Genome Biology

RESEARCH

Open Access

Decoding exon inclusion in the human brain reveals more divergent splicing mechanisms in neurons than glia

Lieke Michielsen[1,2,3,4], Justine Hsu[3,4], Anoushka Joglekar[3,4,5], Natan Belchikov[3,4], Marcel J. T. Reinders[1,2], Hagen U. Tilgner[3,4*] and Ahmed Mahfouz[1,2*]

*Correspondence: hut2006@med.cornell.edu; a.mahfouz@lumc.nl

[1] Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
[2] Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands
[3] Center for Neurogenetics, Weill Cornell Medicine, New York, NY, USA
[4] Feil Family Brain and Mind Research Institute, Weill Cornell Medicine, New York, USA
[5] New York Genome Center, New York, NY, USA

Abstract

Background: Alternative splicing contributes to molecular diversity across brain cell types. RNA-binding proteins (RBPs) regulate splicing, but the genome-wide mechanisms underlying cell-type-specific splicing remain poorly understood.

Results: Here, we want to unravel cell-type-specific splicing mechanisms by using RBP binding sites and/or the genomic sequence to predict exon inclusion in neurons and glia as measured by long-read single-cell data in the human hippocampus and frontal cortex. We found that exon inclusion of variable exons is harder to predict in neurons compared to glia in both brain regions. Comparing neurons and glia, the position of RBP binding sites in alternatively spliced exons in neurons differ more from non-variable exons indicating distinct splicing mechanisms. Model interpretation pinpointed RBPs, including QKI, potentially regulating alternative splicing between neurons and glia. Finally, we accurately predict and prioritize the effect of splicing QTLs.

Conclusions: Our results indicate that the splicing mechanisms in variable exons in neurons diverged more from the standard mechanisms. Splicing in neurons might be less sequence-dependent and influenced more by, for instance, chromatin accessibility or methylation. Taken together, these results highlight new insights into the mechanisms regulating cell-type-specific alternative splicing in the brain.

Keywords: Alternative splicing, Brain, Prediction models, Long-read single-cell sequencing

Background

During RNA splicing, introns are removed from the precursor mRNA. Different combinations of exons result in different mRNA isoforms, which may differ in function [1–3]. This mechanism, called alternative splicing, contributes to the complexity of human tissues and cell types; approximately 95% of all human genes are believed to be spliced in multiple ways

[4, 5]. Across different tissues, the brain has the highest levels of exon skipping and one of the most distinctive patterns of alternative splicing [6].

Alternative splicing is partly regulated by RNA-binding proteins (RBPs) [7, 8], which can activate or inhibit spliceosome assembly or splice site recognition. RBFOX proteins, for instance, instruct neuronal differentiation by regulating splicing of *NIN* which in turn affects the localization of the corresponding Ninein protein [9, 10]. Additionally, splicing regulation often relies on the combinatorial binding of multiple RBPs. For example, the inclusion of exon 9 of *Gabrg2* depends on the binding of RBFOX and NOVA [11]. Splicing simulators have considered splicing enhancers and silencers [12] and a splicing code for tissue-dependent splicing has been elaborated [13–15]. However, the genome-wide mechanisms regulating splicing across different cell types remain largely unknown.

Long-read sequencing is an emerging technology that has contributed enormously to RNA biology since its inception [16–20]. Long-read single-cell and single-nuclei sequencing in fresh [21, 22] and frozen [23] tissue allows studying alternative splicing at the cell-type level in the brain and other complex tissues. Such analyses revealed that most mouse genes show differential isoform expression across at least one pair of cell types, regions, and/or developmental time points in the brain [24, 25]. In accordance with prior studies [26–28], single-nuclei isoform RNA sequencing (SnISOr-Seq) of the human adult frontal cortex revealed that exons associated with autism spectrum disorder (ASD) are variably included across cell types [23].

