenotype of each individual mutant, we hierarchically clustered the four primary features (structure around 5' splice site, 5' splice site strength, enhancer strength, silencer strength) for the 47 mutants that have been experimentally characterized (Materials and methods). Six categories emerged from the clustering of features (Figure 5C, and Figure 5–figure supplement 1C). For about 51% of mutations, both structure and SRE motif strength were altered in the same direction to either promote or inhibit exon 10 inclusion (Figure 5D). For the remaining mutations, structure and SRE strength changed in opposite directions. For 17% of mutants, structure dominated the direction of splicing. For about 23%, SRE strength was dominant (Figure 5D). Overall, these results support the conclusion that structure and SREs have equally important effects on regulation of splicing at this exon-intron junction.

Figure 5

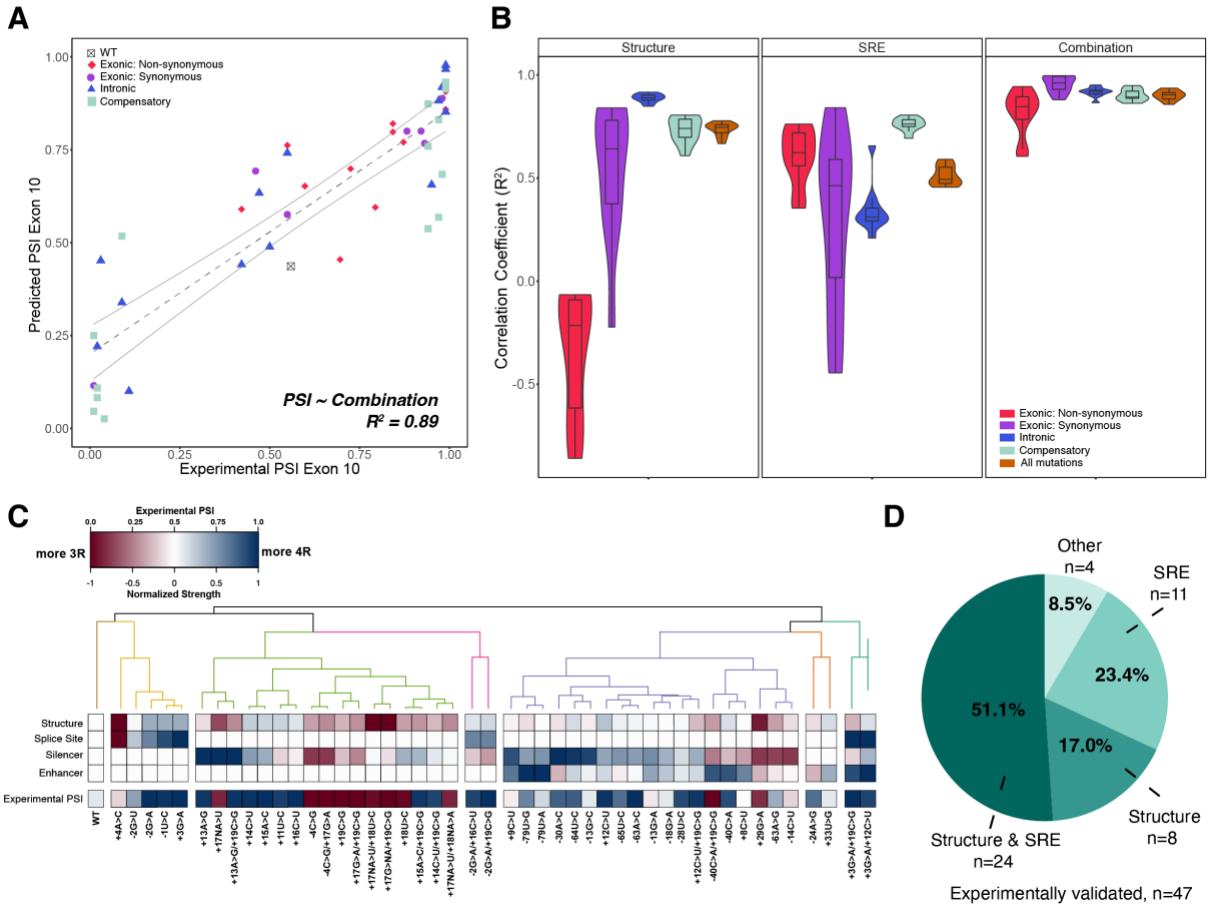

Figure 5. Combining structure and SRE strength into a unified model best predicts exon 10 PSI

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. 7. 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. Combination 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 predicts 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 in the heatmap in Figure 5C. 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

