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

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

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

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

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

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

Abstract:

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

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

Introduction

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

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

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

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

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

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

Results

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

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

118 individual samples, the distance between the two values was only 0.25. Interestingly,
119 although *MAPT*'s function in breast tissue is not understood compared with its function
120 in the brain, for breast tissue, individuals had greater variation in Exon 10 PSI and a
121 lower median PSI compared with the pituitary gland (Figure 1-figure supplement 1B).
122 We also discovered a large amount of variation within tissues of an individual (Figure 1-
123 figure supplement 1C), although there was significantly greater variation between
124 individuals than within a single individual (see Supplementary file 1 for ANOVA table).
125 Overall, 75% of samples were within a standard deviation of the median PSI of 0.54,
126 which confirmed that the 3R to 4R isoform ratio was approximately 1:1 among
127 individuals and within different tissue types. The consistency of this isoform ratio,
128 despite the likely presence of different levels of RBPs, suggest that inherent sequence
129 and structural features regulate splicing at this exon-intron junction. RNA structure
130 regulates alternative splicing around exon-intron junctions (Warf and Berglund 2010;
131 Buratti and Baralle 2004) and a hairpin structure at the exon 10-intron 10 junction is
132 implicated in establishing the 3R to 4R 1:1 isoform ratio (Hutton et al. 1998; Varani et al.
133 1999; Grover et al. 1999; Donahue et al. 2006). Hence, we next used high-throughput
134 chemical mapping techniques to interrogate the endogenous in vivo structure of the
135 *MAPT* junction.
136
137 *Structure of 3R and 4R MAPT mature mRNA isoforms is open and accessibility of*
138 *exons is similar for the two isoforms*
139 Although the structure of the *MAPT* pre-mRNA was previously studied computationally
140 and in vitro (Varani et al. 1999; Lisowiec et al. 2015; Tan et al. 2019; Chen et al. 2019),
141 the structures of the mature 3R and 4R isoforms and *MAPT* pre-mRNA have not been
142 assessed in their endogenous in vivo context. We used dimethyl sulfate (DMS) to
143 chemically probe RNA structure in T47D and neuronal SH-SY5Y cells and primer-
144 amplified the Exon 9-Exon 11 and Exon 9-Exon 10-Exon 11 junctions during library
145 preparation for Mutational-Profiling (-MaP) (Figure 1B; Materials and methods). This
146 approach leverages the read-through aspect of MaP technology to probe the structure
147 of two alternatively spliced isoforms in the same cells. DMS reactivities for replicates,

148 and between T47D and SH-SY5Y MAPT mRNAs were highly correlated (Figure 1-figure
149 supplement 2A; Figure 1-figure supplement 2B; Figure 1-figure supplement 2D; Figure
150 1-figure supplement 2E).
151
152 We also collected in vivo DMS data for the small ribosomal RNA (SSU) whose
153 secondary structure is known from X-ray crystallography (Petrov et al. 2014) (Figure 1-
154 figure supplement 3A). The DMS reactivities for unpaired nucleotides in the SSU were
155 significantly higher than for paired nucleotides (Figure 1-figure supplement 3B),
156 confirming our probing strategy accurately recapitulates RNA secondary structure. We
157 used the SSU in vivo data to calibrate the estimation of equilibrium ensembles as
158 guided by MaP technology (Methods and materials), and we validated that structure
159 prediction guided by experimental DMS reactivities yielded more accurate estimation of
160 the SSU structure (Figure 1-figure supplement 3C). The median DMS reactivity of the
161 mature *MAPT* isoforms was 0.22, significantly greater than the median DMS reactivity of
162 the SSU, 0.0083 (Figure 1-figure supplement 3D); these results suggested that the
163 nucleotides of the mature *MAPT* isoforms were more accessible and unpaired
164 compared with the highly structured SSU, indicating that our endogenous in vivo
165 probing strategy reveals important differences in the structure of cellular RNAs.
166
167 Reactivities of Exon 9 and Exon 11 were highly correlated between the 3R and 4R
168 isoforms (Figure 1-figure supplement 2C). Additionally, computed base-pairing
169 probabilities guided by the experimental data for the two isoforms revealed that,
170 although there were some long-range interactions, 66% of base pairs spanned less than
171 50 nucleotides and were contained within the exon units (Figure 1B). This result
172 suggested that the mature exons function as their own structural unit. However, the
173 mature isoform structures did not suggest how they might regulate splicing of Exon 10.
174 Hence, we next chemically probed the MAPT pre-mRNA.
175
176
177

Figure 1

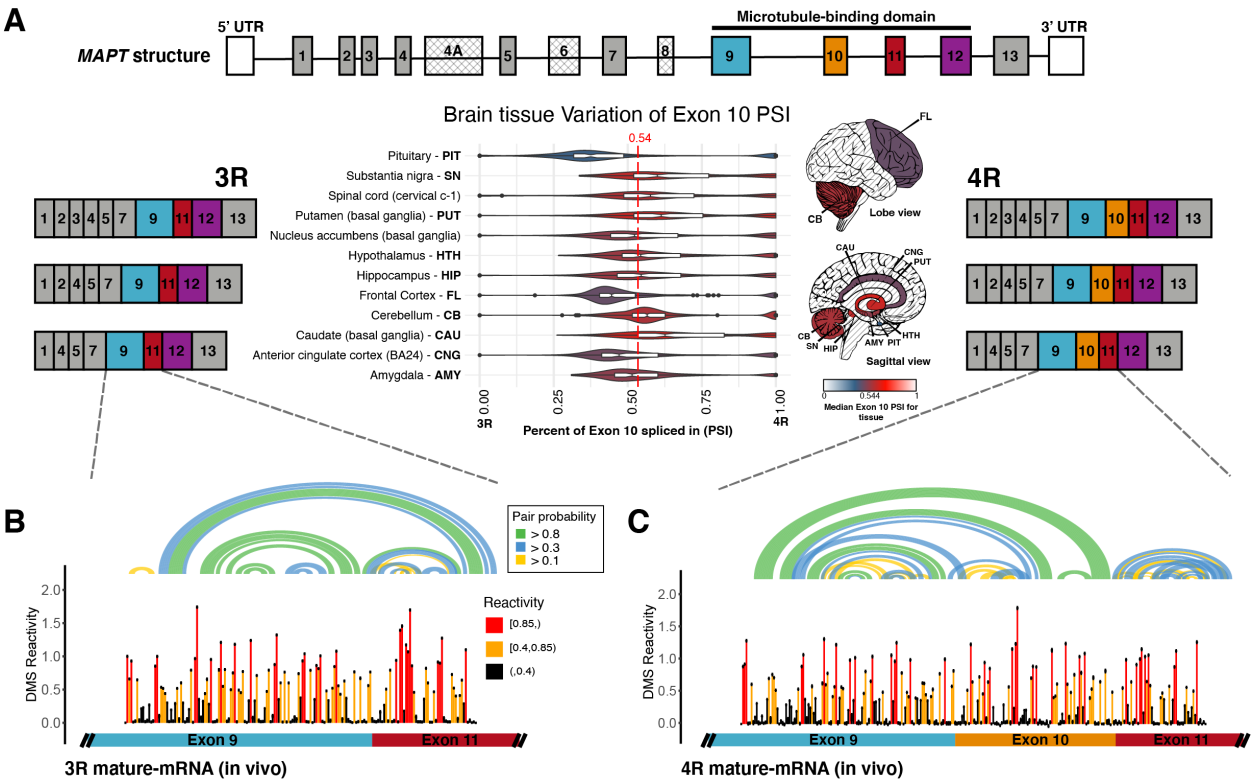


Figure 1: In vivo DMS-MaP structure probing data for 3R and 4R mature *MAPT* transcripts that are expressed in a 1:1 ratio.

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

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

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

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

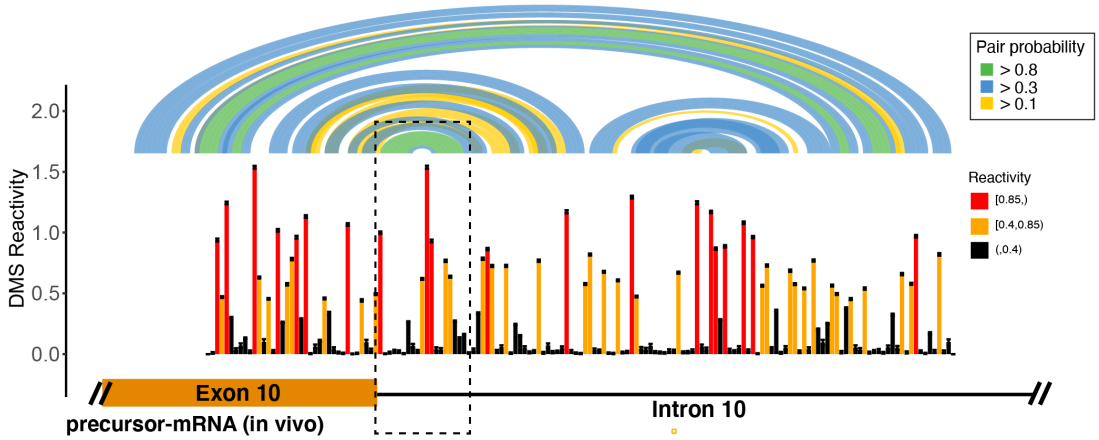
222 *MAPT pre-mRNA Exon 10-Intron 10 junction is more structured compared with the*
223 *mature isoforms*
224 Existence of a hairpin at the *MAPT* Exon 10-Intron 10 junction, implicated in regulating
225 Exon 10 splicing, was established by NMR and in vitro chemical probing (Varani et al.
226 1999; Lisowiec et al. 2015); however, the endogenous in vivo structure of this region
227 has yet to be determined. While collecting data for mature *MAPT* isoform junctions, we
228 simultaneously obtained data for the pre-mRNA Exon 10-Intron 10 junction (Figure 2A;
229 Materials and methods). Replicates were highly correlated (Figure 2-figure supplement
230 1A). Surprisingly, although Exon 10 was still being spliced, the reactivities for Exon 10 in
231 pre-mRNA and the mature 4R isoform were highly correlated (Figure 2-figure
232 supplement 1B). Again, base pairing between nucleotides appeared to be contained
233 within exons, independent of introns. The reactivities were highly correlated between
234 data collected in SH-SY5Y and T47D cells (Figure 2-figure supplement 1C); thus,
235 despite likely differences in RBP concentrations, the structure of the pre-spliced region
236 is the same between cell lines. Additionally, we found lower DMS reactivities for the pre-
237 mRNA Exon 10-Intron 10 junction compared with the mature isoform junctions (Figure
238 2-figure supplement 1D), which suggests that pre-mRNA is more structured than mature
239 mRNA. We uncovered strong evidence for the previously in vitro identified hairpin
240 structure in the DMS reactivity data; pairing probabilities were greater than 0.8 for the
241 entire hairpin stem (Figure 2A).
242
243 *Shifts in structural ensemble of MAPT Exon 10-Intron 10 junction associated with*
244 *disease mutations correlate with changes in splicing level of Exon 10*
245 Many RNAs inhabit multiple conformations in vivo to form a structural ensemble instead
246 of a single rigid structure (Halvorsen et al. 2010; Adivarahan et al. 2018). We posit that
247 a structural ensemble at the *MAPT* Exon 10-Intron 10 junction regulates Exon 10
248 splicing and disease mutations alter the composition of the structural ensemble to
249 disrupt splicing.
250

251 We used Boltzmann sampling of RNA structures guided by DMS reactivity data (Spasic
252 et al. 2018) (Materials and methods) to sample 1000 structures for the wild type and two
253 mutant intronic sequences, +15A>C and +19C>G. The two mutations alter, in opposite
254 directions, the isoform ratio at this junction (Tan et al. 2019). We visualized the
255 structural ensemble for the 3000 structures using t-Distributed stochastic neighbor
256 embedding (t-SNE) (Van Der Maaten and Hinton 2008) (Figure 2B; Materials and
257 methods). Each structure is a dot and is colored by the $\Delta G^{\ddagger}$ of unfolding of the 5' splice
258 site defined as the last three nucleotides of Exon 10 and the first six nucleotides of
259 Intron 10 (Yeo and Burge 2004). The lower the unfolding free energy, the easier to
260 unfold the structure. Overall, although there was a range of unfolding free energies for
261 the three ensembles, there were three distinct populations of free energies for the three
262 sequences (Figure 2-figure supplement 2A). We used k-means clustering to identify
263 representative structures for each cluster (Figure 2B; Figure 2-figure supplement 2B;
264 Materials and methods). We quantified and visualized the density of the clusters (Figure
265 2C; Materials and methods) and revealed distinct regions in the structure space
266 occupied by each sequence. More than 55% of structures in the ensemble of the
267 +19C>G mutation, which shifts the isoform balance entirely 3R (3R mutation) (Figure
268 2C inset), clustered in the lower left quadrant with larger unfolding free energies for the
269 splice site. This result was evidenced by the highly base-paired exon-intron junction in
270 the representative structure for the cluster. Hence, in the presence of the 3R mutation,
271 the structural ensemble of the junction shifted towards more closed structures.
272 Conversely, greater than 50% of structures in the ensemble of the +15 A>C mutation,
273 which shifts the isoform balance entirely 4R (4R mutation) (Figure 2C inset), were
274 clustered in the upper left quadrant with lower unfolding free energies for the splice site.
275 The representative structure for this region was more open and accessible around the
276 exon-intron junction. Correspondingly, the wild-type sequence had structures distributed
277 across the entire space consistent with an ensemble of structures. The exon-intron
278 junction of the representative structure for this region was not as accessible with the 4R
279 mutation, but it had fewer base-pairs than with the 3R mutation, a result recapitulated by

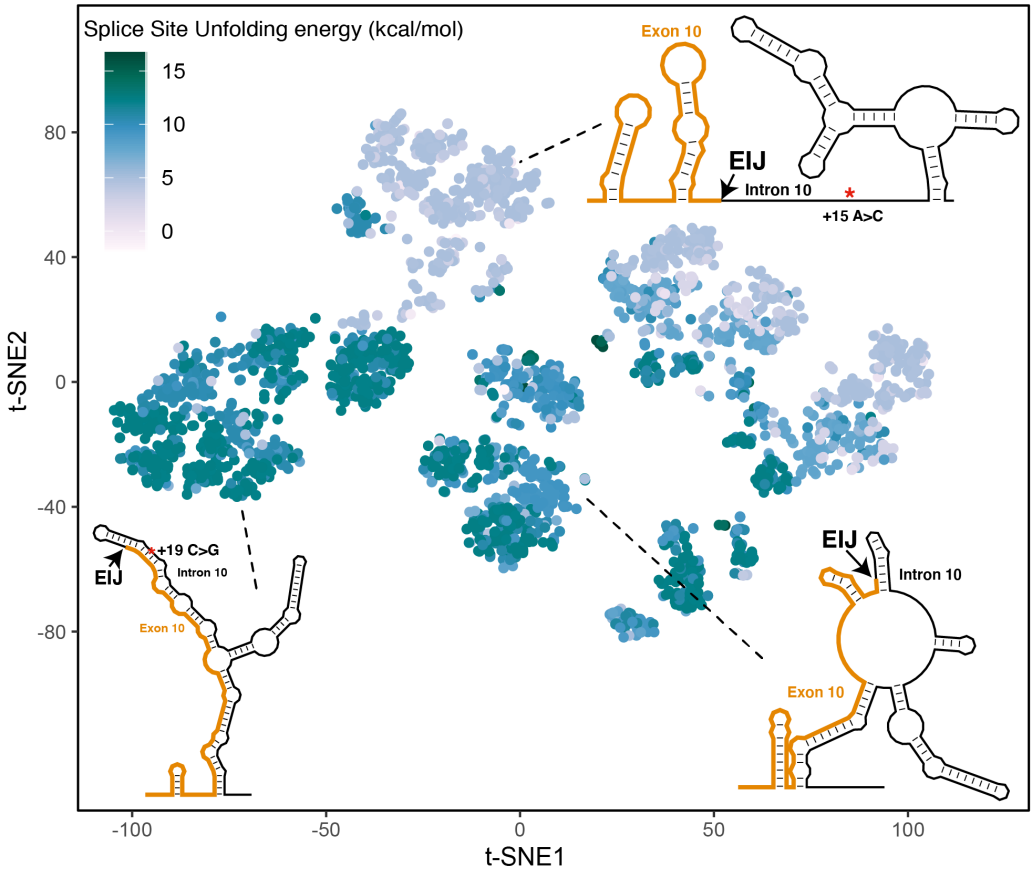
280 the two other representative structures in the right quadrants (Figure 2-figure
281 supplement 2B).
282