To understand (alternative) splicing mechanisms and the influence of RBPs, several computational methods have been developed. AVISPA, for instance, predicts alternative splicing in four tissues by extracting regulatory features, such as the length of the exon or the presence of RBP binding sites, from the mRNA sequence [14]. Other methods, including SpliceAI, DNABERT, Pangolin, and MTSplice, directly use the pre-mRNA sequence as input to their models [29–32]. However, none of the current methods can predict cell-type-specific alternative splicing genome-wide, which is crucial for understanding splicing in heterogeneous tissues such as the brain.

Here, we aim to unravel cell-type-specific alternative splicing mechanisms in the brain. Specifically, we studied the factors driving differential inclusion of variable exons—exons with inclusion rates that differ between neurons and glia—within the hippocampus and frontal cortex. To achieve this, we trained computational models to predict cell-type-specific exon inclusion using the pre-mRNA sequence and/or the presence of RBP binding sites. Using model interpretation, we examined the regulatory mechanisms underlying this splicing specificity. We found that the presence of RBP binding sites in variable exons compared to non-variable exons differs more in neurons than in glia. This indicates that the alternative splicing mechanism in neurons deviates more from the non-variable mechanism. Furthermore, we show that some RBPs, including QKI, have a big effect on exon inclusion in glia, that the regions close to the splice sites are most important for predicting exon inclusion, and that we can correctly predict and prioritize the effect of splicing QTLs and prioritize their effects.

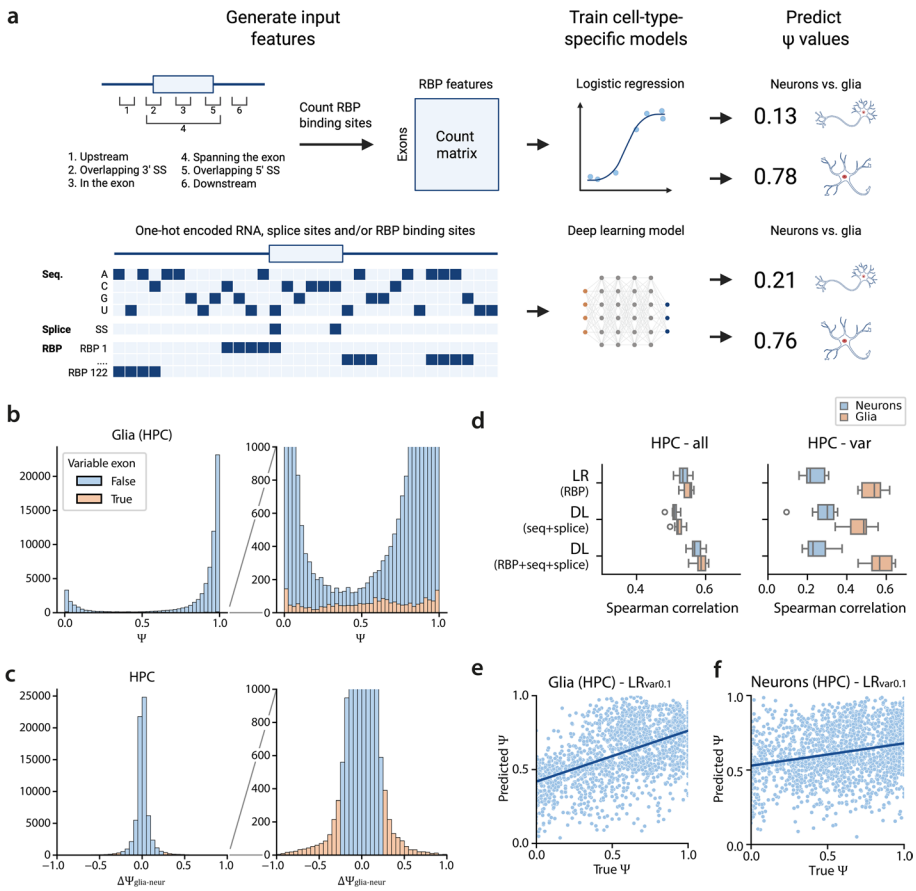

Fig. 1 Overview and performance of the Ψ prediction models. **a** Schematic overview of the models used to predict cell-type-specific Ψ values. **b** Distribution of Ψ values of glia in the hippocampus. **c** Distribution of $\Delta\Psi_{glia-neur}$ for the hippocampus. **d** Performance of the different models during tenfold cross-validation on all exons and the variable exons in glia and neurons in the hippocampus. **e**, **f** Scatterplot showing the predictions of $LR_{var0.1}$ for variable exons in glia and neurons