We next interrogated the model by performing a systematic in silico mutagenesis analysis of the 100 nucleotides spanning the exon 10 – intron 10 junction (Figure 6A). Our model predicted that most mutations should result in inclusion of exon 10. This bias is consistent with the observation that about 75% of known disease-associated mutations in this region induce exon 10 inclusion (Figure 6B and Figure 6–figure supplement 1A). We found that a significantly greater proportion of disease-inducing mutations (76.4%) result in changes to both structure and SRE compared with uncategorized mutations (36.0%) (Figure 6C). Thus, mutations that alter both structure and SREs have a greater likelihood of causing disease than mutations that alter only structure or only an SRE. Intriguingly, mutations overall caused a slight skew toward a more structured exon-intron junction that would be expected to decrease inclusion of exon 10 (Figure 6A, Figure 6–figure supplement 1B); however, these same mutations altered SRE strength in a manner that skewed toward increased inclusion of Exon 10 (Figure 6–figure supplement 1C), indicating that SREs act to counter the effect of structural changes. Our modeling suggests that a complex balance of structure and RBP binding results in the observed 1:1 ratio of the 3R to 4R *MAPT* 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). VUSs 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 used splicing assays to experimentally determine the splicing preference of six instructive variants (Materials and methods): 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 that all six predictions were correct (Figure 6E, Figure 6-figure supplement 1D). The three 3R VUSs caused the region around the exon-intron junction to become more structured while the 4R VUS made this region less structured compared to WT (Figure 6-figure supplement 1E). SRE strength changes correctly predict Exon 10 splicing direction for +30U>C and -6G>A (Figure 6-figure supplement

1F). For +23U>C and +26G>A, we observed changes in the degree of structure around the exon-intron junction and silencer strengths in diverging directions (Figure 6-figure supplement 1E, F) suggesting that these opposing changes preserve the WT 3R/4R ratio.

Figure 6

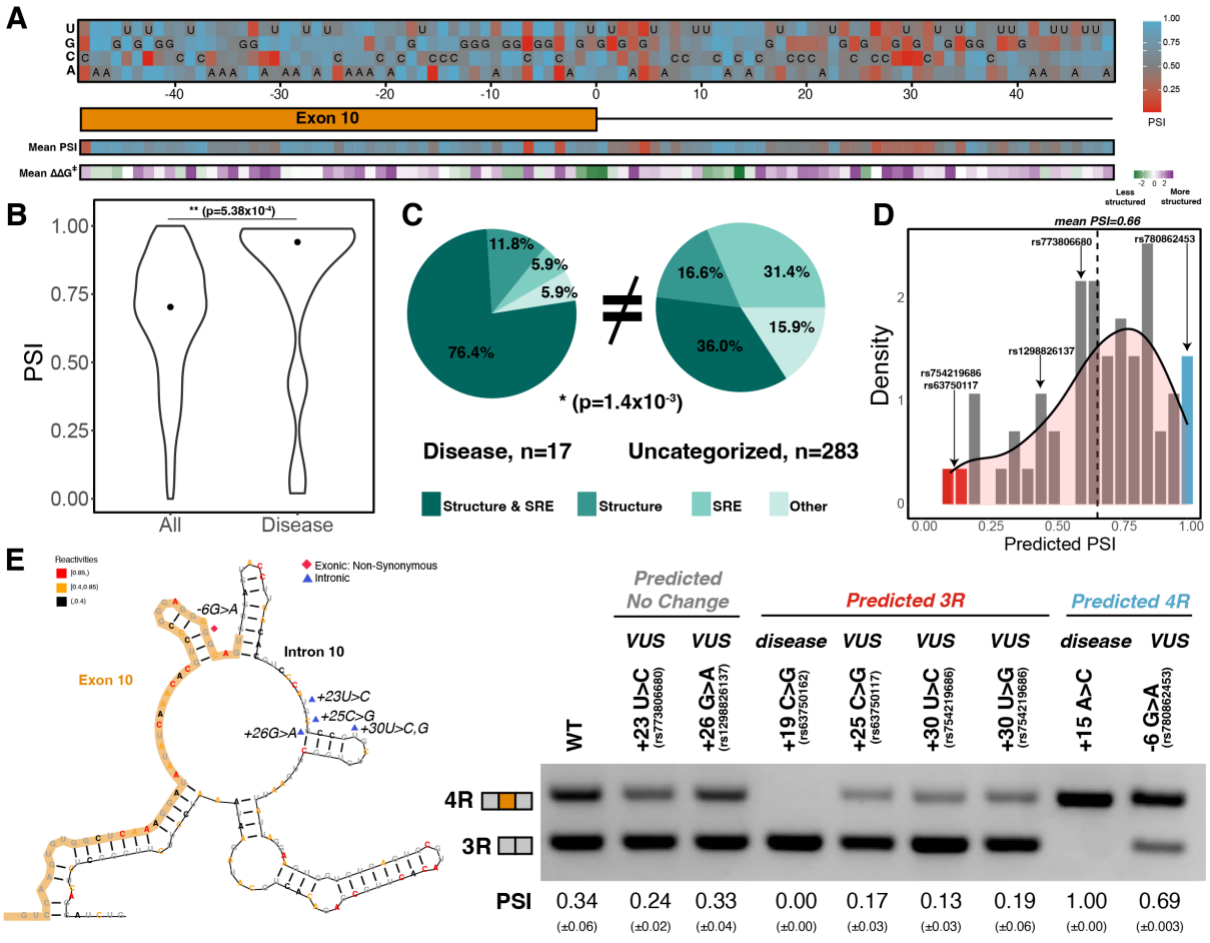

Figure 6. Combined model is predictive of exon 10 inclusion ratios for previously uncharacterized mutations