Figure 2

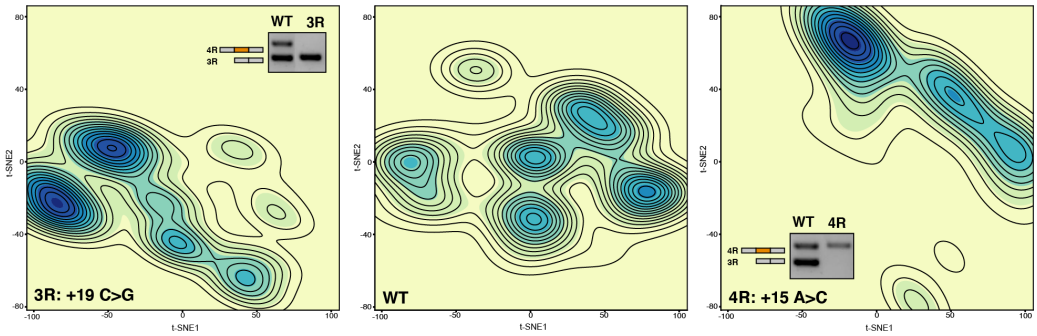
A



B



C



283

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

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

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

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

313 *Unfolding mRNA within the spliceosome B^{act} complex yields best prediction of Exon 10*
314 *splicing level*
315 RNA structure controls alternative splicing by hindering or aiding accessibility of key
316 regulatory regions to spliceosome components (McManus and Graveley 2011; Warf and
317 Berglund 2010). The 5' splice site, defined as the last 3 nucleotides of the exon and first
318 6 nucleotides of the intron, is the minimum region of RNA that must be accessible for
319 base pairing with the U1snRNA (Blanchette and Chabot 1997; Singh, Singh, and
320 Androphy 2007). However, the splicing cycle, orchestrated by the spliceosome,
321 traverses multiple stages to prepare the pre-mRNA and catalyze the two-step splicing
322 reaction (Matera and Wang 2014) (Figure 3A). The RNA itself adopts many
323 conformations as different components of the spliceosome bind to it (L. Zhang et al.
324 2019). In addition to the 5' splice site, a larger segment of RNA likely needs to unpair to
325 accommodate the changing conformations induced by the spliceosome. We analyzed
326 high resolution Cryo-EM structures of the human spliceosome Pre-B, B, Pre-B^{act}, and
327 B^{act} complexes (Charenton, Wilkinson, and Nagai 2019; Bertram et al. 2017; Townsend
328 et al. 2020; X. Zhang et al. 2018) to quantify the number of nucleotides around the 5'
329 splice site for which sufficient density was observed in the cryo-EM structure and which
330 were unpaired (Materials and methods). As can be seen in Figure 3A, the number of
331 unpaired pre-mRNA nucleotides observed in each structure increased through the
332 splicing cycle. Thus, it is likely that RNA structures outside the U1snRNA binding site
333 have to be unfolded to accommodate splicing.
334
335 To evaluate the footprint of the spliceosome that best predicts splicing outcome, we
336 initially focus on predicting 20 synonymous and intronic mutations as a training set
337 (Figure 3-figure supplement 1A). These mutations are most likely to have a structural
338 component to their function (Sharma et al. 2019; Lin, Taggart, and Fairbrother 2016).
339 The distribution of $\Delta G^{\ddagger}$ of unfolding of the splice sites in the presence of these mutations
340 was correlated with Exon 10 PSI (Figure 3-figure supplement 1B). We calculated the
341 $\Delta G^{\ddagger}$ of unfolding of the RNA near the 5' splice site in the four splicing stages' footprints.
342 Features of the unfolding free energy distribution including mean and standard deviation

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

Figure 3

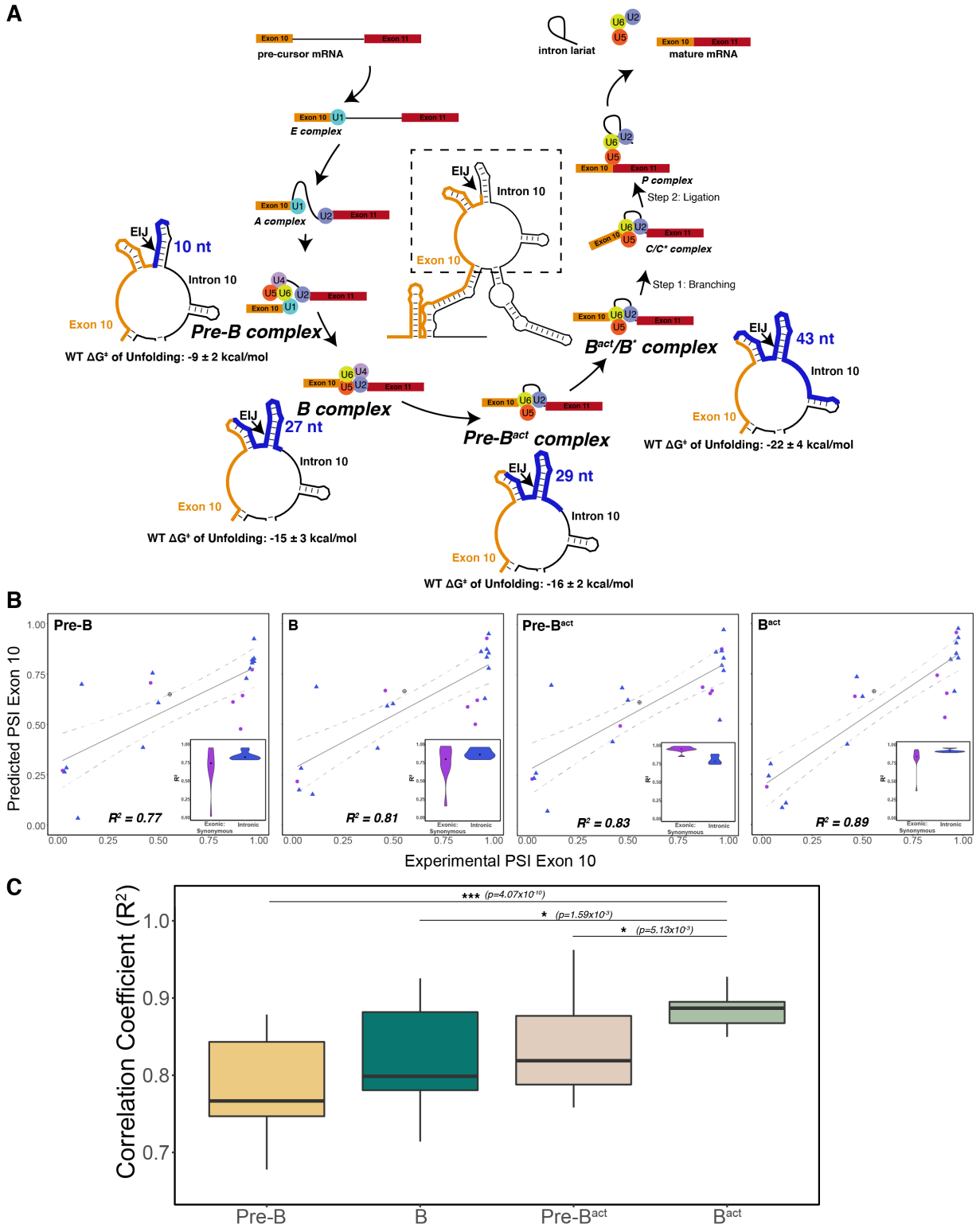


Figure 3: The best predictor of Exon 10 PSI for intronic and synonymous mutations was

375 the unfolding free energy of pre-mRNA during the B^{act} stage of splicing

376 A) Spliceosome footprint on pre-mRNA during splicing cycle. Structure in the center
377 of the cycle is the WT representative structure from Fig 2B. The dotted box
378 indicates the zoomed-in region at each stage of interest. Cryo-EM structures of
379 the human spliceosome complex at stages Pre-B (PDB ID: 6QX9), B (PDB ID:
380 5O9Z), Pre-B^{act} (PDB ID: 7ABF) and B^{act} (PDB ID: 5Z56) are available in the
381 Protein Data Bank. The region around the 5' splice site of pre-mRNA within the
382 spliceosome at each stage is highlighted in blue on the zoomed-in representative
383 structure. The number of nucleotides for each stage is as follows: Pre-B (2
384 exonic, 8 intronic); B (10 exonic, 17 intronic); Pre-B^{act} (9 exonic, 20 intronic); B^{act}
385 (12 exonic, 31 intronic). These values represent the minimum number of
386 nucleotides required to be unfolded to be accessible to the spliceosome. The
387 mean free energy and standard error to unfold RNA within the spliceosome at
388 each stage is calculated for the WT structural ensemble and indicated under the
389 zoomed-in structure.

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

399 C) Overall Pearson correlation coefficients (R^2) calculated for experimental versus
400 predicted Exon 10 PSIs by nonparametric bootstrapping of mutations. Subsets of
401 mutations were randomly sampled 10 times, trained and tested using unfolding
402 free energy of the exon-intron junction region of pre-mRNA within the
403 spliceosome for the respective splicing stage. Pearson's R^2 was calculated by
404 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

435 *Effect of exonic non-synonymous mutations was best predicted by motif strength*
436 *changes of splicing regulatory elements*
437 Exon 10 splicing is highly regulated by differential binding of RBPs to *cis*-SREs within
438 exon 10 and intron 10 (Qian and Liu 2014). While our structure-only model performs
439 moderately well for 47 mutations (R^2=0.74) (see Supplementary file 2 for further details
440 about mutations), *MAPT* Exon 10 PSIs of non-synonymous mutations were poorly
441 predicted (median bootstrapped R^2 = -0.21, Figure 4-figure supplement 1B). Hence, we
442 investigated whether these non-synonymous mutations are predicted better by
443 incorporating changes to the strength of adjacent SREs. We identified SREs by
444 similarity to reported general enhancer and silencer hexamer motifs (Fairbrother et al.
445 2002; Z. Wang et al. 2004; Yang Wang, Ma, et al. 2012; Yang Wang, Xiao, et al. 2012)
446 (Materials and methods). We calculated the changes to splice site, enhancer, and
447 silencer motif strengths in the presence of a mutation (Materials and methods) and
448 visualized the motif strength changes in a heatmap (Figure 4A). We found that using
449 splice site strength as the sole predictor yielded poor prediction of Exon 10 PSI in all
450 mutation categories (Figure 4B; Eq. 3) because most mutations were outside the splice
451 site. We quantified a weak positive correlation between PSI and enhancer strength and
452 a significant negative correlation between PSI and silencer strength (Figure 4A; Figure
453 4-figure supplement 1C). We modeled Exon 10 PSI with the changes to the motif
454 strength of all splicing regulatory elements (Eq. 4) and found an increase in prediction
455 accuracy (R^2=0.51; Figure 4C) compared with using only splice site strength (R^2=0.29).
456 Non-synonymous mutations were predicted more accurately using SRE strength with a
457 median bootstrapped R^2 of 0.75.
458
459 Many RBPs have been identified that regulate *MAPT* Exon 10 splicing (Qian et al. 2011;
460 Ian D'Souza and Schellenberg 2006; Kondo et al. 2004; J. Wang et al. 2004; L. Gao et
461 al. 2007; S. Ding et al. 2012; Broderick, Wang, and Andreadis 2004; Yan Wang et al.
462 2010; Kar et al. 2006, 2011; P. Ray et al. 2011). To determine whether these proteins
463 specific to Exon 10 splicing would improve the model's accuracy, we calculated
464 changes to the strength of their RBP motifs obtained from high throughput sequencing

465 of bound RNAs (Dominguez et al. 2018; D. Ray et al. 2013) (Materials and methods).
466 Unlike SRE motifs, there was no clear pattern or correlation between motif strength
467 change and PSI (Figure 4-figure supplement 2A, B). Subsequently, the model's
468 prediction accuracy was lower (R^2=0.08, Figure 4-figure supplement 2C), and changes
469 to the strength of general SRE motifs were better predictors of Exon 10 PSI.
470
471
472
473
474
475
476
477
478
479
480
481
482
483

Figure 4

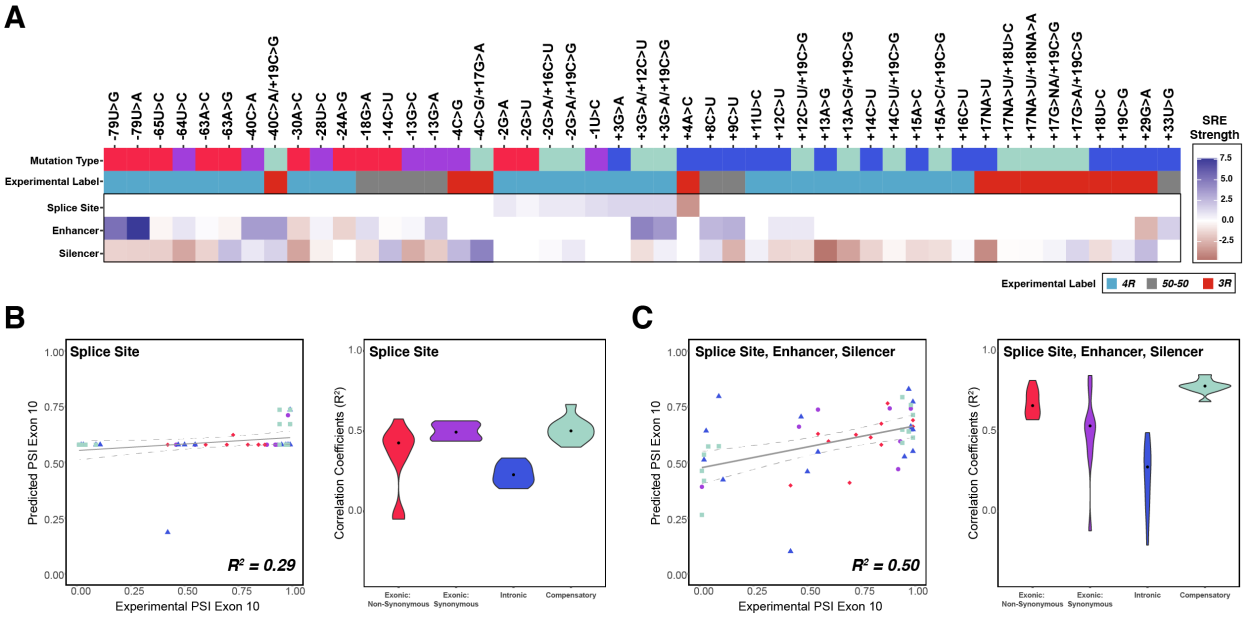


Figure 4: Combining the strength of all splicing regulatory elements improves prediction of Exon 10 PSI by 75% compared with using only splice site strength

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

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

502 parametric bootstrapping; Exonic non-synonymous (n=11), Exonic synonymous
503 (n=7), Intronic (n=15), Compensatory (n=14), Wildtype (n=1).
504 C) Exon 10 PSIs of 47 mutations predicted by combining change in splice site,
505 enhancer, and silencer strength and plotted against experimental PSIs measured
506 in splicing assays. Exon 10 PSIs predicted using Eq. 4.
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531

532 *Model with both structural and SRE motif changes yields best prediction of Exon 10 PSI*
533 Our quantitative models showed that, although SRE motif changes accurately predicted
534 the effects of non-synonymous mutations, structural changes were a better predictor of
535 splicing outcomes of intronic and synonymous mutations. Combining all features (Eq. 6)
536 yielded the highest prediction accuracy (R^2 = 0.89) (Figure 5A). This combined
537 interactive model consistently produced more accurate predictions of Exon 10 PSI
538 compared with a structure-only model and an SRE-only model for all mutation
539 categories (Figure 5B). An additive model (Eq. 7) had lower prediction accuracy (R^2=
540 0.80) (Figure 5-figure supplement 1A), and this lower accuracy resulted primarily from
541 less accurate PSI predictions of non-synonymous mutation effects (Figure 5-figure
542 supplement 1B).
543
544 To determine whether structure or SRE changes were responsible for the splicing
545 changes from each mutation, we hierarchically clustered the four primary features for
546 the 47 experimentally validated mutations (Materials and methods). Six categories
547 emerged from the clustering of features (Figure 5C) where approximately 80% of
548 mutations modified both structure and silencer strength (Figure 5-figure supplement
549 1C). Further, we found that for more than 50% of mutations both structure and SRE
550 motif strength were altered in the same direction and accordingly promoted Exon 10
551 splicing in that direction (Figure 5D). For the remaining mutations in which structure and
552 SRE strength changed in opposite directions, structure dominated the direction of
553 splicing for 18% of mutations, and SRE strength was dominant for 20% (Figure 5D).
554 Overall, these results supported our conclusion that both structure and SREs have
555 equally important effects in regulating splicing at this exon-intron junction.
556
557
558
559
560
561

Figure 5

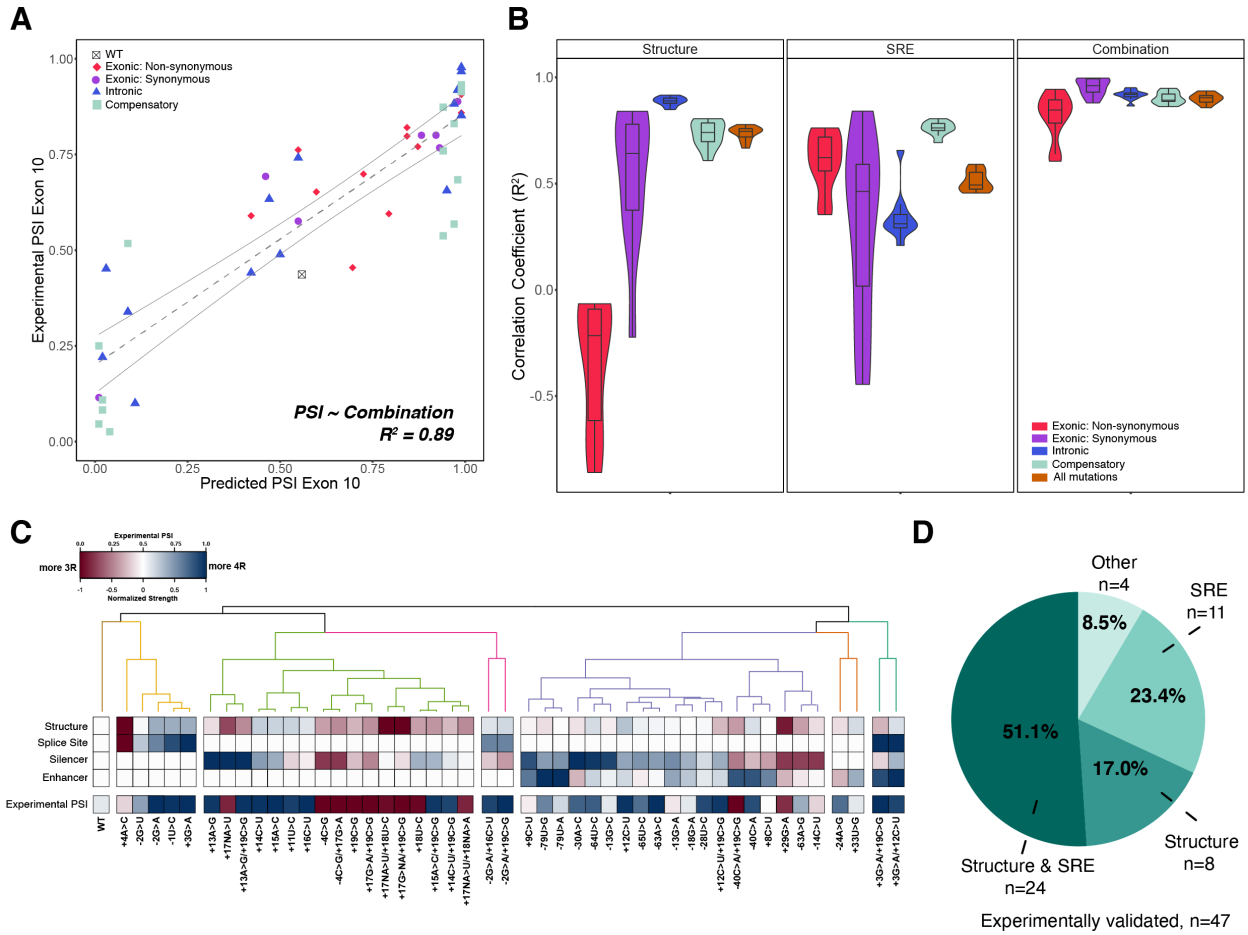


Figure 5: Combining structure and SRE strength into a unified model is the best predictor of Exon 10 PSI