Results

Using ENCODE-derived RBP-binding sites to predict exon inclusion is more difficult in neurons than in glia

To define the rules governing exon inclusion in distinct cell types, we trained different models to predict cell-type-specific percent spliced-in (Ψ) values in the brain (Fig. 1A). We focused on neurons and glia in two human brain regions, hippocampus (HPC) and frontal cortex (FC), and calculated Ψ values per exon by aggregating single-nuclei isoform RNA sequencing (SnISOr-Seq) reads from multiple individuals (Table 1, Methods) [23, 25]. Most exons are either almost always included ($\Psi \approx 1$) or excluded ($\Psi \approx 0$) in

Table 1 The number of measured exons (exons for which at least 10 reads were sequenced in both the neurons and glia) and variable exons ($|\Delta\Psi_{glia-neur}| > 0.25$) in the hippocampus (HPC) and frontal cortex (FC). Numbers between brackets indicate the number of genes for the studied exons

	Individuals	Measured exons (genes)	Variable exons (genes)
HPC [25]	6	68,215 (8,887)	2,244 (1,430)
FC [23]	2	56,427 (8,137)	943 (703)

an mRNA molecule (Fig. 1B, S1A-C). Furthermore, most exons have similar values in neurons and glia (Fig. 1C, S1D). We define exons with different inclusion rates in neurons and glia ($|\Delta\Psi_{glia-neur}| > 0.25$) as variable exons. In HPC and FC, 2,244 and 943 exons are variable respectively (Table 1). In contrast to non-variable exons, these values show a uniform distribution of Ψ (Fig. 1B). Even though we used a minimum of 10 reads per exon to calculate a Ψ value (Methods), we believe these values are reliable. We downsampled the number of reads to 10 reads for variable exons with > 100 reads in both neurons and glia. Calculating the exon inclusion using the original and downsampled reads results in very similar Ψ values (Fig. S2). Furthermore, when comparing the Ψ values of the variable exons per individual in neurons and glia, there is a clear separation between neurons and glia (Fig. S3). Since most exons are almost always included, we downsampled these non-variable exons when training the models (Methods).

First, we used a logistic regression (LR) model to predict Ψ values from RBP binding sites of 122 RBPs from the ENCODE project [8]. These RBPs were measured in two cell lines (K562, HepG2), implying that this data is not brain cell-type-specific. We generated a count matrix, indicating the number of binding sites per exon for each RBP. Since the position of an RBP can influence its function [33, 34], we split these binding sites based on six possible binding locations: 1) upstream of the exon (up to 400 bp), 2) overlapping the 3' splice site, 3) in the exon, 4) spanning the exon, 5) overlapping the 5' splice site, and 6) downstream of the exon (up to 400 bp) (Fig. 1A).

Any model is strongly influenced by its training data. A model trained on all exons might be dominated by the rules governing non-variable exons, while cell-type-specific inclusion effects might be overlooked. Therefore, we trained three different models using tenfold cross-validation and either: A) all exons (LR_{all}), B) exons with $|\Delta\Psi_{glia-neur}| > 0.1$ ($LR_{var0.1}$), or C) exons with $|\Delta\Psi_{glia-neur}| > 0.25$ ($LR_{var0.25}$) as training data (Table S1). When evaluating the models on all exons, LR_{all} showed the highest median Spearman correlation between true and predicted Ψ values on all four datasets followed by $LR_{var0.1}$ and $LR_{var0.25}$ (Fig. 1D, S4). On hippocampal variable exons, however, $LR_{var0.1}$ outperformed the other models. The performance increase when training on variable exons indicates that the splicing mechanism in these variable exons is somewhat different from the mechanism in non-variable exons. In the frontal cortex, the performance on neurons increased when the training data became more specific, while the performance on glia decreased (Fig. S4). Surprisingly, we predicted Ψ values more accurately in glia than neurons in both brain regions (median Spearman correlation of 0.54 vs. 0.23 in HPC, and 0.57 vs. 0.10 in FC) (Fig. 1D-F, S4). Furthermore, using $LR_{var0.25}$ to predict Ψ values of all exons resulted in lower performance for neurons compared to glia in both HPC and FC (Fig. S4). This suggests that the splicing patterns learned for variable exons in neurons do not extend to non-variable exons, likely due to differences in the underlying molecular grammar between the two exon sets.