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 as shown in Figure 5A and then 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). A two-tailed Wilcoxon Rank Sum test was used to determine significance between the two categories. Level of significance: ***p-value < 10^{-6}, **p-value < 0.001, * p-value < 0.01

C) Pie chart showing features that drive Exon 10 splicing for disease and presently uncategorized 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 uncategorized mutations. The difference in proportions was tested with a one-tailed 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

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. The *MAPT* exon 10 – intron 10 junction is a well-studied example of the effect of 5' splice site secondary structure on splicing regulation. A hairpin was initially hypothesized to play a major role in splice site accessibility because disease mutations in this structure, close to the exon-intron junction, shifted the isoform balance to completely exclude or completely include exon 10 in the mature mRNA (Hutton et al. 1998; Grover et al. 1999). NMR, cell-free chemical probing, and computation analyses confirmed the presence of the hairpin (Varani et al. 1999; Chen et al. 2019; Lisowiec et al. 2015). Recent studies have shown that structures determined in cell-free conditions can differ dramatically from those in cells (Sun et al. 2019; Rouskin et al. 2014). Our results suggest that this is not the case for the exon 10 – intron 10 junction region: In-cell chemical probing of the endogenous *MAPT* pre-mRNA provided strong evidence for formation of this hairpin in cells and for structural features not previously captured.

Our analysis also revealed that in cells, exonic regions were less structured than introns, as also observed by Sun et al (Sun et al. 2019). Mature *MAPT* 3R and 4R are less structured in the region of exon 9 through exon 11 than is the pre-mRNA. The high correlations between structures of the exon in different *MAPT* isoforms and our finding that predicted exon 10 folding is only slightly impacted by the presence of intron 10 or exon 11 residues agrees with previous observations of mRNAs. Specifically mRNAs

749 which encode yeast ribosomal proteins that indicate that RNA folding in both pre- and
750 post-spliced exons is highly local and that most base-pairs are intra-exon (Zubradt et al.
751 2016).
752
753 Unlike non-coding RNAs such as the ribosome and tRNA that rely on folding to a single,
754 well-defined structure (Petrov et al. 2014), most RNAs are dynamic, unfolding and
755 refolding within a landscape (Cruz and Westhof 2009; Giegé et al. 2012). We showed
756 that structural ensembles have an important function at the Exon 10 – Intron 10
757 junction. If the 5' splice site was always paired, only the 3R isoform would be produced.
758 However, the presence of 3R and 4R isoforms, usually in a 1:1 ratio implies that the
759 junction is accessible in a subset of the structures. We found disease-causing mutations
760 produced distinct shifts in the ensemble of the *MAPT* exon 10 – intron 10 junction
761 region; these shifts showed strong correlation with changes in the 3R to 4R isoform ratio
762 and confirmed that ensembles are essential at this junction. Our ability to accurately
763 predict the effects of mutations on ensembles significantly improved our quantitative
764 model (Figure 3 - figure supplement 1C).
765
766 The U1 snRNA base pairs with a nine nucleotide sequence around the exon-intron
767 junction (Roca et al. 2012). However, our analysis of cryo-EM structures of the human
768 spliceosomal assembly cycle revealed that a larger region of the pre-mRNA interacts
769 with the spliceosome and must be unfolded during splicing. Our structural model
770 performed most accurately when we required 43 nucleotides around the 5' exon-intron
771 junction to be unfolded, corresponding to the region within the spliceosome B^{act}
772 complex. This observation suggests that a large region of the pre-mRNA is dynamically
773 remodeled by the spliceosome, and that structures distal to the exon-intron junction can
774 regulate splicing. Our finding corroborates evidence that RNA structure near this exon-
775 intron junction is extensive (Tan et al. 2019). Note that we do not claim that all 43
776 nucleotides need to remain fully unpaired during the splicing cycle, as the entire cycle is
777 dynamic and likely involves other intermediate structures. Rather, our model argues that
778 mRNA unfolding and accommodation into the B^{act} complex is a key rate limiting step in
779 splicing, and considering this step is necessary to accurately model splicing outcome for