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

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

575 within spliceosome at B^{act} stage as predictor. SRE strength model uses relative
576 change in SRE strength as predictor. Combined model using both structure and
577 SRE strength and weighs the features based on if mutation is
578 intronic/synonymous or non-synonymous (Materials and methods).
579 C) Heatmap of the normalized changes in structure and SRE strength for each
580 mutation clustered by affected features. Features were normalized such that a
581 value of 1 implied that change in the feature should result in Exon 10 being
582 spliced in (4R isoform, blue), whereas a value of 0 implies Exon 10 should be
583 spliced out (3R isoform, red). Mutations were clustered using hierarchal
584 clustering on normalized features (Materials and methods). Experimental PSIs
585 are plotted for each mutation with a PSI of 1 colored as blue, PSI of 0.5 colored
586 as white and PSI of 0 colored as red.
587 D) Pie chart showing the features that regulate Exon 10 splicing for the 47
588 experimentally validated mutations. The pie chart was generated based on the
589 heatmap in C. Exon 10 splicing for 51.1% of mutations is supported by changes
590 in both structure and SRE, which implies that structure, at least one SRE, and
591 PSI are either all blue or all red. Exon 10 splicing for 23.4% of mutations is
592 supported by changes in SRE wherein one of the SREs is the same color as PSI.
593 For 17.0% of mutations, structural changes support Exon 10 splicing wherein
594 structure and PSI are the same color. For 4 mutations (8.5%), the colors of none
595 of the features match the color of PSI.
596
597
598
599
600
601
602
603
604

605 *Mutations around the MAPT Exon 10-Intron 10 junction skew to Exon 10 inclusion*
606 Having established that our quantitative models accurately predicted Exon 10 PSIs for
607 experimentally validated mutations, we interrogated the model by performing a
608 systematic mutagenic analysis spanning a 100-nucleotide window of the exon-intron
609 junction (Figure 6A). Our model predicts that more mutations result in the inclusion of
610 Exon 10 (4R isoform). This is consistent with the observation that a majority (75%) of
611 known disease associated mutations (Figure 6B) are also 4R; this result is also
612 consistent when categorized by all substitution types (Figure 6-figure supplement 1A).
613 We found that a significantly greater proportion of disease mutations (76.4%) resulted in
614 changes to both structure and SRE compared with non-disease mutations (36.0%)
615 (Figure 6C) suggesting that mutations which affect both structure and SREs have a
616 greater likelihood of causing disease compared with mutations that alter only one of the
617 two factors. Intriguingly, mutations overall caused a slight skew towards a structured
618 exon-intron junction, which would result in decreased inclusion of Exon 10 (Figure 6A,
619 Figure 6-figure supplement 1B). However, changes to SRE strength skewed towards
620 increased inclusion of Exon 10 (Figure 6-figure supplement 1C), which suggested that
621 SREs were acting to counter the effect of structural changes. Our model reveals how a
622 complex balance of structure and SRE RBP binding sites effectively results in the
623 observed 50:50 ratio of the 3R and 4R isoforms.
624
625 To assess the general applicability of our model beyond our mutation training set, we
626 predicted Exon 10 PSIs for 55 variants of unknown significance (VUSs) found in dbSNP
627 (see Supplementary file 3 for further details of VUSs). These are mutations observed in
628 the human population but are not currently associated with disease. The mean Exon 10
629 PSI for VUSs was 0.66, and 70% were within a standard deviation of the mean (Figure
630 6D). We observed that only a few mutations were predicted to have a PSI of zero (3R)
631 (Figure 6D red bar). We therefore experimentally verified with splicing assays (Materials
632 and methods) 6 VUSs: 3 VUSs predicted to be 3R, 1 VUS predicted to be 4R and 2
633 VUSs predicted to maintain the WT splicing ratio (Figure 6D). We found these 6
634 predictions were correct (Figure 6E). The three 3R VUSs made the region around the

635 exon-intron junction more structured while the 4R VUS made the region less structured
636 compared to WT (Figure 6-figure supplement 1D) matching the direction of Exon 10
637 splicing change. Though we see changes to SRE strength match up to Exon 10 splicing
638 direction for +30U>C and -6G>A, this was not the case for +25C>G and +23U>C
639 (Figure 6-figure supplement 1E). For +23U>C and +26G>A, we observed changes in
640 structured-ness around the exon-intron junction and silencer strengths in diverging
641 directions (Figure 6-figure supplement 1D, E) suggesting that these opposing changes
642 would preserve the WT 3R/4R ratio.
643
644
645
646
647
648
649

Figure 6

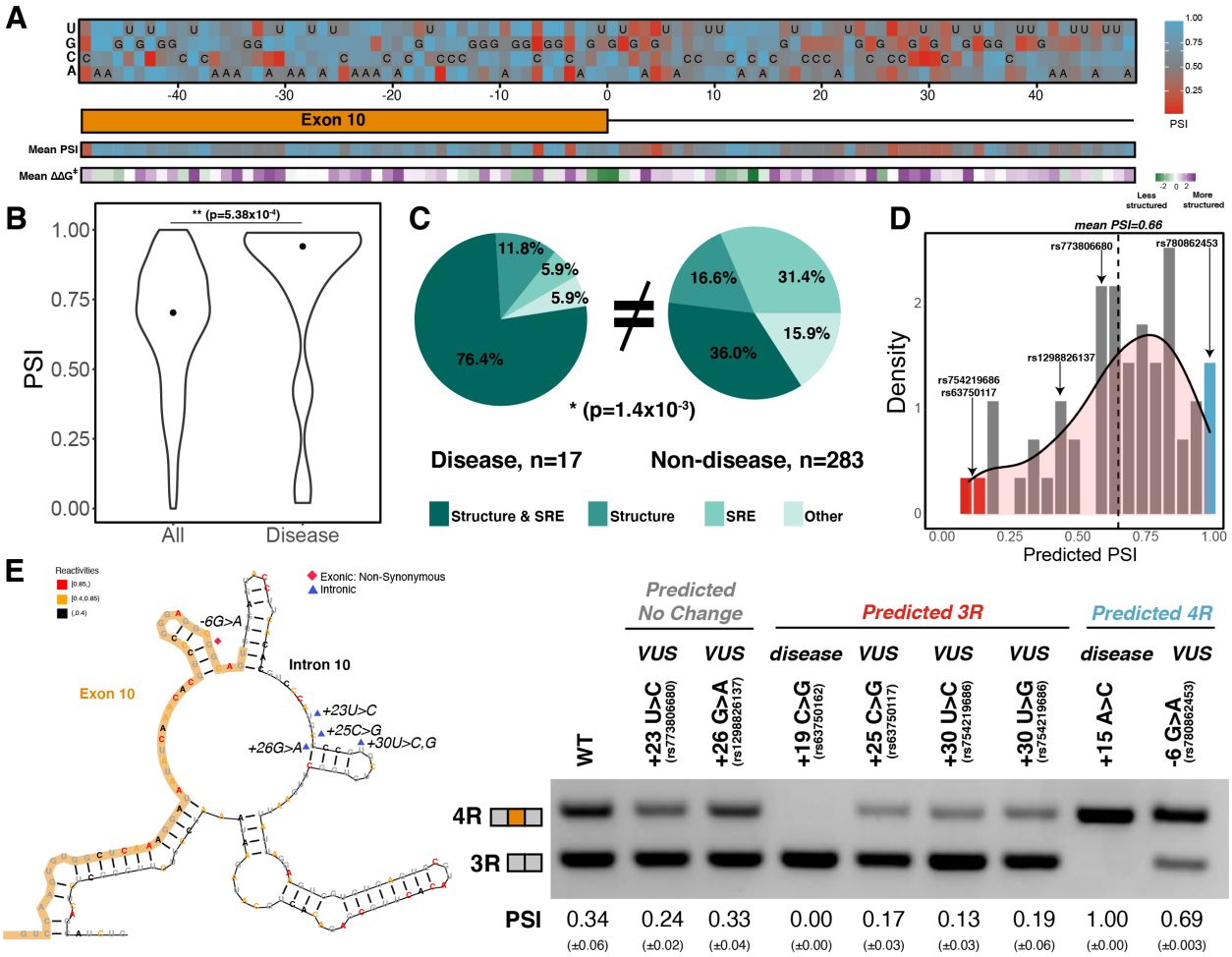


Figure 6: Mutations around Exon 10-Intron 10 junction skew towards inclusion of Exon 10

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

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

663 mutations (n=17). A two-tailed Wilcoxon Rank Sum test was used to determine
664 significance between the two categories. Level of significance: ***p-value < 10^{-6},
665 **p-value < 0.001, * p-value < 0.01
666 C) Pie chart showing features that drive Exon 10 splicing for disease and non-
667 disease mutations. The pie chart was generated by quantifying the number of
668 mutations for which the direction of predicted Exon 10 PSI matched the direction
669 of structure or SRE change. Exon 10 splicing for 76.4% of disease mutations is
670 supported by changes to both structure and SRE compared with only 36.0% of
671 non-disease mutations. The difference in proportions was tested with a one-tailed
672 Fisher's exact test.
673 D) Histogram displaying the distribution of predicted PSIs using the combined model
674 for 55 variants of unknown significance (VUSs) found in dbSNP. Density curve
675 was overlaid on top of histogram showing that predicted PSIs skew away from
676 3R. Dotted line shows mean predicted PSI of 0.66. VUSs tested in splicing
677 assays are indicated by their dbSNP RS IDs.
678 E) Representative gel of RT-PCR data for splicing assay in the presence of VUSs.
679 Splicing reporter was transfected into HEK293 cells. The mean Exon 10 PSI
680 displayed for each variant was calculated from three replicates and standard
681 error is shown in brackets below. Structure diagram on left displays the location
682 of the VUSs tested.
683
684
685
686
687
688
689
690

691 # Discussion

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

693 Splicing specificity is complex (Baralle and Giudice 2017). The spliceosome does not
694 rely on sequence alone to correctly identify 5' and 3' splice sites; other cues ensure
695 correct binding to appropriate locations. In addition, the 5' splice site must be accessible
696 to permit base pairing with the U1snRNA to initiate splicing (Roca et al. 2012). The
697 *MAPT* Exon 10-Intron 10 junction is a well-studied example of the effect of 5' splice site
698 secondary structure in splicing regulation. A hairpin was hypothesized initially because
699 disease mutations close to the exon-intron junction (Hutton et al. 1998; Grover et al.
700 1999) shifted the isoform balance to either completely exclude or include Exon 10.
701 Although NMR, in vitro chemical probing, and computation confirmed the presence of
702 the hairpin (Varani et al. 1999; Chen et al. 2019; Lisowiec et al. 2015), recent studies
703 showed that most RNAs were less structured in vivo and in the nucleus compared with
704 in vitro conditions (Sun et al. 2019; Rouskin et al. 2014). However, our results revealed
705 that this is not the case for the Exon10-Intron10 junction: in vivo chemical probing of the
706 endogenous junction showcased strong evidence of structure.

707

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

715

716 *Changes to structural ensemble around the 5' splice site are strong predictors of Exon*
717 *10 splicing*

718 We showed that structural ensembles have an important function at the Exon 10-Intron
719 10 junction. If the 5' SS was always paired, only one isoform lacking Exon 10 would

720 result. However, the simultaneous presence of 3R and 4R isoforms implies that the
721 junction is accessible in a subset of the structures. Unlike transfer RNAs and ribosomal
722 RNAs that have single structures (Petrov et al. 2014), most RNAs are dynamic,
723 unfolding and refolding within a landscape (Cruz and Westhof 2009; Giegé et al. 2012).
724 We found disease mutations produced distinct shifts in the ensemble of the *MAPT* Exon
725 10-Intron 10 junction; the shifts corresponded to changes in the 3R:4R isoform ratio and
726 confirmed that ensembles are essential at this junction. The activity of ensembles was
727 corroborated by our quantitative model; including free energy features of the structural
728 ensemble produced 1.5 times more accurate prediction of Exon 10 PSI compared with
729 the unfolding free energy of the minimum free energy (MFE) structure.
730
731 *Considering a larger spliceosome footprint on pre-mRNA produced more accurate*
732 *prediction of Exon 10 PSI*
733 The U1snRNA base pairs with the nine nucleotide sequence around the exon-intron
734 junction (Roca et al. 2012). However, our analysis of the Cryo-EM structures of the
735 human spliceosomal assembly cycle revealed that a larger region of the pre-mRNA
736 interacts with the spliceosome during the splicing cycle and is therefore unfolded. Like
737 the other main cellular ribonucleoprotein complex, the ribosome (Ingolia 2016), there is
738 likely a spliceosomal footprint on the pre-mRNA and a minimum span around the
739 splicing signals (5' splice site, 3' splice site and branch point) must be single-stranded
740 for splicing to occur. Accordingly, our structural model performed most accurately when
741 we used the unfolding free energy of 43 nucleotides around the 5' exon-intron junction
742 that exists within the spliceosome B^{act} complex. This suggests that structures distal to
743 the exon-intron junction regulate Exon 10 splicing, a finding that corroborates evidence
744 that RNA structure near this exon-intron junction is more extended than previously
745 determined (Tan et al. 2019). This result, combined with our use of Boltzmann
746 suboptimal sampling demonstrates the key role of pre-cursor mRNA structure in splicing
747 outcome.
748
749 *RNA structure and SREs have complementary functions in MAPT Exon 10 regulation*

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

780 predicted the effect of the three 3R VUSs tested. Interestingly, the systematic
781 computational mutagenesis revealed a hotspot of 3R mutations around 25-30
782 nucleotides downstream of the exon-intron junction (Figure 6A) and indeed the 3R
783 VUSs experimentally validated were located in this region.
784
785 *Quantitative modeling of splicing regulation at exon-intron junctions*
786 Predictive models can measure the contribution of individual factors to an outcome.
787 Structure around the 5' splice site and SRE motifs were excellent predictors of Exon 10
788 splicing in cells. The use of general SRE motifs enables this splicing framework to
789 extend to other exon-intron junctions. By using a common dependent variable of Exon
790 10 PSI, we could use experimentally validated mutation data from disparate sources.
791 Although our model provided an exact PSI prediction for each mutation, its principal
792 utility was in predicting the direction in which the 3R:4R isoform ratio shifted from the
793 wild type balance. On the basis of RNA-sequencing of brain tissue from healthy
794 individuals, we find a range of Exon 10 PSIs between individuals and between tissues
795 within an individual (Figure 1A). Even in individuals with progressive supranuclear palsy,
796 a tauopathy in which *MAPT* variants are implicated, there is variability in Exon 10 PSIs
797 between different brain tissues (Majounie et al. 2013). Ultimately, although it is likely
798 that what is considered the correct ratio for normal brain function varies between
799 tissues, our model provides a means to determine the baseline change of Exon 10
800 splicing simply based on sequence features. Many neurodegenerative diseases are
801 caused by mutations around the *MAPT* Exon 10-Intron 10 junction, and there are no
802 approved therapeutics that target this junction. Our work suggests that it is crucial to
803 consider the larger structural context of the Exon 10-Intron 10 junction and the interplay
804 between structure and SREs when considering the consequences of mutations on
805 splicing regulation and the design of potential therapeutics to alter this ratio.
806
807
808

809 # Materials and methods

810 ### *MAPT* Exon 10 PSIs for GTEx tissue types

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

823

824 ### Culture of T47D and SH-SY5Y cells

825 Mammary gland carcinoma cells (T47D) were cultured in RPMI 1640 medium,
826 supplemented with 10% Fetal Bovine Serum (FBS) and 0.2 units/mL of human insulin at
827 37°C and 5% CO_2. Bone marrow neuroblastoma SH-SY5Y cells were cultured in 1:1
828 mixture of 1X Minimum Essential Medium (MEM) and 1X F12 medium, supplemented
829 with 10% FBS at 37 °C and 5% CO_2.