Interestingly, the number of RBP binding sites and Ψ value is positively correlated in non-variable exons (Spearman correlation = 0.17 in glia, 0.16 in neurons). However, this correlation decreases in variable exons in neurons (Spearman correlation = 0.11), while it remains more stable in variable exons in glia (Spearman correlation = 0.18). Also, variable exons with a higher Ψ value in glia have slightly more RBP binding sites compared to variable exons with a higher Ψ value in neurons (Fig. S5), but this difference is not

statistically significant (two-sided Wilcoxon rank-sum test, p-value = 0.18). This trend may partially be explained by the absence of important neuron-specific RBPs, such as NOVA1 and NOVA2 [11, 34, 35], in the ENCODE eCLIP dataset.

Primary sequence is more informative for neurons

The RBP binding sites used to train the logistic regression models were measured in immune and liver cancer cell lines and are thus not cell-type specific—and may reflect glial more than neuronal splicing as shown above. To test whether this caused the low performance of the models on neurons, we trained sequence-based models—which are independent of any RBP data and comparable across different cell types. We adapted the Saluki model, a hybrid convolutional and recurrent neural network that uses mRNA sequences to predict mRNA degradation rates [36], to predict Ψ values (Methods) (Fig. 1A, S6). The input sequence is 6,144 bp with the exon of interest centered in the middle. Since deep learning models need large training datasets, we trained a model using all exons ($DL_{all\text{-}seq}$) and a model using exons with $|\Delta\Psi_{(glia-neur)}| > 0.1$ ($DL_{var0.1\text{-}seq}$).

In HPC, the LR_{all} model outperformed the $DL_{all\text{-}seq}$ model when evaluating performance on all exons (Fig. 1D). However, on variable exons in neurons, $DL_{all\text{-}seq}$ outperformed $LR_{var0.1}$, the best model for neurons so far (median Spearman correlation of 0.29 vs. 0.23). For the variable exons in neurons, primary sequence is more informative than the measured ENCODE-derived RBP-binding-site data. Even though the performance increases for neurons, the performance gap between neurons and glia remains. Thus, neuronal splicing patterns probably have more complex regulation mechanisms that we do not capture with the current models. In FC, the performance of the DL models on all exons and variable exons was considerably lower compared to HPC (Fig. S7). This is likely related to the size of the training data which is significantly smaller for FC than HPC, which also explains the low performance of the $DL_{var0.1\text{-}seq}$ models (Table S1).

Next, we combined sequence and RBP binding sites by adding a channel for every RBP which indicates the presence of a binding site ($DL_{all\text{-}seq\text{-}RBP}$) (Fig. 1A, S6). For glia, this outperformed the LR models and resulted in the best-performing model (median Spearman correlation of 0.54 vs. 0.57 in HPC, and 0.57 vs. 0.65 in FC) (Fig. 1D, S7). This improvement indicates that we can capture regulatory information from sequence beyond those present in RBP data alone. For neurons, however, $DL_{all\text{-}seq\text{-}RBP}$ had lower performance than $DL_{all\text{-}seq}$, again confirming that the ENCODE RBP data is more informative for glia than neurons.