780 a diverse set of mutations. Broadly, our definition of a functional footprint for splicing
781 parallels a similar idea for translation initiation by the ribosome (Corley et al. 2017;
782 Mustoe, Busan, et al. 2018; Mustoe, Corley, et al. 2018), for which the footprint is
783 roughly 30 nucleotides. Thus, for both translation initiation and splice site selection,
784 there is a region in which RNA structure functions as a rheostat.
785
786 Considerable evidence supports a function for both splicing regulatory elements (and
787 their corresponding RBPs) and RNA structure in controlling alternative splicing of exon
788 10 of *MAPT* (Andreadis 2012). However, the relative importance of these two factors
789 has been controversial. The regression model we developed clarifies that there is a
790 cooperative relationship between RNA structure and SREs in driving splicing outcome.
791 Exonic non-synonymous mutations promote splicing changes primarily by altering SRE
792 motifs, whereas exonic synonymous and intronic mutations altered RNA structure. A
793 combined model that accounted for both structure and SREs was the most accurate
794 predictor of exon 10 PSI (Figure 5D). It was previously proposed that exon 10 is
795 alternatively spliced due to a weak 5' splice site (Ian D'Souza and Schellenberg 2005),
796 and, indeed, we found that mutations that strengthened the splice site increase
797 inclusion of exon 10 (Figure 4A). SRE strength alterations overall skewed more toward
798 increased exon 10 inclusion, which suggest that SREs and the RBPs that bind them
799 buffer the effects of RNA structure to maintain the 1:1 isoform ratio.
800
801 Although structure and SREs had opposite effects on splicing outcomes, disease
802 variants often resulted in a synergistic effect on splicing outcome. The combined model
803 was directly validated by accurate prediction of the effects of six previously untested
804 VUSs on exon 10 splicing (Figure 6E). Few VUSs were predicted to completely exclude
805 exon 10 from the mature mRNA: Only five VUSs had PSIs less than 0.25. Our model
806 accurately predicted the effect of the three with the lowest predicted PSI. Our
807 systematic computational mutagenesis revealed a hotspot for mutations around 25-30
808 nucleotides downstream of the exon-intron junction that were predicted to result in
809 production of only the 3R isoform (Figure 6A). Indeed, the experimentally validated
810 VUSs with PSIs less than 0.25 were in this region.

In principle, our splicing model can be extended to other exon-intron junctions, although RBPs that recognize SRE motifs have different binding contexts (Dominguez et al. 2018) and the exact binding preferences of the RBPs that regulate the junction of interest are currently unknown. Another limitation is that the current model does not consider structural and sequence features around the 3' splice site (in the case of *MAPT* exon 10, the intron 9 – exon 10 junction), that are expected to impact exon 10 splicing regulation. Although our model provides an exact PSI prediction for each mutation, we emphasize that its principal utility is in predicting the direction in which the 3R to 4R isoform ratio shifted from the wild-type ratio.

In brain tissue from healthy individuals, exon 10 PSIs varied 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 was variability in exon 10 PSIs in different brain tissues (Majounie et al. 2013). Thus, although our model combines both structural and sequence features to achieve quantitative prediction accuracy of the 3R to 4R ratio for a wide range of disease mutations (synonymous, non-synonymous, intronic and exonic), it is not clear that PSI alone is predictive of severity of disease for the broad class of tauopathies (Majounie et al. 2013). Disease severity is compounded by other factors including gene-gene interactions and environmental factors. As such, the value of our model stems more from how it incorporates RNA structure in predicting alternative splicing, rather than as a direct predictor of disease severity. 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 this region of the pre-mRNA 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

Analyses of *MAPT* sequencing data for GTEx tissue types

Aligned BAM files of individual samples from the GTEx v8 project for tissue types with *MAPT* transcripts per million (TPM) greater than 10, were accessed in the AnVIL/Terra environment (Kumar 2020a). Reads aligning to *MAPT* were extracted (Kumar 2020b) and downloaded. Exons 2, 4, and 10 PSIs were quantified per BAM file with Outrigger (Song et al. 2017) using the *MAPT* transcriptome reference from Ensembl GRCh38. Only samples with at least 10 reads mapping across the exon-intron junction of interest were considered. For exon 10 PSI, 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 of exon 10 PSI between individuals versus within an individual (for individuals with *MAPT* expression in more than seven tissues) (Supplementary file 1). TPMs for RBPs known to affect the splicing regulation of *MAPT* exon 10 were extracted, and their distributions in brain tissues were plotted using ggplot2.

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-cell DMS-MaP probing of *MAPT* RNA

Approximately 20 million T47D cells and 30 million SHSY-5Y cells were harvested by centrifugation and resuspended in 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 as previously

described (Mustoe et al. 2019). For the negative control (unmodified RNA) ethanol, instead of DMS, was added to cells. After incubation, the reactions were neutralized by addition of equal volume of ice cold 20% β-mercaptoethanol. Total RNA was extracted using Trizol (ThermoFisher Scientific), treated with TURBODNase (ThermoFisher Scientific), purified using Purelink RNA mini kit (ThermoFisher Scientific), and quantified based on absorbance determined with a NanoDrop spectrophotometer.

Cell-free DMS-MaP probing for *MAPT* RNA

Approximately 10 million T47D cells in 10 cm plates were used. Growth media was removed, following which cells were trypsinzed (Tryple, ThermoFisher Scientific) and the pellet was washed with PBS. Total RNA was extracted by Trizol (ThermoFisher Scientific), chloroform and isoamyl alcohol (24:1, Sigma-Aldrich) using phase lock heavy tubes (5PRIME Phase Lock Gel) followed by Purelink RNA mini kit purification (ThermoFisher Scientific) and on-column DNase digestion (PureLink DNase, ThermoFisher Scientific). RNA was quantified by NanoDrop™ spectrophotometer. 10 ug of RNA was resuspended in 90 uL of bicine buffer (200 mM Bicine pH 8, 100 mM NaCl and 10 mM MgCl2) with 20 U of RNase inhibitor (NEB) and incubated at 37°C for 10 minutes. Samples were treated with 10 uL of DMS diluted in ethanol (1:10) for 5 min at 37°C. For the negative control (unmodified RNA), instead of DMS, an equivalent amount of ethanol was added to the extracted RNA. After incubation, all reactions were neutralized by addition of 100 uL of ice cold 20% by volume β-mercaptoethanol and kept on ice for 5 minutes. Reaction by-products were removed by RNA purification with the Purelink RNA mini kit (ThermoFisher Scientific) before error-prone reverse transcription.