830

831 ### In vivo DMS-MaP probing for *MAPT* RNA

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

838 mercaptoethanol. Total RNA was extracted by Trizol (ThermoFisher Scientific), treated
839 with TURBODNase (ThermoFisher Scientific), purified using Purelink RNA mini kit
840 (ThermoFisher Scientific) and quantified with NanoDrop™ spectrophotometer.
841
842 **DMS-MaP cDNA synthesis, library construction and sequencing for *MAPT* RNA**
843 Purified RNA (9 μg) was reverse transcribed using Random Primer 9 (NEB) and
844 SuperScript II reverse transcriptase under error prone conditions as described in Smola
845 et al., 2015. The resultant cDNA was purified using G50 column (GE healthcare) and
846 subjected to second strand synthesis (NEBNext Second Strand Synthesis Module).
847 Supplementary file 4 lists PCR primers used for library generation. The cDNA was
848 amplified by the NEB Q5 HotStart polymerase (NEB). Secondary PCR was performed
849 to introduce TrueSeq barcodes (Smola et al. 2015). All samples were purified using the
850 Ampure XP (Beckman Coulter) beads and quantification of the libraries was performed
851 with Qubit dsDNA HS Assay kit (ThermoFisher Scientific). Final libraries were run on
852 Agilent Bioanalyzer for quality check. TrueSeq libraries were then sequenced as
853 necessary for their desired length as paired end 2×151 and 2×301 read multiplex runs
854 on MiSeq platform (Illumina) for pre-cursor and mature MAPT isoforms respectively.
855 Sequenced reads have been uploaded to the NCBI SRA database under BioProject ID
856 PRJNA762079.
857
858 **In vivo DMS-MaP probing for *SSU ribosome***
859 For in vivo ribosomal structure data, we used approximately 10 million T47D cells in 10
860 cm plates for each condition. We removed the growth media, added 1.8 mL of bicine
861 buffered growth medium (200 mM Bicine pH 8.3) followed by treatment at 37°C with 200
862 uL of DMS diluted in ethanol (1.25% final DMS) for 5 min. For the negative control
863 (unmodified RNA), instead of DMS, an equivalent amount of ethanol was added to the
864 same number of T47D cells. After incubation, all reactions were neutralized by
865 addition of ice cold 20% by volume β-mercaptoethanol and kept on ice for 5 minutes.
866 Total RNA was extracted by Trizol (ThermoFisher Scientific), chloroform and isoamyl
867 alcohol using phase lock heavy tubes (5PRIME Phase Lock Gel). RNA was purified

868 using Purelink RNA mini kit (ThermoFisher Scientific), treated with TURBODNase
869 (ThermoFisher Scientific) and quantified with NanoDrop™ spectrophotometer.
870
871 **DMS-MaP cDNA synthesis, library construction and sequencing for *SSU***
872 ***ribosome***
873 Purified RNA (5 ug) was reverse transcribed using Random Primer 9 (NEB) and
874 SuperScript II reverse transcriptase under error prone conditions as described Smola et
875 al., 2015. The resultant cDNA was purified using G50 column (GE healthcare) and
876 subjected to second strand synthesis (NEBNext Second Strand Synthesis Module).
877 Standard Nextera DNA library protocol (Illumina) was used to fragment the cDNA and
878 add sequencing barcodes. All samples were purified using Ampure XP (Beckman
879 Coulter) beads and quantification of the libraries was performed with Qubit dsDNA HS
880 Assay kit (ThermoFisher Scientific). Final libraries were run on Agilent Bioanalyzer for
881 quality check. Libraries were sequenced as paired end 2×151 read multiplex runs on
882 MiSeq platform (Illumina). Sequenced reads have been uploaded to the NCBI SRA
883 database under BioProject ID PRJNA762079.
884
885 **DMS-MaP analysis**
886 Sequenced reads were analyzed using the ShapeMapper pipeline(Busan and Weeks
887 2018), version (v2.1.4) which calculates the DMS reactivity of each nucleotide *i* as
888 follows:
889 $$R_i = mutr_S - mutr_U$$
890 where $mutr_S$ is the mutation rate in the sample treated with DMS, $mutr_U$ is the mutation
891 rate in the untreated control. DMS reactivities were normalized within a sample and per
892 nucleotide type (A, C, U, G) using the normalization method described in Busan and
893 Weeks, 2018. DMS probing data were collected for the Exon 9-Exon 11 and Exon 9-
894 Exon 10-Exon 11 junctions using a single pair of primers listed in Supplementary file 4.
895 The ShapeMapper pipeline ran for the two junctions in a single run with reference
896 sequences for both junctions entered in one FASTA file. For the SSU, sequenced reads
897 were first aligned to the SSU ribosome sequence using Bowtie2 parameters from Busan

898 and Weeks, 2018. Using samtools, alignments with MAPQ score greater than 10 were
899 kept, sorted, and converted back into FASTQ files after which the ShapeMapper
900 pipeline was executed.
901
902 **Updating DMS parameters for RNAstructure using SSU ribosome data from T47D**
903 **cells**
904 To use DMS data to guide secondary structure prediction by the Rsample (Spasic et al.
905 2018) component of RNAstructure (Reuter and Mathews 2010), we calibrated the
906 expected DMS reactivities per nucleotide. Using the SSU ribosome mapping data and
907 the known secondary structure (Petrov et al. 2014), we determined histograms for DMS
908 reactivities for unpaired nucleotides, nucleotides paired at helix ends, and nucleotides
909 paired in base pairs stacked between two other base pairs. These DMS histograms can
910 be invoked by Rsample with the "--DMS" command line switch as part of RNAstructure
911 6.4 or later. The histograms had long tails to relatively high reactivities. We empirically
912 found that limiting reactivities in the histograms and in the input data to a reactivity of 5
913 (where higher values are set to 5) gave the best performance at improving SSU rRNA
914 secondary structure. The "--max 5" parameter is used with Rsample to apply this
915 limitation. Pre-release of Rsample code including in vivo DMS parameters is included
916 as a zip file for review, and will be included in RNAstructure 6.4.
917
918 **Base-pairing probabilities for SSU**
919 The partition function for the SSU was generated using Rsample, using either the
920 sequence or using the sequence and the DMS reactivities. All possible *i-j* base pairing
921 probabilities were summed for each nucleotide *i* to generate a base pairing probability
922 per nucleotide *i*.
923
924 **ROC curves for predicting SSU base pairs**
925 Using the known secondary structure of the SSU, we assigned a nucleotide as either 0
926 or 1 if it was paired or unpaired. DMS reactivities were used to predict whether a
927 nucleotide was paired; the higher the DMS reactivity, the more likely a nucleotide is

928 unpaired. Base pairing probabilities were subtracted from 1 to obtain the probability that
929 a base was unpaired, with 0 implying base was paired and 1 implying that base was
930 unpaired. ROC curves and AUC values were generated using the plotROC (Sachs
931 2017) R package.
932
933 **Arc plots**
934 Arc plots were generated using Superfold (Siegfried et al. 2014) modified to process
935 DMS reactivity data.
936
937 **Generating structural ensemble of Exon 10-Intron 10 *MAPT* junction**
938 The partition function of the Exon 10-Intron 10 MAPT junction for wild type (WT) and
939 mutations was calculated with DMS reactivities as restraints using Rsample (Spasic et
940 al. 2018). The DMS reactivities, which were collected for the WT sequence, were also
941 used for the mutations to restrain the structural space with the reactivity made NA at the
942 nucleotide where the mutation occurred. The program stochastic (Reuter and Mathews
943 2010) was used to sample 1000 structures (in CT format) from the Boltzmann
944 distribution wherein the likelihood a structure is sampled was proportional to the
945 probability that it occurred in the distribution (Y. Ding and Lawrence 2003).
946
947 **t-SNE visualization of structural ensembles of WT, 3R and 4R mutations of Exon**
948 **10-Intron 10 *MAPT* junction**
949 Structural ensembles were generated as described above for WT, 3R, and 4R
950 mutations. For each sequence, the 1000 structures in CT format were converted to dot-
951 bracket (db) format with ct2dot (Reuter and Mathews 2010), after which the db structure
952 was transformed into the element format using rnaConvert in the Forgi package
953 (Kerpedjiev, Höner Zu Siederdissen, and Hofacker 2015). In the element format, every
954 base is represented by the subtype of RNA structure in which it is found: stem (s),
955 hairpin (h), loop(m), 5'end(f), and 3'end(t). Hence, each db structure is a string of
956 characters. These characters were digitized (f, t:0, s:1, h:2, m:3) to create a numerical
957 matrix with 1000 rows and 234 columns, the length of the exon-intron junction.

958 Combining the matrices for the three sequences resulted in a 3000×234 matrix. This
959 matrix was entered into the tSNE function from the scikit-learn python package
960 (Pedregosa et al. 2011) and dimensionality was reduced to a 3000×2 matrix which was
961 then plotted with ggplot2 (Wickham 2016) in R. The $\Delta G^{\ddagger}$ of unfolding of the splice site
962 was calculated for each of the 3000 structures as described below. Source file for
963 Figure 3B lists t-SNE reduced data with corresponding free energies.
964
965 **Determining representative structures for clusters in t-SNE plot**
966 The 3000×2 matrix, the result of t-SNE dimensionality reduction, was clustered using k-
967 means clustering with the k-means function from the scikit-learn python package
968 (Pedregosa et al. 2011). The value of k was set to 5 as determined visually. A custom
969 python script was used to deduce the representative structure for each cluster by first
970 calculating the most common RNA structure subtype at each position. The actual
971 structure in the ensemble, most similar to the RNA structure with the most common
972 subtypes at each position, was then determined and deemed to be the representative
973 structure of that cluster.
974
975 **Visualizing density of structures in t-SNE plot**
976 A custom python script was written. For the WT and, 3R, and 4R mutant sequences, a
977 meshgrid was created for the three matrices using a 1000-point interpolation and
978 NumPy (Harris et al. 2020) meshgrid function which returns two two-dimensional arrays
979 that represent all the possible x-y coordinates for the three matrices. A gaussian kernel
980 was next fit and evaluated for each 1000×2 matrix with SciPy gaussian_kde function
981 (Virtanen et al. 2020) to smoothen over the meshgrid. Contour lines were generated for
982 the smoothed data with Matplotlib contour function (Hunter 2007) and contourf was
983 used to plot the data.
984
985 **Quantifying nucleotides around the 5' splice site in cryo-EM structure**
986 The Protein Databank (PDB) files for Pre-B (PDB ID: 6QX9), B (PDB ID: 5O9Z), Pre-
987 B^{act} (PDB ID: 7ABF) and B^{act} (PDB ID: 5Z56) complexes were downloaded from the

988 PDB website. A custom python script was used to extract pre-mRNA from each PDB
989 file. The number of nucleotides were counted for mRNA found near the 5' splice site.
990 The result was visually confirmed by visualizing the PDB on PyMol.
991
992 **Calculating $\Delta G^{\ddagger}$ of unfolding of a region of interest**
993 The $\Delta G^{\ddagger}$ of a structure was calculated using the efn2 program in RNAstructure (Reuter
994 and Mathews 2010). This represents the non-equilibrium unfolding energy of the region
995 as the sequence is not allowed to refold after unfolding (Mustoe et al. 2018). The base
996 pairs within a region of interest were removed using a custom python script. The $\Delta G^{\ddagger}$ of
997 the "unfolded" structure was next re-calculated with efn2. The $\Delta G^{\ddagger}$ of unfolding of a
998 region was the subtraction of the $\Delta G^{\ddagger}$ of the original structure from the $\Delta G^{\ddagger}$ of the
999 unfolded structure. For example, for determining the $\Delta G^{\ddagger}$ of unfolding of the splice site,
1000 we removed all base pairs within the last 3 nucleotides of the exon and the first 6
1001 nucleotides of the intron.
1002
1003 **Calculating the change in strength of SRE motifs**
1004 *Splice Site*: Strength of the WT splice site was calculated with MaxEntScan (Yeo and
1005 Burge 2004). Strength was recalculated in the presence of splice site mutations either in
1006 the last 3 bases of Exon 10 or first 6 bases of Intron 10. WT strength was subtracted
1007 from the mutant strength: a 0 implied no change in splice site strength, positive values
1008 implied that a mutation made splice site stronger, resulting in increased inclusion of
1009 Exon 10, and negative values implied that a mutation made splice site weaker and
1010 decreased inclusion of Exon 10.
1011 *Enhancers and Silencers*: Overrepresented hexamers in cell-based screens of general
1012 exonic and intronic splicing enhancers (ESEs, ISEs) and silencers (ESSs, ISSs) were
1013 obtained from Fairbrother et al., 2002, Wang et al., 2004, Wang, Ma et al., 2012 and
1014 Wang, Xiao et al., 2012. Position weight matrices (PWMs) of hexamers for each
1015 category were re-calculated as described in Fairbrother et al., 2002 and collated in
1016 Supplementary file 5. There were 8 clusters of ESE motifs, 7 of ESS motifs, 7 of ISE
1017 motifs, and 8 clusters of ISS motifs; each cluster had an associated PWM. For each

1018 PWM, a threshold strength was found by taking the 95th percentile value of strength of
1019 all possible k-mers of PWM length. This threshold was used to determine whether there
1020 was a valid SRE motif at a particular position. The strength of the PWM motif was
1021 calculated across the exon-intron junction using a sliding window for both WT sequence
1022 and per mutation. The only windows that differed were around the location of the
1023 mutation, and only windows with strength above the threshold were considered. The
1024 WT strength was subtracted from the mutation strength for each window, and all
1025 windows were then summed to yield a Δstrength for every PWM per mutation. The
1026 average of the non-zero Δstrengths was calculated for ESE, ESS, ISE and ISS
1027 categories. The ESE and ISE Δstrengths were summed to obtain an enhancer strength,
1028 and the ESS and ISS Δstrengths were summed to obtain a silencer strength.
1029 Supplementary file 6 presents all SRE Δstrengths for the 47 mutations and 55 VUSs.
1030

1031 **Calculating the change in strength of RBP motifs**

1032 Position Frequency Matrices (PFMs) were available from Ray et al. 2013 for the
1033 following RBPs: SRSF1, SRSF2, SRSF7, SRSF9, SRSF10, PCBP2, RBM4 and SFPQ.
1034 PFMs were converted into PWMs by normalizing frequencies to 0.25 (Prior probability
1035 for nucleotide frequency) and calculating the log2 value. Overrepresented hexamers
1036 were available from Dominguez et al., 2018 for the following RBPs: SRSF11, SRSF4,
1037 SRSF5 and SRSF8. PFMs for those RBPs were calculated as described in Fairbrother
1038 et al., 2002. Δstrength in RBP motifs were calculated the same way as SRE motifs. The
1039 average of non-zero values of RBPs implicated in either the inclusion or exclusion of
1040 Exon 10 was computed separately. All RBP Δstrengths for the 47 mutations are found in
1041 Supplementary file 6.
1042

1043 **Models and bootstrapping**

1044 Exon 10 PSI was limited to values between 0 and 1 with 0 signifying that no transcripts
1045 had Exon 10 and 1 implying that all transcripts had Exon 10. Hence, standard linear
1046 regression was no longer appropriate and features were fit with a beta regression model
1047 to Exon 10 PSI. Regression parameters were determined using the betareg package

1048 (Cribari-Neto and Zeileis 2010) in R. Bootstrapping was performed by sampling without
1049 replacement 70% of the mutations to train and test the model and calculating the
1050 Pearson correlation coefficient (R^2) between true values and predictions of the sample.
1051 This bootstrapping was executed 10 times resulting in a range of R^2s, ensuring that no
1052 subset of mutations skewed model performance. Since there were only 4 mutations that
1053 maintained the wildtype 3R to 4R ratio in our training set, we added 3 additional variants
1054 of unknown significance (VUSs) from the Single Nucleotide Polymorphism database
1055 (dbSNP) which we experimentally verified preserved the wildtype splicing pattern (see
1056 Supplementary file 7 for gel). WT VUSs tested and added to the training set were
1057 assigned a PSI of 0.5 to indicate equivalence to the WT sequence. Eq. **1**, the structure
1058 ensemble model, uses four characteristics describing ***X***, the ΔG° of unfolding of the
1059 region of interest around the exon-intron junction for 1000 structures in the ensemble.
1060 Eq. **2**, the minimum free energy (MFE) model, uses just ***Y***, the ΔG° of unfolding of the
1061 exon-intron junction found within the spliceosome at the B^{act} stage for the single MFE
1062 structure. Eq. **3**, the splice site model, uses the difference in splice site strength
1063 between WT sequence and a mutation where *SS* represents splice site. Eq. **4**, the
1064 combined SRE model, uses the difference in SRE strength between WT sequence and
1065 a mutation where *SS* represents splice site, *E* represents enhancer, and *S* represents
1066 silencer. Eq. **5**, the RBP model, uses the difference in RBP motif strength between WT
1067 sequence and a mutation where *Ex* represents RBPs involved in the exclusion of Exon
1068 10 and *In* represents RBPs involved in the inclusion of Exon 10. Eq. **6** is the interactive
1069 model between structure and SRE, and Eq.**7** is the additive model. *isNonSynonymous*,
1070 *isSynonymous* and *isIntronic* represent the category of mutation and is either 0 or 1.
1071 Supplementary file 6 summarizes the performance of the models and features utilized.
1072

1073 $$PSI \sim Mean(X) + SD(X) + Skew(X) + Kurtosis(X) \qquad [\mathbf{1}]$$

1074

1075 $$PSI \sim Y \qquad [\mathbf{2}]$$

1076

1077 $$PSI \sim \Delta SS \qquad [\mathbf{3}]$$

1078

1079 $$PSI \sim \Delta E + \Delta S + \Delta SS \qquad [\mathbf{4}]$$

1080

1081 $$PSI \sim \Delta Ex + \Delta In \qquad [\mathbf{5}]$$

1082

1083 $$PSI \sim [Mean(X) + SD(X) + Skew(X) + Kurtosis(X)] * [isSynonymous + isIntronic]$$

1084 $$+ [\Delta E + \Delta S + \Delta SS] * [isNonSynonymous] \qquad [\mathbf{6}]$$

1085

1086 $$PSI \sim [Mean(X) + SD(X) + Skew(X) + Kurtosis(X)] + [\Delta E + \Delta S + \Delta SS] \qquad [\mathbf{7}]$$

1087

1088 **Clustering changes in structural and SRE features**

1089 For each feature, non-zero values greater than the 95th percentile value were set to the
1090 95th percentile or, if less than the 5th percentile value, were set to the 5th percentile for
1091 visualization, after which all values were normalized to the maximum absolute value.
1092 Silencer Δstrength and mean ΔG° of unfolding of exon-intron junction of ensemble were
1093 inverted to follow the visualization such that values closer to 1 would result in greater
1094 Exon 10 inclusion and values closer to 0 would result in lower Exon 10 inclusion.
1095 Features were then assigned values 0 or 1 depending on whether the feature changed
1096 at all in the presence of the mutation. These digitized features were first clustered by
1097 hierarchal clustering resulting in 6 clusters. Each individual cluster was clustered again
1098 by hierarchal clustering but using the normalized feature values instead of 0s and 1s.