Next, we trained DL models that do not use splice sites or only use RBPs as input for the neurons and glia in HPC to understand how the input channels affect performance (Fig. S8). Omitting splice sites only slightly decreased the performance, which indicates that the model can recognize the splice sites quite easily from the sequence itself. For glia, using the RBPs as the only input feature results in a comparable performance to the LR_{all} model (median Spearman correlation of 0.55 vs. 0.54) and an even better performance than sequence and splice sites only (median Spearman correlation of 0.49). However, for neurons, we observe the opposite; using RBP binding sites reduces performance compared to the $DL_{all\text{-}seq}$ model (median Spearman correlation of 0.23 vs. 0.30).

Exon inclusion mechanisms are conserved between human and mouse

As cell-type-specific alternative splicing is partially conserved between humans and mice [25], we hypothesized that adding mouse data to our model would increase performance. We combined human HPC data with mouse HPC [25]. Since mouse FC data is not available, we combined human FC with data from the mouse visual cortex (VisC). While these two regions are not identical, both lie within the neocortex. Especially in mouse HPC, few exons are variable (528) compared to VisC (1,404) (Table S2, Fig. S9). Although $DL_{all\text{-}seq\text{-}RBP}$ performed best in glia, we only trained models with sequence and splice sites as input channels ($DL_{all\text{-}seq\text{-}m}$) since RBP binding sites were not measured in mouse cell lines. In HPC, the performance on variable exons of both cell types slightly increased by adding the mouse data (Fig. S10A). On FC, the performance on all exons increased as well (Fig. S10C), supporting our hypothesis that not enough training data was available to train these models on human exons alone. Similar to the human data, glial Ψ values were easier to predict than neuronal ones in mice (Fig. S10B, D).

Cell line data supports divergence of neuronal splicing regulation

Both the human and the mouse data, however, are generated using a similar protocol (ScISOr-seq or SnISOr-seq) [21, 23]. To ensure that this doesn't bias our conclusions, we trained our models as well on bulk long-read sequencing data from PGP1-derived astrocytes and PGP1-derived excitatory neurons from ENCODE4 [37]. Compared to the hippocampus and frontal cortex data, the ENCODE4 data contains considerably less variable exons (205 compared to e.g., 2,244 in the hippocampus data) (Table S3), which could be explained by the fact that these cell lines are not mature and interactions with other cell types are missing [38, 39]. For the LR models, we observe a similar pattern as in the hippocampus and frontal cortex data: performance on variable exons is higher in astrocytes than in excitatory neurons (Fig. S11). Furthermore, also consistent with earlier findings, the $LR_{var0.25}$ model shows a reduced performance on all exons in excitatory neurons. A performance drop is also seen in astrocytes, but it is less pronounced. However, the DL models were more challenging to train and showed reduced performance across all exons, especially for models using only sequence and splice site features, most likely due to the smaller training set. Overall, the ENCODE4 data supports our hypothesis that splicing regulation in variable exons of neurons diverged more from standard splicing mechanisms.

Microexons and frame consistent exons are more included in neurons

We input RBP binding sites and the sequence to our models to learn the regulatory mechanisms of exon inclusion. However, other factors, such as exon length or intron size, may affect exon inclusion. Since correlations between model predictions vary considerably (Fig. S12-13), each model might perform well on a distinct set of exons.

For example, alternatively spliced microexons ($<=27$ bp) are more often included in neurons compared to other tissues [26]. We observe the same trend in our hippocampus data: of the 82 variable microexons, 80.5% show a higher Ψ value in neurons.

Inclusion of variable exons in other length groups (28–100, 101–200, > 200 bp), is not biased towards neurons or glia (49.9%, 49.3%, and 47.5% respectively).

Therefore, we assessed whether 1) exon length, 2) frame consistency (i.e., whether exon length is a multiple of three), 3) upstream/downstream intron length, and 4) the number of exons within the input window affect model performance on variable exons in the hippocampus dataset. Since some subgroups are rather small (e.g., the microexons), we evaluate using mean squared error (MSE) instead of Spearman correlation.