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

Purified RNA (9 μg) was reverse transcribed using Random Primer 9 (NEB) and SuperScript II reverse transcriptase under MaP conditions as described previously (Smola et al., 2015). A no-reverse transcriptase control was also prepared. The resultant cDNA was purified over a 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 with the NEB Q5 HotStart polymerase. Secondary PCR was performed to introduce TrueSeq barcodes (Smola et al. 2015). All samples were purified using the Ampure XP beads (Beckman Coulter), 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 paired-end 2×151 and 2×301 read multiplex runs on MiSeq platform (Illumina) for pre-mRNA and mature mRNA, respectively. Sequenced reads have been uploaded to the NCBI SRA database under BioProject ID PRJNA762079 for in-cell data and PRJNA812003 for cell-free data.

In-cell DMS-MaP probing of SSU

For in-cell rRNA structure data, approximately 10 million T47D cells were used for each condition. Growth media was removed, followed by addition of 1.8 mL of 200 mM bicine, pH 8.3 and treatment at 37 °C with 200 μL of DMS diluted in ethanol (1.25% final DMS) for 5 min. For the negative control ethanol was added instead of DMS. After incubation, all reactions were neutralized by addition of equal volume ice cold 20% β-mercaptoethanol and kept on ice for 5 min. Total RNA was extracted using Trizol (ThermoFisher Scientific) and chloroform and isoamyl alcohol using phase lock heavy tubes (5PRIME Phase Lock Gel). RNA was purified using a Purelink RNA mini kit (ThermoFisher Scientific), treated with TURBODNase (ThermoFisher Scientific), and quantified.

Cell-free DMS-MaP probing of SSU

Approximately 10 million T47D cells were trypsinzed (Tryple, ThermoFisher Scientific), and the pellet was washed with PBS. Total RNA was extracted using Trizol (ThermoFisher Scientific) and chloroform and isoamyl alcohol (24:1, Sigma-Aldrich) using phase lock heavy tubes (5PRIME Phase Lock Gel) followed by purification using a Purelink RNA mini kit purification (ThermoFisher Scientific) and on-column DNase digestion (PureLink DNase, ThermoFisher Scientific). RNA was quantified based on absorbance determined using NanoDrop spectrophotometer. For each sample, 10 μg of RNA was resuspended in 90 μL of 200 mM bicine, pH 8, 100 mM NaCl, and 10 mM

$MgCl_2$ with 20 U of RNase inhibitor (NEB) and incubated at 37 °C for 10 min. Samples were treated with 10 μL of DMS diluted in ethanol (1:10) for 5 min at 37 °C. For the negative control, samples were treated with ethanol. After incubation, all reactions were neutralized by addition of 100 μL of ice cold 20% β-mercaptoethanol and kept on ice for 5 min. Reaction by-products were removed using a Purelink RNA mini kit (ThermoFisher Scientific) before error-prone reverse transcription.

DMS-MaP cDNA synthesis, library construction, and sequencing of SSU

Purified RNA was reverse transcribed using Random Primer 9 (NEB) and SuperScript II reverse transcriptase under error prone conditions (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). A standard Nextera DNA library protocol (Illumina) was used to fragment the cDNA and add sequencing barcodes. Samples were purified using Ampure XP beads (Beckman Coulter), 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. Gel purification (GeneJET, ThermoFisher Scientific) was performed as needed to remove primer dimer bands from libraries before sequencing. 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 BioProject ID PRJNA762079 for in-cell data and PRJNA812003 for cell-free data.

DMS-MaP analysis

Sequenced reads were analyzed using the ShapeMapper pipeline (Busan and Weeks 2018), version (v2.1.4). 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 rRNA sequence using Bowtie2 parameters from Busan and Weeks (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.

Per-nucleotide mutation rates were obtained from the profile file output by ShapeMapper. Raw DMS reactivities are computed as:

$$R_i = mutr_S - mutr_U$$

where $mutr_S$ is the mutation rate in the sample treated with DMS, $mutr_U$ is the mutation rate in the untreated control. Raw reactivities were then normalized within a sample and per nucleotide type (A, C, U, G). For each nucleotide type, reactivity rates were normalized by dividing the mean reactivity of the top 10% of reactivities after the most reactive 2% were removed (Busan and Weeks, 2018). The Gs and Us were further normalized to have a maximum value of 0.1 to minimize their impact on plotting reactivities. Though only As and Cs were only used in structure modeling, all 4 nucleotide types were plotted in the reactivity figures and reported in the supplementary material.