1099

1100 **Splicing Assays**

1101 HEK-293 cells (ATCC CRL-1573) were grown at 37°C in 5% CO_2 in Dulbecco's
1102 Modified Eagle Medium (Gibco) supplemented with 10% FBS (Omega Scientific) and
1103 0.5% Penicillin Streptomycin (Gibco). The wild type splicing reporter plasmid was
1104 generously provided by the Roca lab and is described in Tan et al., 2019. Single-
1105 nucleotide point mutations were generated using a Q5 site-directed mutagenesis kit
1106 (NEB) and confirmed by Sanger sequencing, or custom ordered directly from GenScript.
1107 Reporter plasmids (2 μg) were transfected into HEK-293 cells in 6-well plates at ~60-

1108 90% confluency using Lipofectamine 3000 (ThermoFisher Scientific). Cells were
1109 harvested after 1 day by aspirating the media and resuspending the cells in 1 mL Trizol
1110 reagent (ThermoFisher Scientific). RNA was isolated using the PureLink RNA Isolation
1111 Kit (ThermoFisher Scientific) with on-column DNase treatment, following manufacturer's
1112 instructions. RNA (1 μg) was reverse transcribed to cDNA using Superscript VILO
1113 reverse transcriptase (ThermoFisher Scientific). Reverse transcriptions were performed
1114 by annealing (25°C 10 minutes), extension (50°C 10 minutes), and inactivation (85°C 10
1115 min) steps. Heat-inactivated controls were prepared by heating the reaction without
1116 RNA at 85°C for 10 minutes prior to adding RNA, then following the described reaction
1117 conditions. The cDNA was PCR amplified with NEB Q5 HotStart polymerase (NEB)
1118 using splicing assay primers from IDT (AGACCCAAGCTGGCTAGCGTT forward,
1119 GAGGCTGATCAGCGGGTTTAAAC reverse) with 25 cycles. PCR product was purified
1120 and concentrated using the PureLink PCR micro clean up kit (ThermoFisher Scientific),
1121 following manufacturer's instructions. Splicing products were visualized by loading ~200
1122 ng of DNA on a 2% agarose gel in 1X tris-acetate EDTA (TAE) buffer and staining with
1123 ethidium bromide. Gel images were quantified with ImageJ.

1124

1125 **Supplementary files, figure source files, SNRNASMs and code are available at**
1126 **GitHub repository: https://git.io/JuSW8**

1127

1128 ## Acknowledgements

1129 This work was supported by the US National Institutes of Health R01 HL111527 and
1130 R35 GM 140844 to A.L. and R01 GM076485 to D.M. The authors wish to thank the
1131 Roca Lab for providing wildtype splicing reporter plasmids, Dr. Zefeng Wang for intronic
1132 splicing enhancer and silencer motifs, and Drs. Peter Castaldi, John Platig and Kevin
1133 Weeks for insightful discussions.

1134

1135 ## Competing Interests

1136 The authors have declared that no competing interests exist.

1137 # References

1138 Adivarahan, Srivathsan, Nathan Livingston, Beth Nicholson, Samir Rahman, Bin Wu, Olivia S. Rissland,
1139 and Daniel Zenklusen. 2018. “Spatial Organization of Single MRNPs at Different Stages of the Gene
1140 Expression Pathway.” *Molecular Cell* 72 (4): 727-738.e5.
1141 https://doi.org/10.1016/J.MOLCEL.2018.10.010.

1142 Andreadis, Athena. 2005. “Tau Gene Alternative Splicing: Expression Patterns, Regulation and
1143 Modulation of Function in Normal Brain and Neurodegenerative Diseases.” *Biochimica et Biophysica*
1144 *Acta - Molecular Basis of Disease* 1739 (2): 91–103. https://doi.org/10.1016/j.bbadis.2004.08.010.

1145 ———. 2012. “Tau Splicing and the Intricacies of Dementia.” *Journal of Cellular Physiology* 227 (3):
1146 1220–25. https://doi.org/10.1002/jcp.22842.

1147 Bahl, Ethan, Tanner Koomar, and Jacob J Michaelson. 2017. “CerebroViz: An R Package for Anatomical
1148 Visualization of Spatiotemporal Brain Data.” *Bioinformatics (Oxford, England)* 33 (5): 762–63.
1149 https://doi.org/10.1093/bioinformatics/btw726.

1150 Baralle, Francisco E, and Jimena Giudice. 2017. “Alternative Splicing as a Regulator of Development and
1151 Tissue Identity.” *Nature Reviews Molecular Cell Biology* 18 (7): 437–51.
1152 https://doi.org/10.1038/nrm.2017.27.

1153 Bertram, Karl, Dmitry E Agafonov, Olexandr Dybkov, Berthold Kastner, Reinhard Lü, and Holger Stark
1154 Correspondence. 2017. “Cryo-EM Structure of a Pre-Catalytic Human Spliceosome Primed for
1155 Activation.” *Cell* 170: 701-706.e11. https://doi.org/10.1016/j.cell.2017.07.011.

1156 Blanchette, Marco, and Benoit Chabot. 1997. “A Highly Stable Duplex Structure Sequesters the 5’ Splice
1157 Site Region of HnRNP A1 Alternative Exon 7B.” *RNA*, no. 3: 405–19.

1158 Broderick, Jennifer, Junning Wang, and Athena Andreadis. 2004. “Heterogeneous Nuclear
1159 Ribonucleoprotein E2 Binds to Tau Exon 10 and Moderately Activates Its Splicing.” *Gene* 331 (1–2):
1160 107–14. https://doi.org/10.1016/j.gene.2004.02.005.

1161 Buratti, Emanuele, and Francisco E Baralle. 2004. “Influence of RNA Secondary Structure on the Pre-
1162 MRNA Splicing Process.” *MOLECULAR AND CELLULAR BIOLOGY* 24 (24): 10505–14.
1163 https://doi.org/10.1128/MCB.24.24.10505–10514.2004.

1164 Busan, Steven, and Kevin M Weeks. 2018. “Accurate Detection of Chemical Modifications in RNA by
1165 Mutational Profiling (MaP) with ShapeMapper 2.” *RNA* 24 (2): 143–48.
1166 https://doi.org/10.1261/rna.061945.117.

1167 Catarina Silva, M., and Stephen J. Haggarty. 2020. “Tauopathies: Deciphering Disease Mechanisms to
1168 Develop Effective Therapies.” *International Journal of Molecular Sciences* 21 (23): 1–49.
1169 https://doi.org/10.3390/ijms21238948.

1170 Charenton, Clément, Max E. Wilkinson, and Kiyoshi Nagai. 2019. “Mechanism of 5’ Splice Site Transfer
1171 for Human Spliceosome Activation.” *Science* 364 (6438): 362–67.

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

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

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

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

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

D’Souza, 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.

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

1208 Prediction." *Nucleic Acids Research* 31 (24): 7280–7301. https://doi.org/10.1093/nar/gkg938.
1209 Dominguez, Daniel, Peter Freese, Maria S. Alexis, Amanda Su, Myles Hochman, Tsultrim Palden,
1210 Cassandra Bazile, et al. 2018. "Sequence, Structure, and Context Preferences of Human RNA
1211 Binding Proteins." *Molecular Cell* 70(5): 854–867.e9. https://doi.org/10.1016/j.molcel.2018.05.001.
1212 Donahue, Christine P, Christina Muratore, Jane Y Wu, Kenneth S Kosik, and Michael S Wolfe. 2006.
1213 "Stabilization of the Tau Exon 10 Stem Loop Alters Pre-MRNA Splicing." *Journal of Biological*
1214 *Chemistry* 281 (33): 23302–6. https://doi.org/10.1074/jbc.C600143200.
1215 Fairbrother, William G., Ru-Fang Yeh, Phillip A. Sharp, and Christopher B. Burge. 2002. "Predictive
1216 Identification of Exonic Splicing Enhancers in Human Genes." *Science* 297: 1007–13.
1217 http://dx.doi.org/10.1126/science.1073774.
1218 Ferrari, Silvia L.P., and Francisco Cribari-Neto. 2004. "Beta Regression for Modelling Rates and
1219 Proportions." *Journal of Applied Statistics* 31 (7): 799–815.
1220 https://doi.org/10.1080/0266476042000214501.
1221 Gao, Lei, Junning Wang, Yingzi Wang, and Athena Andreadis. 2007. "SR Protein 9G8 Modulates Splicing
1222 of Tau Exon 10 via Its Proximal Downstream Intron, a Clustering Region for Frontotemporal
1223 Dementia Mutations." *Molecular and Cellular Neuroscience* 34 (1): 48–58.
1224 https://doi.org/10.1016/j.mcn.2006.10.004.
1225 Gao, Qing Sheng, John Memmott, Robert Lafyatis, Stefan Stamm, Gavin Screaton, and Athena
1226 Andreadis. 2000. "Complex Regulation of Tau Exon 10, Whose Missplicing Causes Frontotemporal
1227 Dementia." *Journal of Neurochemistry* 74 (2): 490–500. https://doi.org/10.1046/j.1471-
1228 4159.2000.740490.x.
1229 Giegé, Richard, Frank Jühling, Joern Pütz, Peter Stadler, Claude Sauter, and Catherine Florentz. 2012.
1230 "Structure of Transfer RNAs: Similarity and Variability." *Wiley Interdisciplinary Reviews: RNA* 3 (1):
1231 37–61. https://doi.org/10.1002/WRNA.103.
1232 Goedert, M., M. G. Spillantini, R. A. Crowther, S. G. Chen, P. Parchi, M. Tabaton, D. J. Lanska, et al.
1233 1999. "Tau Gene Mutation in Familial Progressive Subcortical Gliosis." *Nature Medicine* 5 (4): 454–
1234 57. https://doi.org/10.1038/7454.
1235 Goedert, M., M G. Spillantini, R. Jakes, D Rutherford, and R A Crowther. 1989. "Multiple Isoforms of
1236 Human Microtubule-Associated Protein Tau: Sequences and Localization in Neurofibrillary Tangles
1237 of Alzheimer's Disease." *Neuron* 3 (4): 519–26. https://doi.org/10.1016/0896-6273(89)90210-9.
1238 Grover, Andrew, Henry Houlden, Matt Baker, Jennifer Adamson, Jada Lewis, Guy Prihar, Stuart
1239 Pickering-Brown, Karen Duff, and Mike Hutton. 1999. "5' Splice Site Mutations in Tau Associated
1240 with the Inherited Dementia FTDP-17 Affect a Stem-Loop Structure That Regulates Alternative
1241 Splicing of Exon 10*." *Journal of Biological Chemistry* 274 (21): 15134–43.
1242 http://www.jbc.org/content/274/21/15134.full.pdf.
1243 Halvorsen, Matthew, Joshua S Martin, Sam Broadaway, and Alain Laederach. 2010. "Disease-Associated

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1316 Morris, Michael A. Nalls, et al. 2013. “Variation in Tau Isoform Expression in Different Brain Regions
1317 and Disease States.” *Neurobiology of Aging* 34 (7): 1922.e7-1922.e12.
1318 https://doi.org/10.1016/j.neurobiolaging.2013.01.017.
1319 Matera, A. Gregory, and Zefeng Wang. 2014. “A Day in the Life of the Spliceosome.” *Nature Reviews*
1320 *Molecular Cell Biology* 15 (2): 108–21. https://doi.org/10.1038/nrm3742.
1321 McManus, C. Joel, and Brenton R. Graveley. 2011. “RNA Structure and the Mechanisms of Alternative
1322 Splicing.” *Current Opinion in Genetics and Development*. https://doi.org/10.1016/j.gde.2011.04.001.
1323 Mirra, Suzanne S ;, Jill R ; Murrell, Marla ; Gearing, and Maria G Spillantini. 1999. “Tau Pathology in a
1324 Family with Dementia and a P301L Mutation in Tau.” *Journal of Neuropathology and Experimental*
1325 *Neurology* 58 (4): 335.
1326 https://search.proquest.com/docview/229718949/fulltextPDF/AA5FABCB40F44E3FPQ/1?accountid
1327 =14244.
1328 Mustoe, Anthony M., Steven Busan, Greggory M. Rice, Christine E. Hajdin, Brant K. Peterson, Vera M.
1329 Ruda, Neil Kubica, Razvan Nutiu, Jeremy L. Baryza, and Kevin M. Weeks. 2018. “Pervasive
1330 Regulatory Functions of MRNA Structure Revealed by High-Resolution SHAPE Probing.” *Cell* 173
1331 (1): 181-195.e18. https://doi.org/10.1016/j.cell.2018.02.034.
1332 Park, Sun Ah, Sang Il Ahn, and Jean Marc Gallo. 2016. “Tau Mis-Splicing in the Pathogenesis of
1333 Neurodegenerative Disorders.” *BMB Reports* 49 (8): 405–13.
1334 https://doi.org/10.5483/BMBRep.2016.49.8.084.
1335 Pedregosa, F, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, et al. 2011. “Scikit-
1336 Learn: Machine Learning in Python.” *Journal of Machine Learning Research* 12: 2825–30.
1337 Petrov, Anton S., Chad R. Bernier, Burak Gulen, Chris C. Waterbury, Eli Hershkovits, Chiaolong Hsiao,
1338 Stephen C. Harvey, et al. 2014. “Secondary Structures of RRNAs from All Three Domains of Life.”
1339 *PLoS ONE* 9 (2): e88222. https://doi.org/10.1371/journal.pone.0088222.
1340 Qian, Wei, Khalid Iqbal, Inge Grundke-Iqbal, Cheng Xin Gong, and Fei Liu. 2011. “Splicing Factor SC35
1341 Promotes Tau Expression through Stabilization of Its MRNA.” *FEBS Letters* 585 (6): 875–80.
1342 https://doi.org/10.1016/j.febslet.2011.02.017.
1343 Qian, Wei, and Fei Liu. 2014. “Regulation of Alternative Splicing of Tau Exon 10.” *Neuroscience Bulletin*
1344 30 (2): 367–77. https://doi.org/10.1007/s12264-013-1411-2.
1345 Ray, Debashish, Hilal Kazan, Kate B Cook, Matthew T Weirauch, Hamed S Najafabadi, Xiao Li, Serge
1346 Gueroussov, et al. 2013. “A Compendium of RNA-Binding Motifs for Decoding Gene Regulation.”
1347 *Nature* 499 (7457): 172–77. https://doi.org/10.1038/nature12311.
1348 Ray, Payal, Amar Kar, Kazuo Fushimi, Necat Havlioglu, Xiaoping Chen, and Jane Y. Wu. 2011. “PSF
1349 Suppresses Tau Exon 10 Inclusion by Interacting with a Stem-Loop Structure Downstream of Exon
1350 10.” *Journal of Molecular Neuroscience* 45 (3): 453–66. https://doi.org/10.1007/s12031-011-9634-z.
1351 Reuter, Jessica S, and David H Mathews. 2010. “RNAstructure: Software for RNA Secondary Structure

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

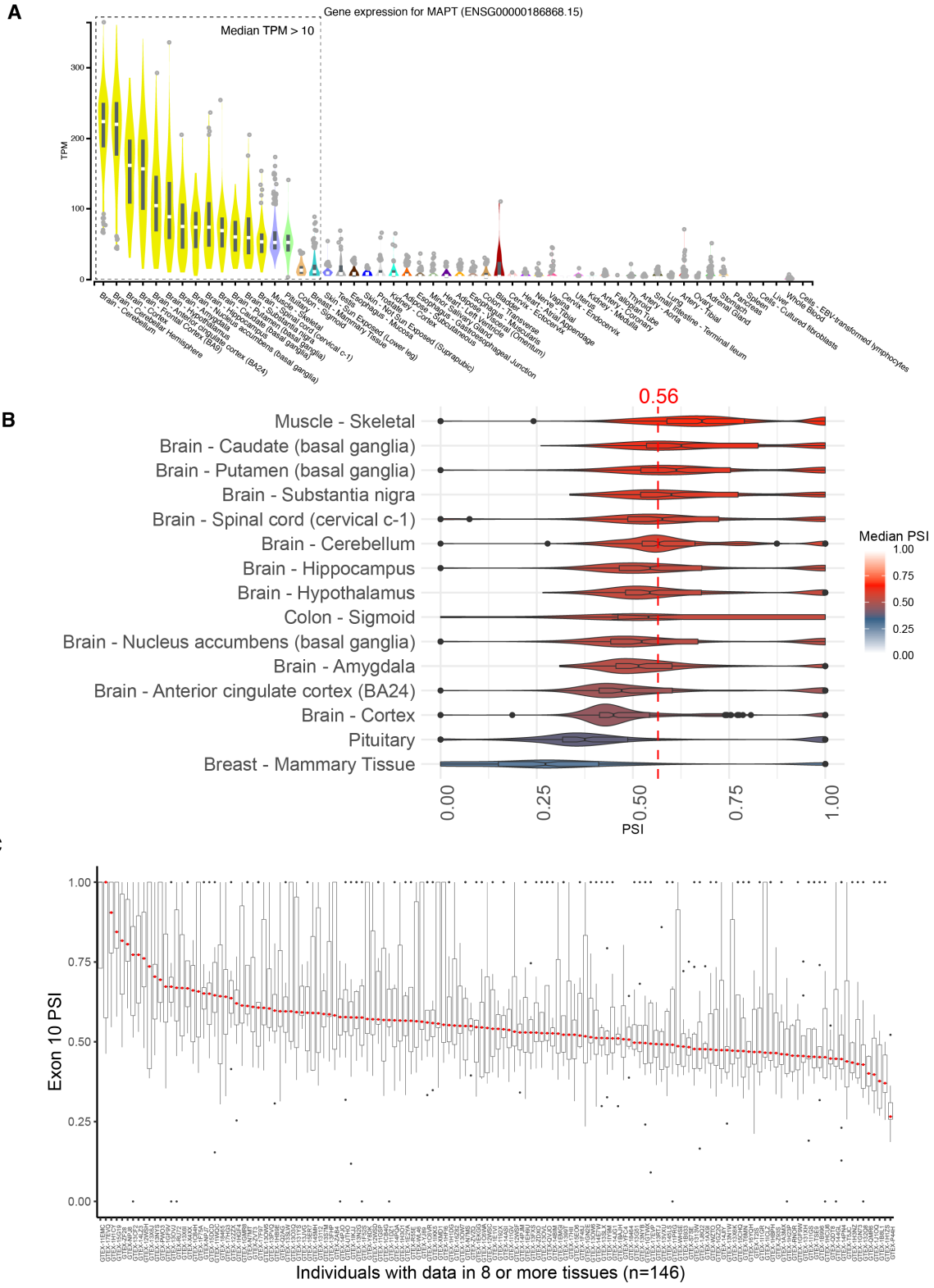
Varani, Luca, Masato Hasegawa, Maria Grazia Spillantini, Michael J Smith, Jill R Murrell, Bernardino Ghetti, Aaron Klug, Michel Goedert, and Gabriele Varani. 1999. “Structure of Tau Exon 10 Splicing Regulatory Element RNA and Destabilization by Mutations of Frontotemporal Dementia and Parkinsonism Linked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