First, the models' performance on variable exons in glia is very similar for all length groups (Fig. S14A). For neurons, however, models trained on all exons ($DL_{all\text{-}seq}$, $DL_{all\text{-}seq\text{-}RBP}$, LR_{all}) exhibit a higher MSE compared to models trained on variable exons. Second, variable exons are enriched for frame consistency (49.6% frame consistent in variable vs. 37.3% in non-variable exons). Frame consistent variable exons tend to have a higher Ψ value in neurons (58.1% is more included in neurons), while for non-frame consistent variable exons, we observe the opposite trend (43.5% more included in neurons). More inclusion of frame consistent exons appears to be a neuron-specific characteristic and might explain the models' slightly lower performance on this group in neurons (Fig. S14B). Models trained on variable exons ($LR_{var0.1}$ and $LR_{var0.25}$) seem to learn this specific characteristic better and show a reduced performance gap between the two groups. Similarly, $DL_{all\text{-}seq}$ exhibits slightly less variability indicating that the sequence is more informative for learning this characteristic compared to combining sequence and RBP binding sites. In glia, we observe no difference in model performance between frame consistent and non-frame consistent exons. Third, we tested the effect of intron length (< 101, 101–1000, > 1000 bp) and found no clear differences across most models (Fig. S14C-D). The only exception was $DL_{all\text{-}seq\text{-}RBP}$ in neurons, which performs poorly overall. Finally, we observed the same for the number of exons within the input window (Fig. S14E). We only tested this for DL models since the input window of the LR models was limited to 400 bp upstream and downstream of the exon.

Decreased performance of the models trained on all exons on neuron-specific groups (i.e., variable microexons and frame-consistent exons) supports our hypothesis that the splicing mechanisms for variable exons in neurons diverged from non-variable exons. Especially exon length affects the MSE, although the number of microexons in the data is low (only 4.5%) and thus limiting the effect on overall performance.

We also checked whether the models trained on all exons are skewed towards predicting extreme values ($\Psi = 0$ or $\Psi = 1$) due to this bias in the training data (Fig. 1B). We assessed the performance of LR_{all}, $DL_{all\text{-}seq}$, and $DL_{all\text{-}seq\text{-}RBP}$ on non-variable exons with intermediate Ψ values ($0.3 < \Psi < 0.7$) in the hippocampus dataset (1896 exons in glia, 1798 in neurons). Especially $DL_{all\text{-}seq\text{-}RBP}$, and to a lesser extent $DL_{all\text{-}seq}$, mainly predict values close to 0 or 1 (Fig. S15). LR_{all} is less skewed, but still has low predictive power on exons with intermediate Ψ values.

Adding more data does not improve model performance

To assess how model performance scales with additional data, we included long-read single-nucleus data from the frontal cortex of 10 control individuals [40]. This data was generated using a custom-designed enrichment array for 3,630 genes related to various brain diseases and synaptic function. As a result, this data contains less measured exons

compared to the full transcriptome hippocampus and frontal cortex data. But, for the targeted genes, we can evaluate how adding more individuals (and thus increasing the read depth) affects model performance. The number of reads increases when adding more individuals, but the read depth varies per sample (Fig. S16A). The number of measured exons (i.e., exons with at least 10 reads for both neurons and glia) and the number of variable exons increase when more individuals are added, but seem to saturate after five individuals (Fig. S16B-C). We trained the $LR_{var0.1}$ model on data subsets of increasing size. Interestingly, model performance on all exons decreases slightly as more individuals are added, while performance on the variable exons remains stable (Fig. S16D-E). This small decrease may be due to a reduction in the fraction of exons with binary PSI values (i.e., PSI = 0 or 1) (Fig. S16F), which are easier to predict. In summary, increasing the data increases the number of measured exons and variable exons but does not affect model performance.

Splicing mechanisms for variable exons in neurons diverged more from the mechanisms of non-variable exons compared to those in glia

Our above results show that neuronal Ψ values are harder to predict than glial regardless of model or input data. Hence, splicing mechanisms in neurons might be different than in glia and more complex. However, Ψ values could be biased, making it easier to predict in glia. To exclude the latter, we used the hippocampus data to assess whether glia and neurons are similar in terms of 1) Ψ-value distributions, 2) heterogeneity within each cell type, and 3) variation across individuals.