DMS-reactivity guided structure prediction

Previous versions of Rsample only utilized SHAPE parameters for calculation of the partition function. In order to use DMS data to guide secondary structure modeling by Rsample (Spasic et al. 2018)(Reuter and Mathews 2010), we used our SSU data to calibrate the expected relationship between DMS reactivity and base-pairing status. Since DMS primarily reacts with As and Cs, we only used reactivity data for these two nucleotides in all of our structure modeling. Using the SSU in-cell data and the known secondary structure (Petrov et al. 2014), we determined distributions for DMS reactivities for unpaired nucleotides, nucleotides paired at helix ends, and nucleotides paired in base pairs stacked between two other base pairs, which provide the sampling distributions needed for Rsample calculations (Spasic et al. 2018). These DMS data can be invoked by Rsample by the "--DMS" command line switch as part of RNAstructure 6.4 or later. The distributions 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

995 secondary structure prediction. The “--max 5” parameter is used with Rsample to apply
996 this limitation.
997
998 **Base-pairing probabilities for SSU**
999 The partition function for the SSU was generated using Rsample, using either the
1000 sequence or using the sequence and the DMS reactivities. All possible *i-j* base pairing
1001 probabilities were summed for each nucleotide *i* to generate a base pairing probability
1002 per nucleotide *i*.
1003
1004 **ROC curves for predicting SSU base pairs**
1005 Using the known secondary structure of the SSU, we assigned a nucleotide as either 0
1006 or 1 if it was paired or unpaired. Only As and Cs were considered. DMS reactivities
1007 were used to predict whether a nucleotide was paired; the higher the DMS reactivity, the
1008 more likely a nucleotide is unpaired. ROC curves and AUC values were generated
1009 using the plotROC (Sachs 2017) R package.
1010
1011 **Minimum free energy and base-pairing probability modeling**
1012 Minimum free energy and base-pairing probability “arc” plots were generated using
1013 Superfold (Siegfried et al. 2014; Smola et al. 2015) modified to process DMS reactivity
1014 data. The original Superfold function used SHAPE parameters (Deigan et al. 2009) to
1015 fold an RNA sequence using the Fold and Partition functions of the RNAstructure
1016 package. In our modified version of Superfold, base pairing probabilities were computed
1017 using Rsample with DMS-specific parameters for A and C nucleotides. MFE structures
1018 were computed using Fold from RNAstructure with DMS reactivities input using the “--
1019 DMS” option. For structure modeling applications, G and U reactivites were set to -999
1020 (no data).
1021
1022 **In silico co-transcriptional folding of exon 10 – exon 11 and exon 10 – intron 10**
1023 **regions**
1024 We folded exon 10 using the modified Superfold function as described above, inputting
1025 a truncated DMS reactivity map file containing just reactivities of exon 10. We then

added nucleotides from exon 11 or intron 10 one at a time and re-ran Superfold after each addition inputting the DMS reactivity map file modified to only the folded nucleotides. At every additional nucleotide, we calculated the number of base pairs within exon 10 and the total number of base pairs for the current sequence. We plotted the percentage of intra-exon/intron base pairs of total number of base pairs at every additional nucleotide.

Generating a structural ensemble of the exon 10 – intron 10 region of *MAPT*

The partition function of the exon 10 – intron 10 region of *MAPT* for wild type and mutants was calculated with DMS reactivities from the wild-type pre-mRNA as restraints using Rsample (Spasic et al. 2018). For modeling mutant sequences, DMS reactivities collected for the WT sequence were used to restrain the structural space with the reactivity at each mutation site set to -999. The program stochastic (Reuter and Mathews 2010) was used to sample 1000 structures 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

For each sequence, the 1000 structures in CT format for each ensemble 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 10 – intron 11 region. 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 $\Delta G^{\ddagger}$ 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.

Identification of representative structures for clusters

The 3000×2 matrix obtained after 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. Boundaries for each cluster were marked and colored using the ggscatter function in the R ggpubr package. 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 nucleotide. The structure in the ensemble that was most similar to the RNA structure with the most common subtypes at each position, was chosen as representative of that cluster.

Visualizing density of structures in t-SNE plot

To evaluate density of structures in clusters, a meshgrid was created for the three matrices corresponding to WT, 3R and 4R mutant structures using the meshgrid function of NumPy (Harris et al. 2020) with a 1000-point interpolation, which returns two-dimensional arrays that represent all the possible x-y coordinates for the three matrices. A Gaussian kernel was 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 upstream and downstream of the 5' exon-intron junction. The result was visually confirmed by visualizing the PDB on PyMol.