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

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

1424 Wang, Yan, Lei Gao, Sze Wah Tse, and Athena Andreadis. 2010. “Heterogeneous Nuclear
1425 Ribonucleoprotein E3 Modestly Activates Splicing of Tau Exon 10 via Its Proximal Downstream
1426 Intron, a Hotspot for Frontotemporal Dementia Mutations.” *Gene* 451 (1–2): 23–31.
1427 https://doi.org/10.1016/j.gene.2009.11.006.
1428 Wang, Yang, Meng Ma, Xinshu Xiao, and Zefeng Wang. 2012. “Intronic Splicing Enhancers, Cognate
1429 Splicing Factors and Context-Dependent Regulation Rules.” *Nature Structural & Molecular Biology*
1430 19 (10): 1044–52. https://doi.org/10.1038/nsmb.2377.
1431 Wang, Yang, Xinshu Xiao, Jianming Zhang, Rajarshi Choudhury, Alex Robertson, Kai Li, Meng Ma,
1432 Christopher B Burge, and Zefeng Wang. 2012. “A Complex Network of Factors with Overlapping
1433 Affinities Represses Splicing through Intronic Elements.” *Nature Structural & Molecular Biology* 20
1434 (1): 36–45. https://doi.org/10.1038/nsmb.2459.
1435 Wang, Zefeng, and Christopher B Burge. 2008. “Splicing Regulation: From a Parts List of Regulatory
1436 Elements to an Integrated Splicing Code.” *RNA* 14: 802–13. https://doi.org/10.1261/rna.876308.
1437 Wang, Zefeng, Michael E. Rolish, Gene Yeo, Vivian Tung, Matthew Mawson, and Christopher B. Burge.
1438 2004. “Systematic Identification and Analysis of Exonic Splicing Silencers.” *Cell* 119 (6): 831–45.
1439 https://doi.org/10.1016/j.cell.2004.11.010.
1440 Warf, M Bryan, and J Andrew Berglund. 2010. “Role of RNA Structure in Regulating Pre-MRNA Splicing.”
1441 *Trends in Biochemical Sciences*. https://doi.org/10.1016/j.tibs.2009.10.004.
1442 Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
1443 https://ggplot2.tidyverse.org.
1444 Yeo, Gene, and Christopher B Burge. 2004. “Maximum Entropy Modeling of Short Sequence Motifs with
1445 Applications to RNA Splicing Signals.” *Journal of Computational Biology* 11 (2–3): 377–94.
1446 https://doi.org/10.1089/1066527041410418.
1447 Zhang, Lingdi, Anne Vielle, Sara Espinosa, and Rui Zhao. 2019. “RNAs in the Spliceosome: Insight from
1448 CryoEM Structures.” *Wiley Interdisciplinary Reviews: RNA* 10 (3): 1–11.
1449 https://doi.org/10.1002/wrna.1523.
1450 Zhang, Xiaofeng, Chuangye Yan, Xiechao Zhan, Lijia Li, Jianlin Lei, and Yigong Shi. 2018. “Structure of
1451 the Human Activated Spliceosome in Three Conformational States.” *Cell Research* 28 (3): 307–22.
1452 https://doi.org/10.1038/cr.2018.14.
1453 Zubradt, Meghan, Paromita Gupta, Sitara Persad, Alan M Lambowitz, Jonathan S Weissman, and Silvi
1454 Rouskin. 2016. “DMS-MaPseq for Genome-Wide or Targeted RNA Structure Probing in Vivo.”
1455 *Nature Methods* 14 (1): 75–82. https://doi.org/10.1038/nmeth.4057.
1456
1457
1458

Supplementary Figures

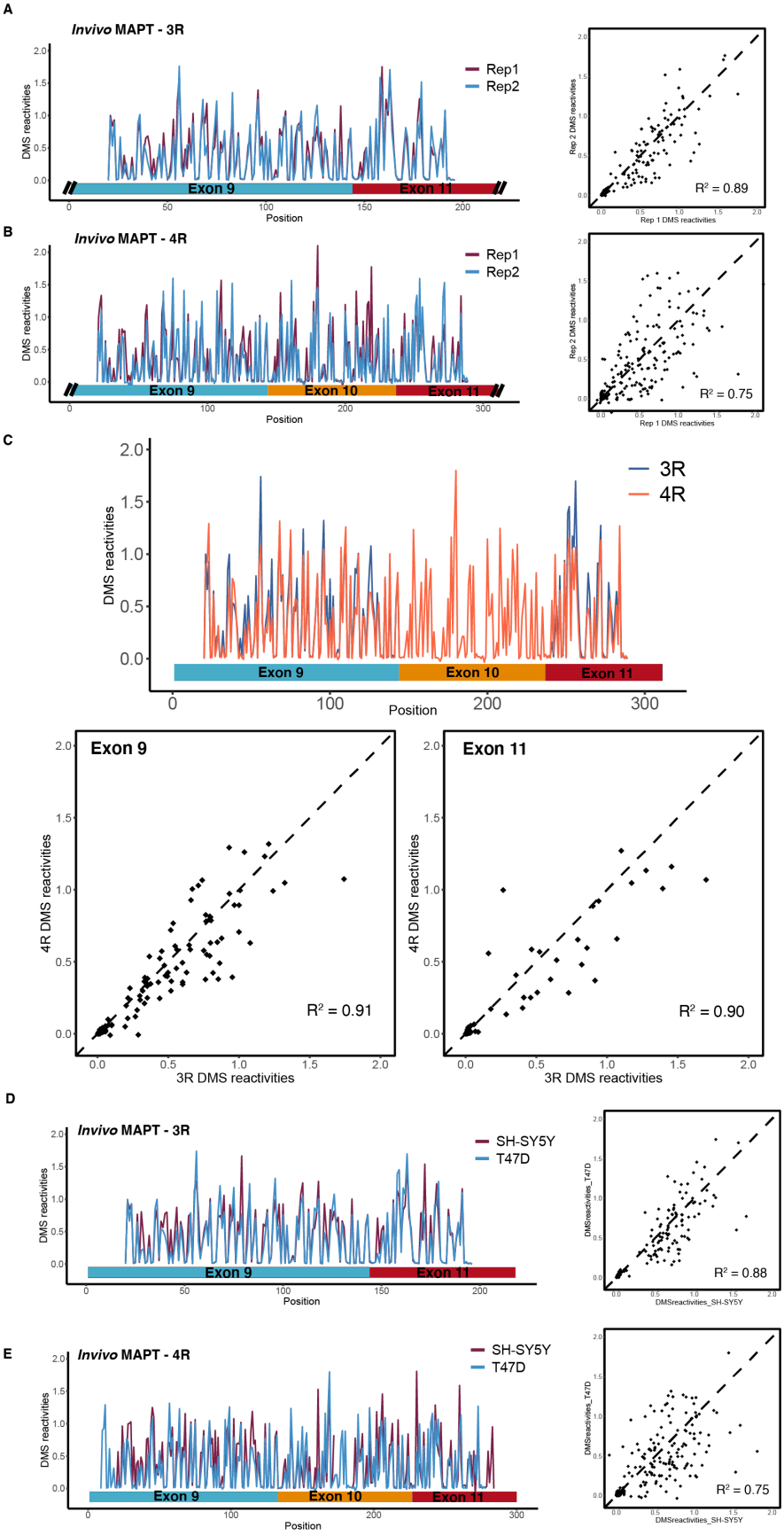


1461 **Figure 1-figure supplement 1**: Distribution of Exon 10 PSIs calculated for RNA-seq
1462 data from GTEx database.

1463 A) Distribution of TPM values of *MAPT* gene expression for tissues in the GTEx
1464 database sorted by median TPM. Dotted box indicates tissues with median TPM
1465 greater than 10. MAPT is highly expressed in the brain, and there is little
1466 expression in other tissues. Figure was downloaded from the GTEx website.

1467 B) Distribution of Exon 10 PSI for 12 central nervous system, muscle-skeletal,
1468 colon-sigmoid, and breast-mammary tissue types. Percent Spliced In (PSI) of
1469 Exon 10 was calculated from RNA-seq data for 2,962 tissue samples among 15
1470 tissue types collected from 818 individuals in GTEx v8 database. The violin plot
1471 for each tissue type and the corresponding region on the brain diagram is colored
1472 by the median PSI for all samples of that type. The median PSI of 0.56 for all
1473 tissue samples is indicated by the red dotted line.

1474 C) Distribution of Exon 10 PSI for tissues per individual. Only individuals with *MAPT*
1475 expression data in 8 or more tissues were plotted. Median PSI for each individual
1476 is labelled by red dot on box plot.


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

1477

1478 **Figure 1-figure supplement 2**: DMS structure probing data for mature *MAPT* 3R and
1479 4R isoforms

1480 A) DMS reactivity data from T47D cells for two biological replicates for Exon 9-Exon
1481 11 junction (3R isoform). Structure probing data for junctions of interest were
1482 obtained using primers (Supplementary file 4) following RT of extracted RNA.
1483 DMS reactivity is plotted for each nucleotide across spliced junctions for both
1484 replicates overlaid in plot on the left. For scatter plot on the right, DMS reactivity
1485 for Rep 1 vs Rep 2 is plotted per nucleotide with Pearson's correlation coefficient
1486 displayed.

1487 B) DMS reactivity data from T47D cells for two biological replicates for Exon 9-Exon
1488 10-Exon 11 junction (4R isoform).

1489 C) Comparison of DMS reactivity data for 3R vs 4R isoforms. Replicates 1 and 2
1490 were pooled for each isoform. Top plot shows DMS reactivity plotted for each
1491 nucleotide with isoforms overlaid. No data were collected for Exon 10 for the 3R
1492 isoform because Exon 10 is spliced out. Bottom left scatter plot shows DMS
1493 reactivities for Exon 9 in the 3R vs 4R context, whereas bottom right scatter plot
1494 shows DMS reactivities for Exon 11 in the 3R vs 4R context. Pearson's
1495 correlation coefficient is shown for each comparison.

1496 D) DMS reactivity data from T47D and SH-SY5Y cells for Exon 9-Exon 11 junction
1497 (3R isoform). Replicates from T47D cells were pooled. DMS reactivity is plotted
1498 for each nucleotide across spliced junctions for both replicates overlaid in plot on
1499 the left. For scatter plot on the right, DMS reactivity for T47D vs SH-SY5Y is
1500 plotted per nucleotide with Pearson's correlation coefficient displayed.

1501 E) DMS reactivity data from T47D and SH-SY5Y cells for Exon 9-Exon 10-Exon 11
1502 junction (4R isoform).

1503

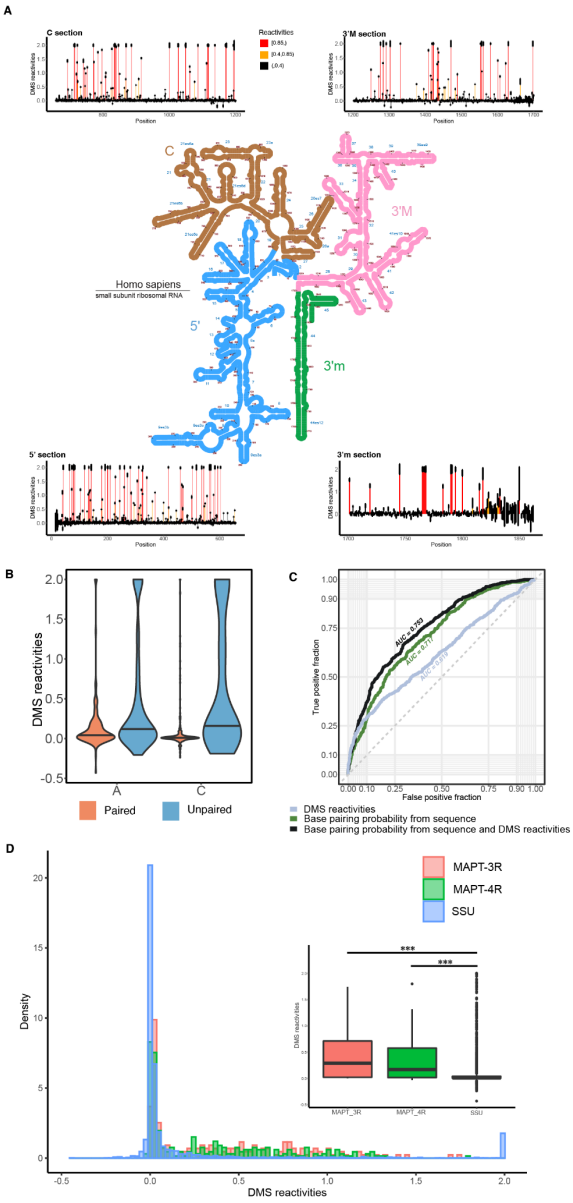
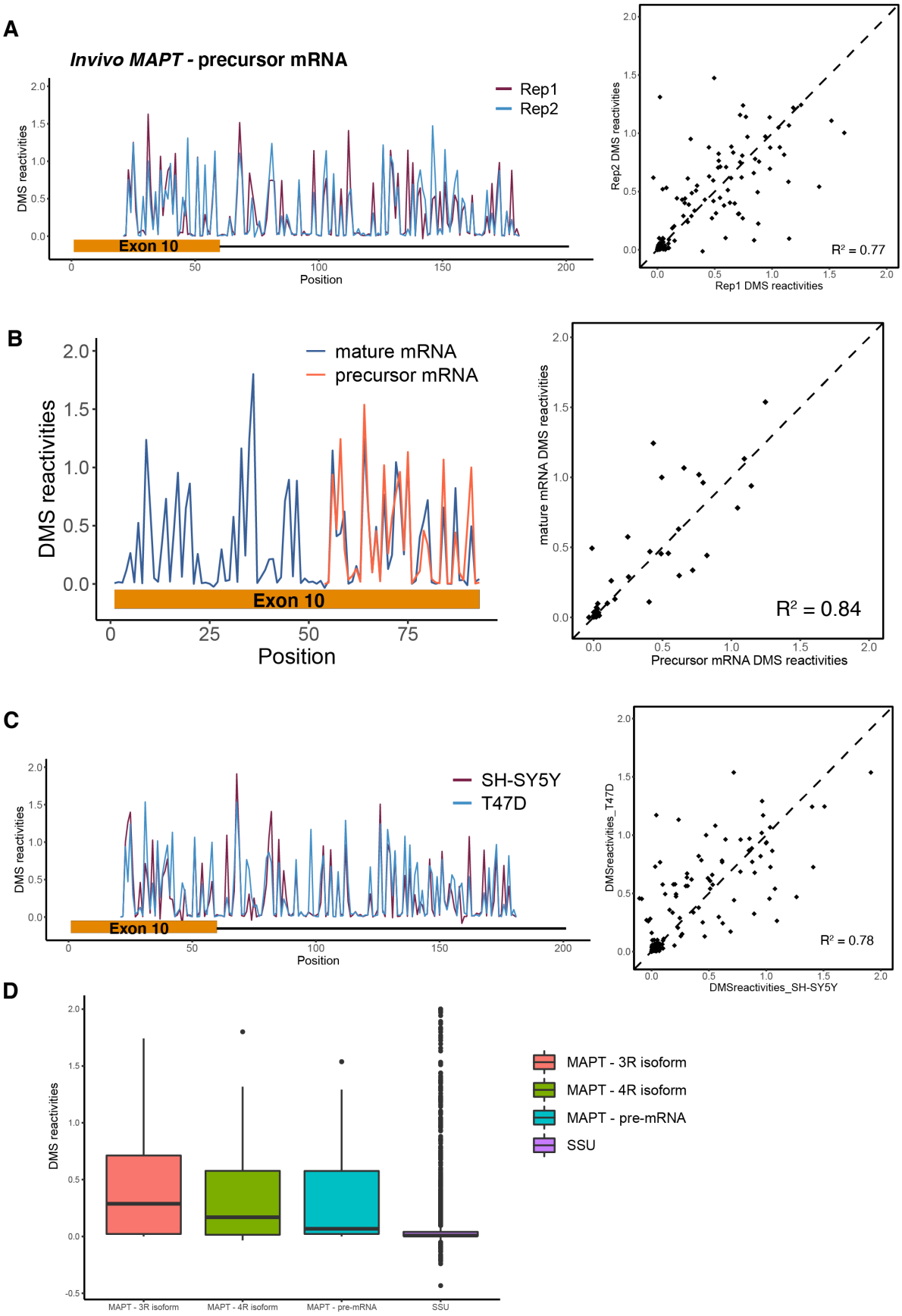


Figure 1-figure supplement 3: DMS structure probing data for human small subunit ribosome RNA (SSU) collected from T47D cells.

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

1514 lab Ribosome Gallery website
1515 (http://apollo.chemistry.gatech.edu/RibosomeGallery/eukarya/H%20sapiens/SSU/ind
1516 ex.html).
1517 B) Violin plots showing distribution of DMS reactivities for adenines and cytosines
1518 partitioned by paired versus unpaired nucleotides. Pairing status of nucleotides was
1519 determined from the known secondary structure of the SSU. Median DMS reactivity
1520 is indicated by thick horizontal black line on violin plot.
1521 C) ROC curves for predicting whether a nucleotide in the SSU is paired. Three different
1522 parameters were used for each of the three curves: DMS reactivities, base pairing
1523 probabilities predicted from SSU sequence, and base pairing probabilities for SSU
1524 sequence that were guided by DMS reactivities. The area under the curve (AUC) for
1525 each curve was calculated with AUCs closer to 1 corresponding to higher accuracy
1526 of predictions. Dotted line indicates AUC of 0.5 which corresponds to a model
1527 making random predictions.
1528 D) Comparison of distribution of DMS reactivities between SSU, *MAPT* 3R and 4R
1529 isoforms. Larger plot shows a density histogram of the DMS reactivities for each
1530 RNA. Inset boxplots display distribution of DMS reactivities. Level of significance:
1531 ***p-value $< 10^{-6}$


A
Invivo MAPT - precursor mRNA
Rep1
Rep2
DMS reactivities
Exon 10
Position
Rep2 DMS reactivities
Rep1 DMS reactivities
R² = 0.77
B
mature mRNA
precursor mRNA
DMS reactivities
Exon 10
Position
mature mRNA DMS reactivities
Precursor mRNA DMS reactivities
R² = 0.84
C
SH-SY5Y
T47D
DMS reactivities
Exon 10
Position
DMSreactivities_T47D
DMSreactivities_SH-SY5Y
R² = 0.78
D
DMS reactivities
MAPT - 3R isoform
MAPT - 4R isoform
MAPT - pre-mRNA
SSU

1532

1533 **Figure 2-figure supplement 1**: DMS structure probing data for precursor *MAPT* Exon
1534 10-Intron 10 junction

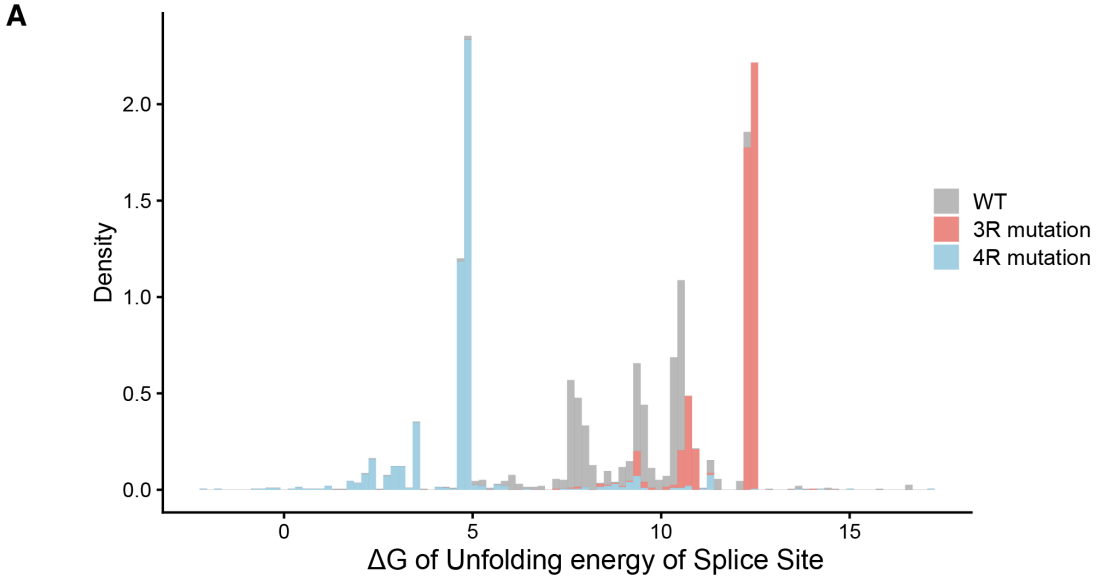
1535 A) DMS reactivity data from T47D cells for two biological replicates for Exon 10-
1536 Intron 10 junction. Structure probing data for junctions of interest were obtained
1537 using primers following RT of extracted RNA. DMS reactivity is plotted for each
1538 nucleotide across spliced junctions for both replicates overlaid in plot on the left.
1539 For scatter plot on the right, DMS reactivity for Rep 1 vs Rep 2 is plotted per
1540 nucleotide with Pearson's correlation coefficient displayed.

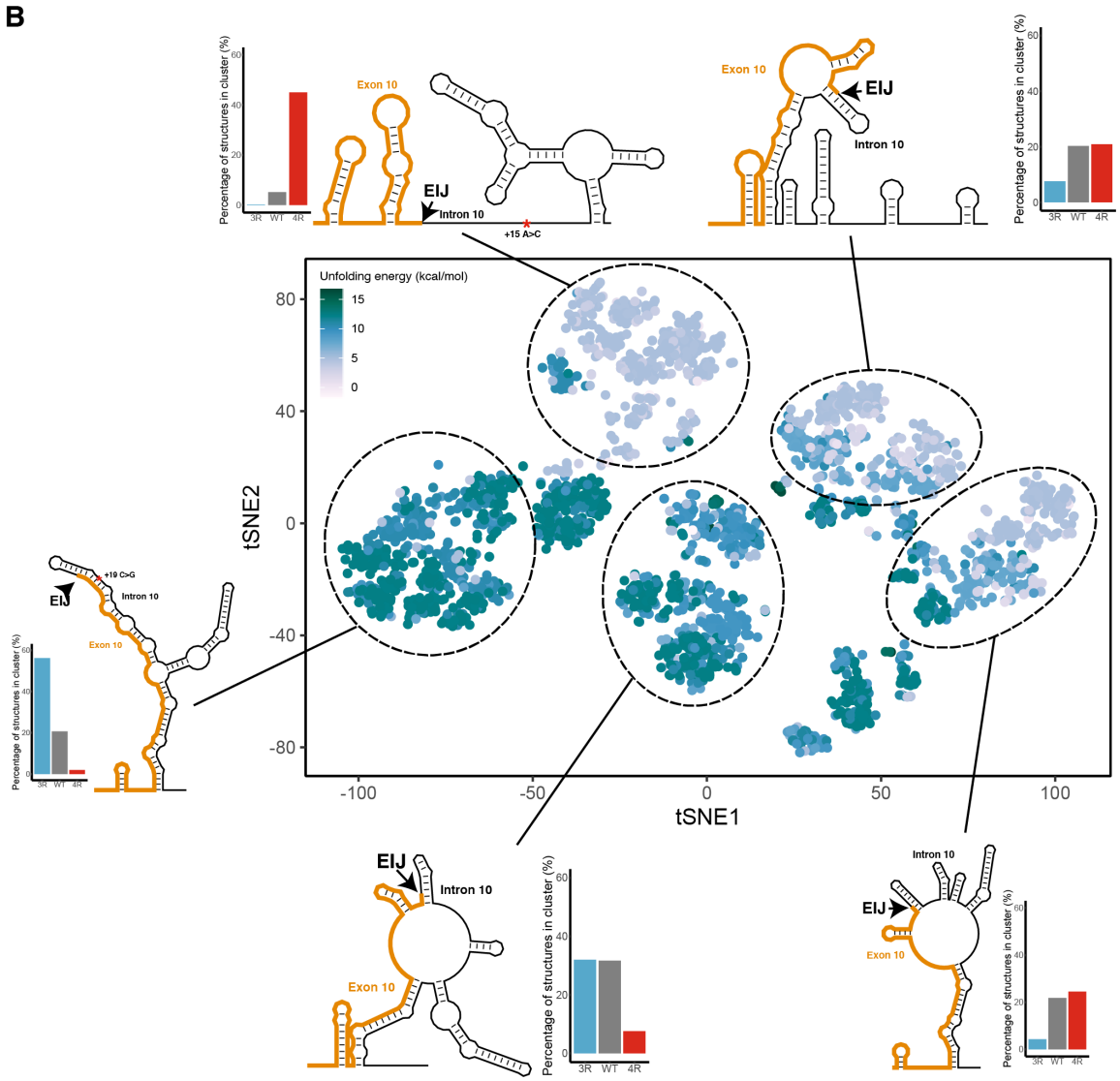
1541 B) DMS reactivity data comparing Exon 10 in precursor vs mature transcript.
1542 Replicates 1 and 2 were pooled for each transcript. Right plot shows DMS
1543 reactivity plotted for each nucleotide with mature and precursor RNAs overlaid.
1544 DMS data for all of Exon 10 could not be collected for the precursor RNA due to
1545 the position of primers chosen for amplification. Scatter plot on the left shows
1546 DMS reactivities for Exon 10 in the precursor vs mature mRNA context with
1547 Pearson's correlation coefficient shown for the comparison.

1548 C) DMS reactivity data from T47D and SH-SY5Y cells for Exon 10-Intron 10
1549 junction. Replicates from T47D cells were pooled. DMS reactivities are plotted for
1550 each nucleotide across exon-intron junctions for both cell types overlaid in plot on
1551 the left. For scatter plot on the right, DMS reactivity for T47D vs SH-SY5Y is
1552 plotted per nucleotide with Pearson's correlation coefficient displayed.

1553 D) Boxplots of distribution of DMS reactivities between SSU, *MAPT* 3R isoform, 4R
1554 isoform and pre-cursor mRNA.

1555


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


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

1556

1557 **Figure 2-figure supplement 2**: The 3R and 4R mutations shift the WT structural
1558 ensemble towards less and more accessible exon-intron junctions, respectively.

1559 A) Density histogram showing the distribution of unfolding free energies of the splice
1560 site (defined as last 3 nucleotides of exon, first 6 nucleotides of intron) for all
1561 structures in the ensemble for WT, 3R and 4R mutated sequence. Distributions
1562 for each sequence are colored according to the legend.

1563 B) t-SNE Visualization of structural ensemble of wild type (WT) and, 3R (+19C>G)
1564 and 4R (+15A>C) mutations. Structures were predicted using Boltzmann
1565 suboptimal sampling and guided by DMS reactivity data (in Figure 2A). Data
1566 were visualized using t-Distributed Stochastic Neighbor Embedding (t-SNE).
1567 Shown are 3000 structures corresponding to 1000 structures per category. Each
1568 dot represents a single structure and was colored by calculated unfolding free
1569 energy of splice site at exon-intron junction (3 exonic bases, 6 intronic bases).
1570 Data were clustered by k-means clustering and representative structures for five
1571 clusters are shown. Bar plots next to the representative structure show the
1572 proportion of the cluster in WT, 3R and 4R. The exon-intron junction is marked by
1573 EIJ on each structure. Position of 3R and 4R mutations are marked by a red
1574 asterisk on their respective representative structures. There are two additional
1575 representative structures shared by WT and 4R sequence which have similar
1576 structural contexts around the EIJ as the representative WT structure in Figure
1577 2B.

1578

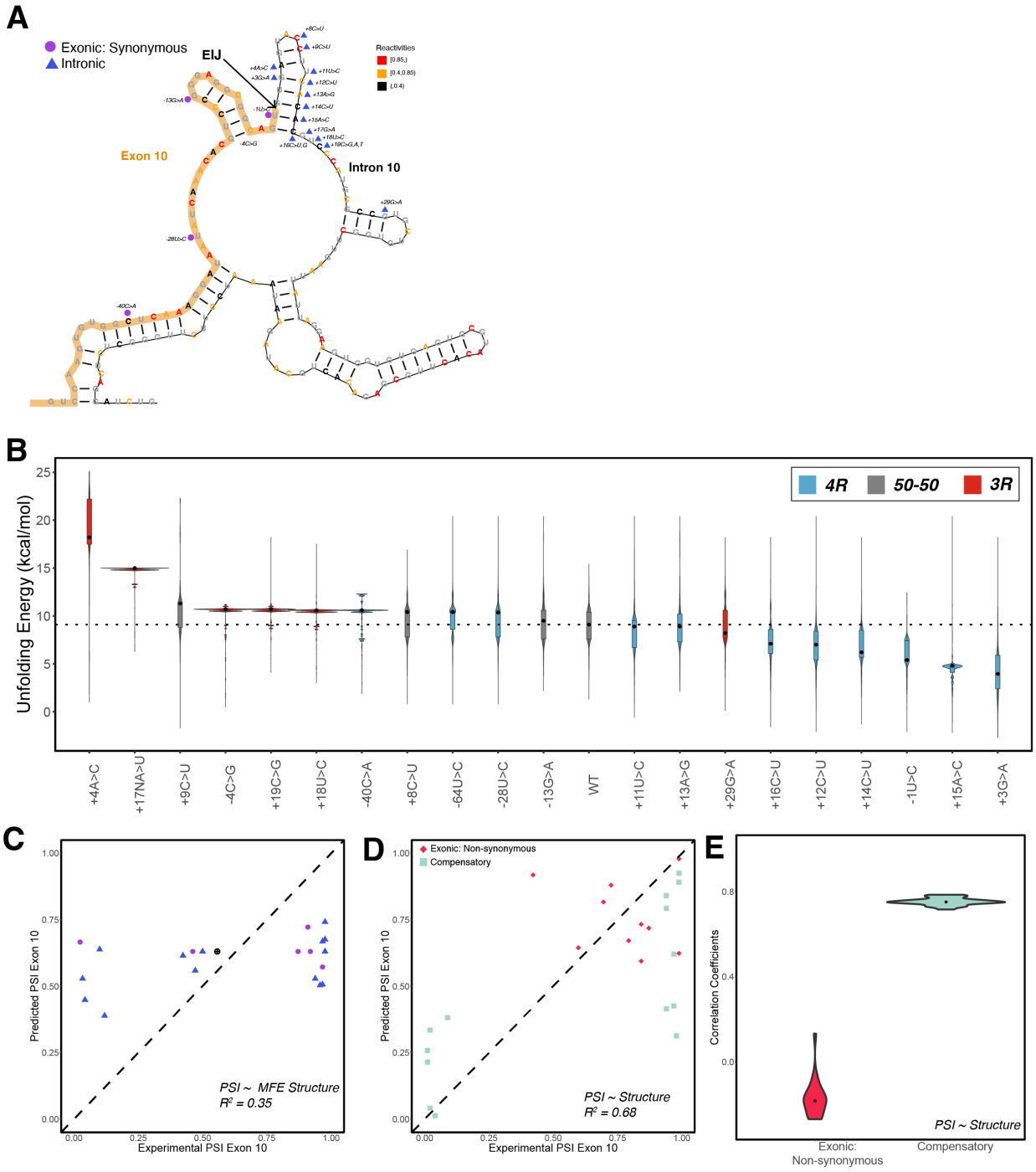


Figure 3-figure supplement 1: Structure is a poor predictor of Exon 10 PSI for exonic non-synonymous mutations

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

1584 from Figure 2B. Each nucleotide is colored by its corresponding reactivity value
1585 based on the color scale.
1586 B) Violin plots showing the distribution of unfolding free energy of the exon-intron
1587 pre-mRNA in the spliceosome B^{act} complex for the 1000 structures in the
1588 ensembles of the 22 intronic and synonymous mutations. Each violin plot is
1589 colored by whether the mutation promotes the 3R or 4R isoform ratio or the ratio
1590 remains 50:50.
1591 C) Exon 10 PSIs of 22 mutations predicted using unfolding free energy of the exon-
1592 intron pre-mRNA in B^{act} complex of the spliceosome for the single minimum free
1593 energy (MFE) structure and plotted against experimental PSIs measured in
1594 splicing assays. Exon 10 PSIs predicted using Eq. 2. Each point on the
1595 scatterplot represents a mutation and is colored by mutation category. Dotted
1596 diagonal line is the x=y line, and the closer the points are to the diagonal, the
1597 more accurate the prediction. Pearson correlation coefficient (R^2) of experimental
1598 to predicted PSIs was calculated.
1599 D) Exon 10 PSIs of non-synonymous and compensatory mutations predicted using
1600 the unfolding free energy of pre-mRNA within the spliceosome B^{act} stage plotted
1601 against corresponding experimental PSIs measured in splicing assays. Exon 10
1602 PSIs were predicted using Eq. 1.
1603 E) Violin plots show R^2s calculated for each mutation category by training and
1604 testing on subsets of all mutations by non-parametric bootstrapping; Non-
1605 synonymous (n=10), Compensatory (n=14).

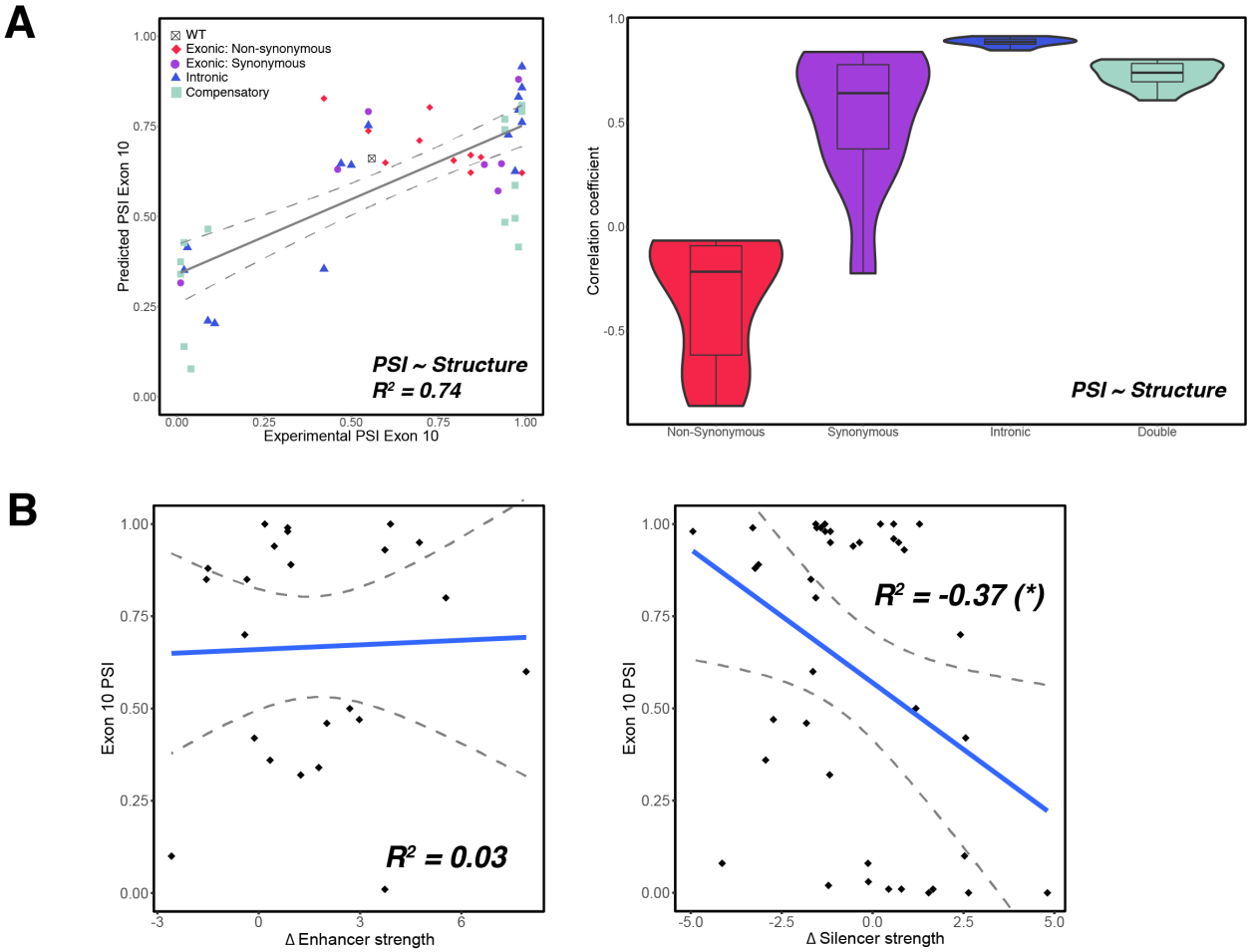


Figure 4-figure supplement 1: RBP binding motif strength is a poor predictor of Exon 10 PSI for all mutations