Calculating $\Delta G^{\ddagger}$ of unfolding of a region of interest

The $\Delta G^{\ddagger}$ of unfolding energies of regions of interest were calculated using a custom Python script. The non-equilibrium unfolding energy of the region, defined as the energy require to unfold a specific region without allowing refolding (Mustoe, Busan, et al. 2018) is defined as follows:

$$\Delta G^{\ddagger} = \Delta G^{fold} - \Delta G^{unfold} \qquad [\mathbf{1}]$$

The ΔG of the original folded structure (ΔG^{fold}) was calculated with the efn2 program in RNAstructure (Reuter and Mathews 2010). Next, the base pairs within a region of interest were made single stranded by setting the base pair column value to be 0 in the CT file. From this modified CT file, we evaluate the ΔG of the unfolded structure (ΔG^{unfold}) with efn2. This was done for every suboptimal structure in the Boltzmann ensemble. For example, to determine the $\Delta G^{\ddagger}$ of unfolding of the splice site, we made all nucleotides within the last 3 nucleotides of the exon and the first 6 nucleotides of the intron single stranded.

Calculating changes in strength of splice site and SRE motifs

The strength of the WT splice site was calculated with MaxEntScan (Yeo and Burge 2004). Strength was recalculated if mutations were located in the last 3 bases of exon 10 or first 6 bases of intron 10. WT strength was subtracted from the mutant strength. A value of 0 implied no change in splice site strength, positive values implied that a mutation made the 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.

Overrepresented hexamers in cell-based screens of general exonic and intronic splicing enhancers (ESEs and ISEs) and silencers (ESSs and ISSs) were obtained from previous reports (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, calculated as described (Fairbrother et al., 2002), are collated in Supplementary file 5. There were eight clusters of ESE motifs, seven of ESS motifs,

seven of ISE motifs, and eight 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. The only windows that differed between the WT and mutants 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 ∆strength for every PWM per mutation. The average of the non-zero ∆strengths was calculated for ESE, ESS, ISE, and ISS categories. The ESE and ISE ∆strengths were summed to obtain an enhancer strength, and the ESS and ISS ∆strengths were summed to obtain a silencer strength. Supplementary file 6 presents all SRE ∆strengths for the 47 mutations and 55 VUSs.

Calculating the change in strength of RBP motifs

Previously determined position frequency matrices for SRSF1, SRSF2, SRSF7, SRSF9, SRSF10, PCBP2, RBM4, and SFPQ (Ray et al. 2013) were converted into PWMs by normalizing frequencies to 0.25 (the prior probability for nucleotide frequency) and calculating the log2 value. Position frequency matrices were calculated based on previously reported overrepresented hexamers for SRSF11, SRSF4, SRSF5, and SRSF8 (Dominguez et al., 2018). PFMs for these RBPs were calculated as described previously (Fairbrother et al., 2002). ∆strength values for RBP motifs were calculated as described for SRE motifs. The averages of non-zero values of RBPs implicated in either the inclusion or exclusion of exon 10 were computed separately. All RBP ∆strengths for the 47 mutations are listed 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 that all transcripts had exon 10. Hence, standard linear regression was not 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 mutants. Pearson R^2 values between true values and predictions of the sample were calculated for the training set. 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 four mutants that maintained the wild-type 1:1 3R to 4R ratio in our training set, we added three VUSs from dbSNP, which we experimentally verified preserved the wild-type splicing pattern (Supplementary file 7). The VUSs tested and added to the training set were assigned a PSI of 0.5 to indicate equivalence to the WT sequence. Eq. 2, the structure ensemble model, uses four characteristics describing $\boldsymbol{X}$, the $\Delta G^{\ddagger}$ of unfolding of the region of interest around the exon-intron junction for 1000 structures in the ensemble. The mean, standard deviation (SD), skew, and kurtosis were calculated for the $\Delta G^{\ddagger}$ values of 1000 structures in each ensemble. Eq. 3, the minimum free energy model, uses just $\boldsymbol{Y}$, the $\Delta G^{\ddagger}$ of unfolding of the exon-intron junction found within the spliceosome at the B^{act} stage for the single minimum free energy structure. Eq. 4, the splice site model, uses the difference in splice site strength between WT sequence and a mutation where *SS* represents splice site. Eq. 5, 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. 6, 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. 7 is the interactive model between structure and SRE, and Eq. 8 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.

$$PSI \sim Mean(X) + SD(X) + Skew(X) + Kurtosis(X) \qquad [\mathbf{2}]$$

$$PSI \sim Y \qquad [\mathbf{3}]$$

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

$$PSI \sim \Delta E + \Delta S + \Delta SS \qquad [\mathbf{5}]$$

$$PSI \sim \Delta Ex + \Delta In \qquad [\mathbf{6}]$$

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

$$PSI \sim [Mean(X) + SD(X) + Skew(X) + Kurtosis(X)] + [\Delta E + \Delta S + \Delta SS] \qquad [\mathbf{8}]$$

Clustering of 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 $\Delta G^{\ddagger}$ of unfolding of ensembles 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 clustered by hierarchal clustering resulting in six clusters. Each individual cluster was then clustered again by hierarchal clustering using the normalized feature values instead of values of 0 and 1.