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

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

1620 lines indicating the 95% confidence interval. Pearson correlation coefficient (R^2)
1621 is shown. The negative correlation between silencer strength and Exon 10 PSI is
1622 statistically significant with a p-value of 0.004.
1623
1624
1625

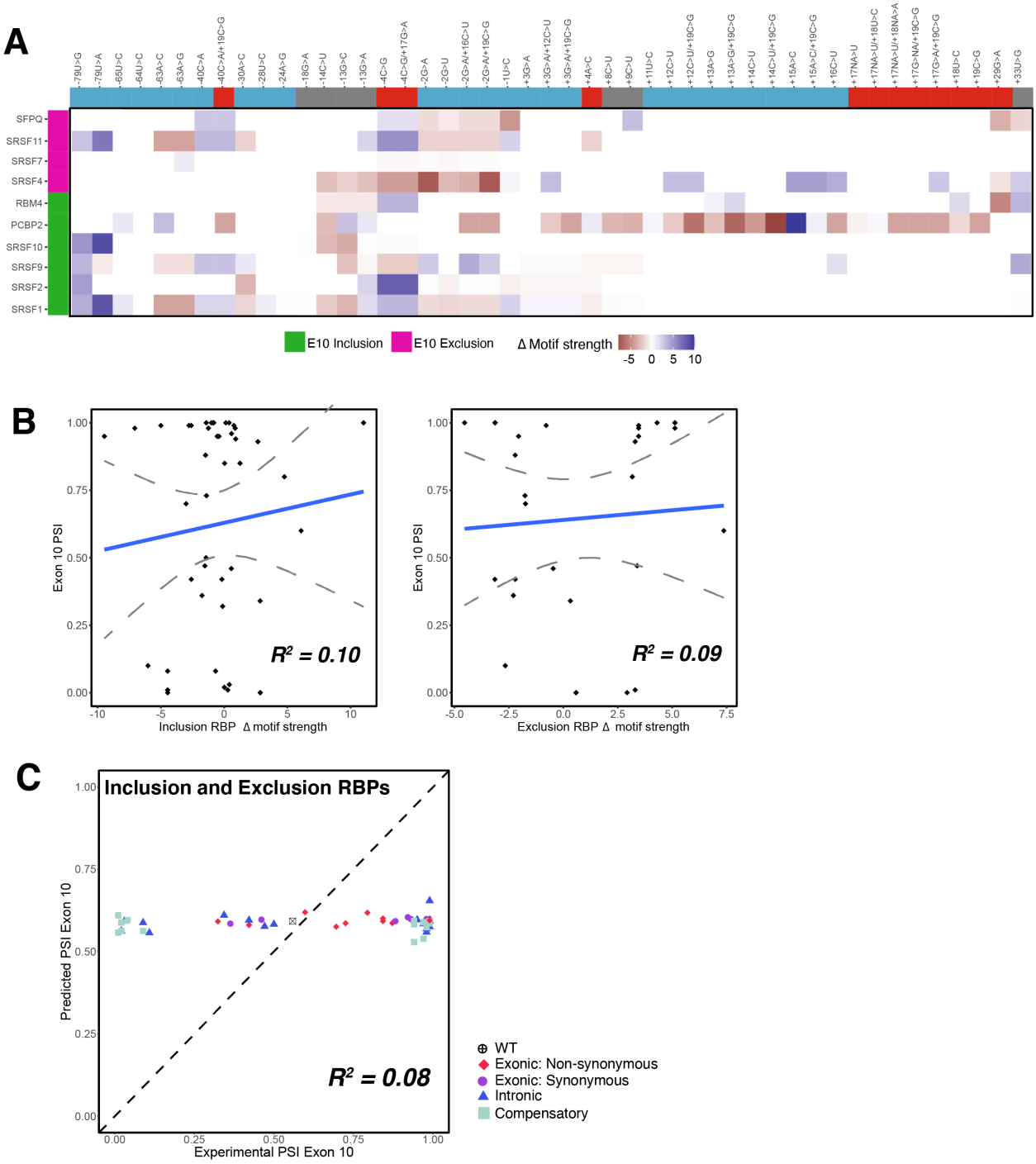


Figure 4-figure supplement 2: RBP binding motif strength is a poor predictor of Exon 10 PSI for all mutations

A) Heatmap of relative RBP binding motif strengths compared to wild type for 44 mutations. A value of 0 indicates that the mutation does not change RBP binding motif strength, a positive value indicates increase in RBP binding motif strength,

1632 and a negative value indicates weaker strength. RBPs implicated in the
1633 regulation of Exon 10 splicing were collected from Qian & Liu, 2014 and the RBP
1634 binding motifs were from Dominguez et al., 2018 and Ray et al., 2013. RBPs on
1635 the left, implicated in the splicing inclusion of MAPT Exon 10, are highlighted in
1636 pink, and RBPs involved in the exclusion of Exon 10 are highlighted in green.
1637 Mutations are marked based on whether they promote the 3R or 4R isoform ratio
1638 or the ratio remains 50:50.

1639 B) Scatter plot displaying change in RBP motif strength versus Exon 10 PSI,
1640 categorized based on whether the RBP is implicated in exclusion or inclusion of
1641 Exon 10. Neither correlation coefficient is statistically significant.

1642 C) Exon 10 PSIs of 44 mutations and wild type predicted using change in RBP motif
1643 strength and plotted against experimental PSIs measured in splicing assays.
1644 Exon 10 PSIs predicted using Eq. 5. Each point represents a mutation and is
1645 colored by category of mutation. Dotted diagonal line is the x=y line, and the
1646 closer the points are to the diagonal, the more accurate the prediction. Pearson
1647 correlation coefficient (R^2) of experimental to predicted PSIs was calculated.

1648

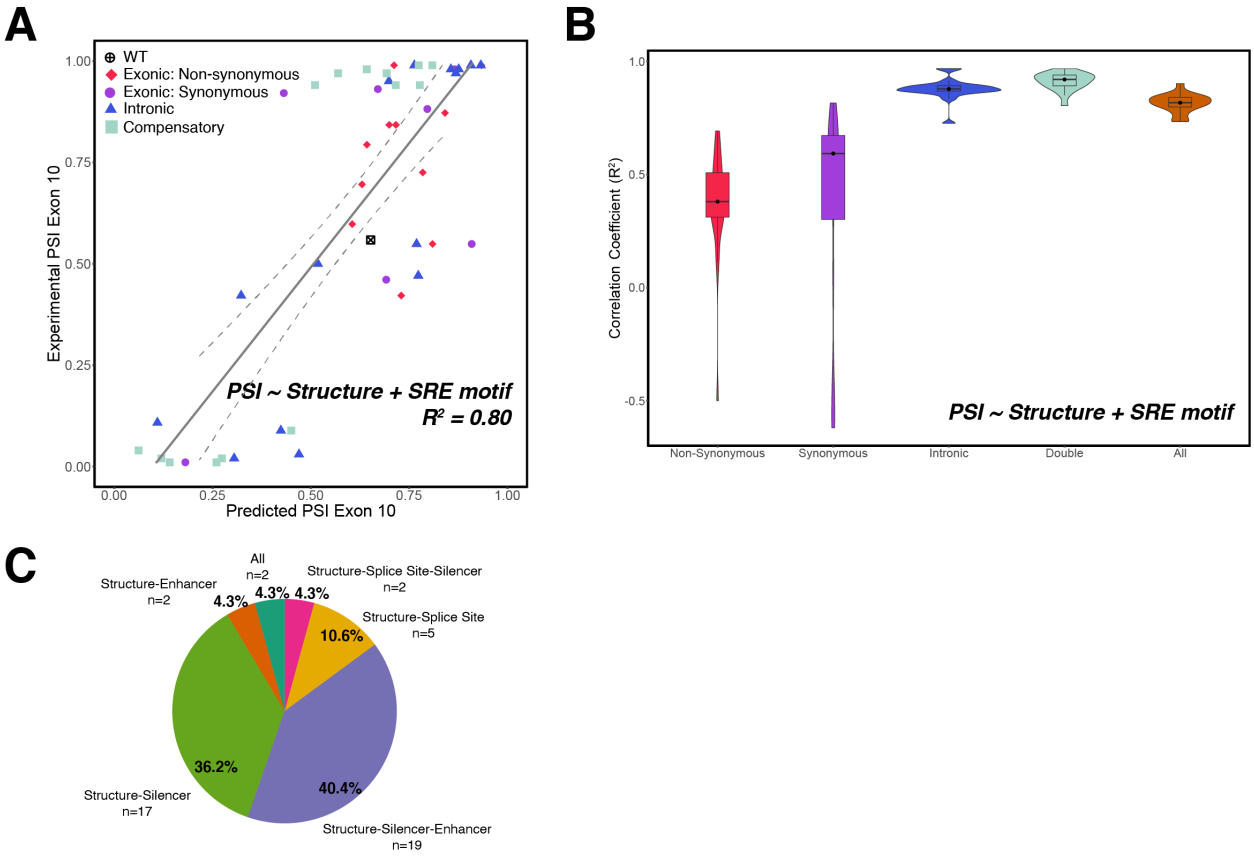
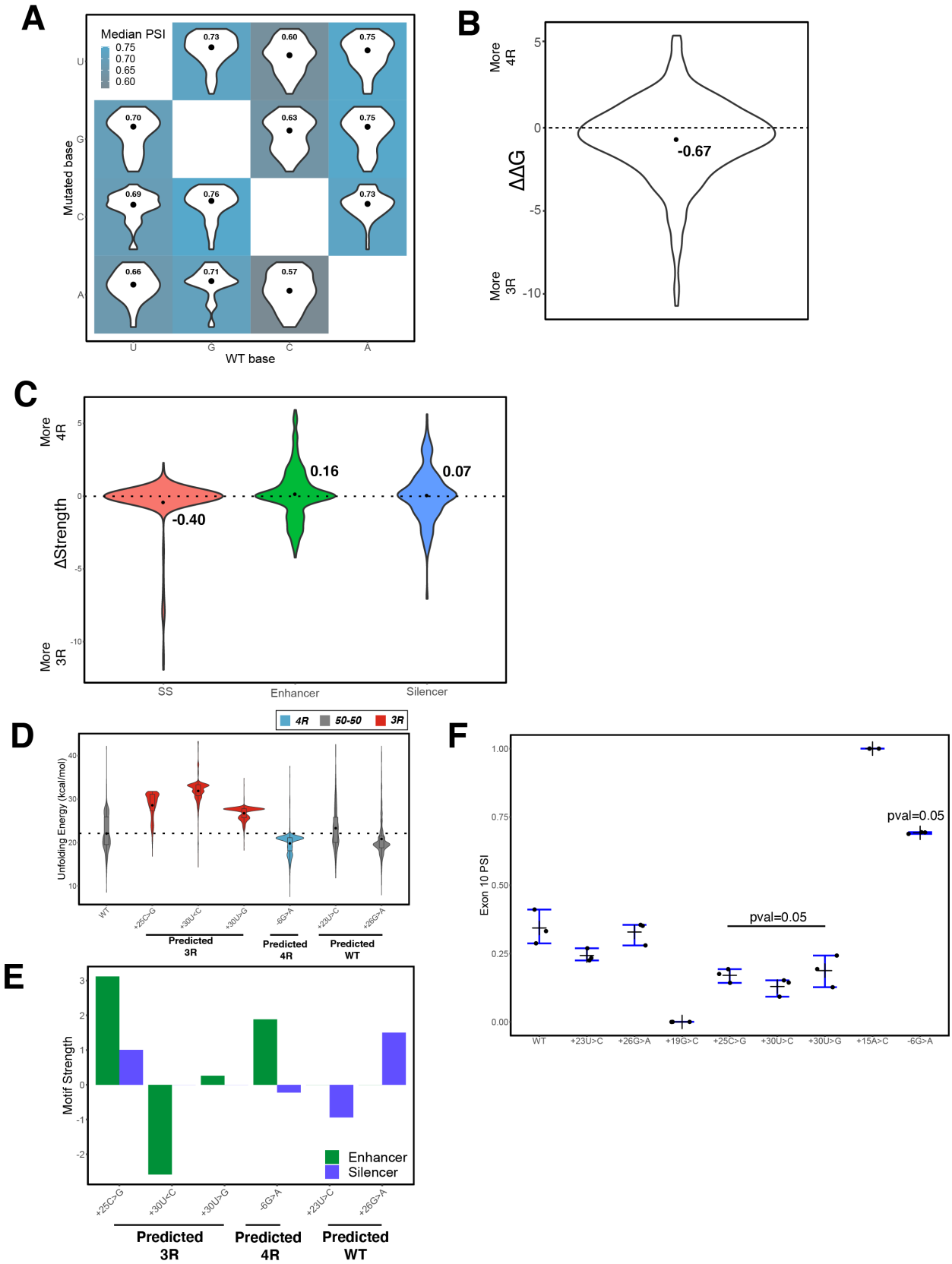


Figure 5-figure supplement 1: Additive model of structure and SRE has poorer predictive performance compared with an interactive model specifically for synonymous and non-synonymous mutations

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

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

1664 C) Pie chart of number and proportion of experimentally validated mutations in each
1665 cluster for heatmap in Fig 5B. Color of segment of pie chart matches up to the
1666 color of dendrogram branch in Fig 5B.


A
Median PSI
0.75
0.70
0.65
0.60
Mutated base
WT base
B
More 4R
More 3R
ΔΔG
-0.67
C
More 4R
More 3R
ΔStrength
-0.40
0.16
0.07
SS
Enhancer
Silencer
D
4R
50-50
3R
Unfolding Energy (kcal/mol)
Predicted 3R
Predicted 4R
Predicted WT
E
Motif Strength
Enhancer
Silencer
Predicted 3R
Predicted 4R
Predicted WT
F
Exon 10 PSI
pval=0.05
pval=0.05
WT
+23U>C
+26G>A
+19G>C
+25C>G
+30U>C
+30U>G
+15A>C
-6G>A

1667

1668 **Figure 6-figure supplement 1**: Complete mutagenesis of 100-nucleotide window
1669 spanning Exon 10-Intron 10 junction

1670 A) Heatmap of mean predicted Exon 10 PSIs grouped by wild type and mutant
1671 nucleotide. Mutations were grouped by wild type and mutant nucleotide, and
1672 mean predicted PSIs were calculated by group and colored according to color
1673 scale. Violin plots of the distribution of PSI per group are shown in tile
1674 corresponding to group. On each tile, mean PSI is indicated by dot and labeled
1675 within violin plot.

1676 B) Violin plot of the distribution of normalized change in unfolding free energy of the
1677 exon-intron pre-mRNA in the spliceosome B^{act} complex from WT for all mutations
1678 around a 100-nucleotide window of exon-intron junction. Mean of -0.67 is
1679 indicated by dot. Dotted line represents the 0 value where there is no difference
1680 between WT and mutant unfolding free energy. Positive values imply region
1681 becomes less structured and has increased inclusion of Exon 10 (4R isoform);
1682 negative values are interpreted as more structured and decreased inclusion of
1683 Exon 10 (3R isoform).

1684 C) Violin plots showing the distribution of normalized change in splice site,
1685 enhancer, and silencer strength compared with WT for all mutations spanning a
1686 100-nucleotide window of exon-intron junction. Mean is indicated by large black
1687 dots on violin plots. Dotted lines represent the 0 value where there is no
1688 difference from WT strength for mutation. Positive values suggest increased
1689 inclusion of Exon 10 (4R isoform), whereas negative values are interpreted as
1690 decreased inclusion of Exon 10 (3R isoform).

1691 D) Violin plot shows the distribution of unfolding free energy of the exon-intron pre-
1692 mRNA in the spliceosome B^{act} complex for the 1000 structures in the ensembles
1693 of wild type and the 6 VUSs experimentally tested. Each violin plot is colored by
1694 whether the mutation promotes the 3R or 4R isoform ratio or the ratio remains
1695 50:50. The dotted line indicates the median unfolding free energy of the WT
1696 ensemble.

1697 E) Bar plots display the change in enhancer and silencer strength of the 6 VUSs
1698 compared with WT.
1699 F) Quantification of Exon 10 PSI of three replicates for splicing assay gels for 6
1700 VUSs. One tailed Wilcoxon Rank Sum test was used to calculate significance of
1701 Exon 10 PSI of VUS of interest compared to WT.
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726

1727 # Supplementary Files

1728

1729 **Supplementary file 1**: ANOVA table for between individuals and within individuals
1730 Exon 10 PSI comparison

1731

1732 **Supplementary file 2**: Details on 47 experimentally tested *MAPT* mutations used in
1733 training model

1734

1735 **Supplementary file 3**: Details on 55 variants of unknown significance (VUSs) in *MAPT*
1736 from dbSNP

1737

1738 **Supplementary file 4**: Primers used for amplification of exon-exon or exon-intron
1739 junctions

1740

1741 **Supplementary file 5**: Re-calculated Position Weight Matrices for ESEs, ESSs, ISEs,
1742 ISSs

1743

1744 **Supplementary file 6**: Details on beta regression model results and features used for
1745 each training and test set

1746

1747 **Supplementary file 7**: Gel of RT-PCR data for splicing assay for new WT VUSs

1748

1749

1750

1751

1752

1753

1754

1755