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 (Tan et al., 2019). Single-nucleotide point mutations were generated using a Q5 site-directed mutagenesis kit (NEB) and confirmed by Sanger sequencing or were custom ordered directly from GenScript. Reporter plasmids (2 μg) were transfected into HEK-293 cells in 6-well plates when cells were 60-90% confluent using Lipofectamine 3000 (ThermoFisher Scientific). Cells were harvested after 1 day by aspirating the media. Cells were resuspended 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 min), extension (50 °C, 10 min), and inactivation (85 °C, 10 min) steps. Heat-inactivated controls were prepared by heating the reaction without RNA at 85 °C for 10 min 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., R01 GM076485 to D.M., and R35 GM122532 to K.M.W. A.M.M. is a CPRIT Scholar. The authors wish to thank the Roca Lab for providing wild-type splicing reporter plasmids, Dr. Zefeng Wang for intronic splicing enhancer and silencer motifs, and Drs. Peter Castaldi and John Platig for insightful discussions.

Competing Interests

K.M.W. is an advisor to and holds equity in Ribometrix. All other 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.

Bubenik, Jodi L., Melissa Hale, Ona Mcconnell, Eric T. Wang, Maurice S. Swanson, Robert C. Spitale, and J. Andrew Berglund. 2020. "RNA Structure Probing to Characterize RNA-Protein Interactions on a Low Abundance Pre-MRNA in Living Cells." *RNA (New York, N.Y.)* 27 (3): 343–58. https://doi.org/10.1261/RNA.077263.120.

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.

Corley, Meredith, Amanda Solem, Gabriela Phillips, Lela Lackey, Benjamin Ziehr, Heather A Vincent, Anthony M Mustoe, et al. 2017. "An RNA Structure-Mediated, Posttranscriptional Model of Human

α-1-Antitrypsin Expression." *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1706539114.

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, I, 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.

Deigan, Katherine E, Tian W Li, David H Mathews, and Kevin M Weeks. 2009. "Accurate SHAPE-Directed RNA Structure Determination." *Proceedings of the National Academy of Sciences* 106 (1): 97–102. https://doi.org/10.1073/pnas.0806929106.

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.

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.

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.

Lackey, Lela, Aaztli Coria, Chanin Woods, Evonne McArthur, and Alain Laederach. 2018. “Allele-Specific SHAPE-MaP Assessment of the Effects of Somatic Variation and Protein Binding on MRNA Structure.” *RNA (New York, N.Y.)* 24 (4): 513–28. https://doi.org/10.1261/rna.064469.117.

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.

Lin, Hai, Katherine A Hargreaves, Rudong Li, Jill L Reiter, Yue Wang, Matthew Mort, David N Cooper, et al. 2019. “RegSNPs-Intron: A Computational Framework for Predicting Pathogenic Impact of Intronic Single Nucleotide Variants.” *Genome Biology* 20 (1). https://doi.org/10.1186/s13059-019-1847-4.

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.

Liu, Zhenshan, Qi Liu, Xiaofei Yang, Yueying Zhang, Matthew Norris, Xiaoxi Chen, Jitender Cheema, Huakun Zhang, and Yiliang Ding. 2021. “In Vivo Nuclear RNA Structurome Reveals RNA-Structure Regulation of MRNA Processing in Plants.” *Genome Biology* 22 (1): 1–22. https://doi.org/10.1186/S13059-020-02236-4/FIGURES/5.

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.

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.

Mustoe, Anthony M., Meredith Corley, Alain Laederach, and Kevin M. Weeks. 2018. “Messenger RNA Structure Regulates Translation Initiation: A Mechanism Exploited from Bacteria to Humans.” *Biochemistry* 57 (26): 3537–39. https://doi.org/10.1021/acs.biochem.8b00395.

Mustoe, Anthony M., Nicole N. Lama, Patrick S. Irving, Samuel W. Olson, and Kevin M. Weeks. 2019. “RNA Base-Pairing Complexity in Living Cells Visualized by Correlated Chemical Probing.” *Proceedings of the National Academy of Sciences*. https://doi.org/10.6084/m9.figshare.

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* 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.

Supek, Fran, Belén Miñana, Juan Valcárcel, Toni Gabaldón, and Ben Lehner. 2014. “Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers.” *Cell* 156 (6): 1324–35. https://doi.org/10.1016/J.CELL.2014.01.051/ATTACHMENT/3FCA62C2-C601-45EF-A9A7-821D41838468/MMC3.XLSX.

Tan, Jiazi, Lixia Yang, Alan Ann Lerk Ong, Jiahao Shi, Zhensheng Zhong, Mun Leng Lye, Shiyi Liu, et al. 2019. “Wrong- A Disease-Causing Intronic Point Mutation C19G Alters Tau Exon 10 Splicing via RNA Secondary Structure Rearrangement - Wrong Listing.” *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 Splicinḡintronic Mutations̄stem-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.

Woods, Chanin T., Lela Lackey, Benfeard Williams, Nikolay V. Dokholyan, David Gotz, and Alain Laederach. 2017. “Comparative Visualization of the RNA Suboptimal Conformational Ensemble In Vivo.” *Biophysical Journal* 113 (2): 290–301. https://doi.org/10.1016/j.bpj.2017.05.031.

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.

1602

1603

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