First, comparing Ψ distributions, more values are close to 0 or 1 in glia than neurons (Fig. S17AB), which is most apparent for the non-variable exons (two-sided Kolmogorov–Smirnov test, p-value < 2.2e-16). For variable exons (Fig. S17B), however, both distributions are not different (two-sided Kolmogorov–Smirnov test, p-value = 0.44). Thus, data distribution differences cannot explain all observed differences between neurons and glia.

Second, to quantify the heterogeneity within a cell type, we measured the difference in Ψ values between finer cell-type classifications. For neurons, we compared the inhibitory and excitatory neurons, and for glia, we compared oligodendrocytes and astrocytes. Within glia, we have more variable exons ($|\Delta\Psi| > 0.25$) compared to neurons (831 vs. 745). In neurons, more exons have an extreme difference ($|\Delta\Psi| > 0.5$) (92 vs. 70) (Fig. S17CD). Compared to the total exon number defined for both cell types in neurons and glia (28,296 and 27,047 respectively), both numbers are small. Thus, this cannot explain the difference in performance between neurons and glia.

Third, to compare the variance across individuals for glia and, separately, for neurons, we calculated Ψ values per individual instead of using the aggregated counts. We expect that a higher variance across individuals would make the Ψ values harder to predict. We calculated the variance for an exon only if ≥ 3 individuals have ≥ 10 reads for that exon in both neurons and glia. For both non-variable and variable exons, the variance is higher in glia (two-sided paired Wilcoxon signed-rank test, p-value = 1.3e-28 and 8.9e-5 respectively) (Fig. S17E). Thus, the data do not explain observed differences in performance between neurons and glia.

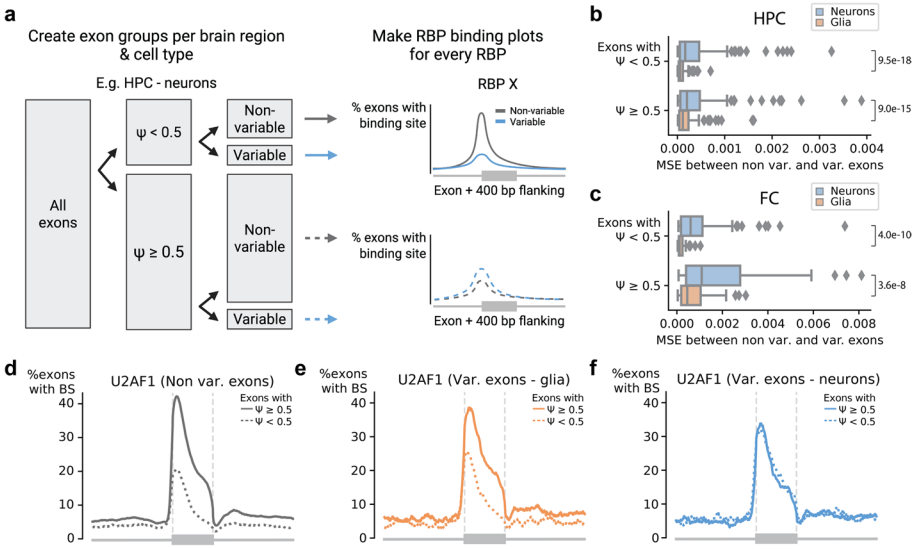

Fig. 2 The difference in RBP binding profiles between non-variable and variable exons. **a** Schematic overview showing how to generate the RBP binding profiles of non-variable ($|\Delta\Psi_{glia-neur}| \leq 0.25$) and variable ($|\Delta\Psi_{glia-neur}| > 0.25$) exons in neurons in the hippocampus. We generated these RBP binding profiles for every RBP and split the exons based on their Ψ value (threshold = 0.5) and their variability. We calculated the mean-squared error (MSE) between the profiles in non-variable and variable exons. We do this for the exons with a high and low Ψ value resulting in four comparisons per RBP. **b**, **c** Boxplot showing the MSE between the RBP profiles in non-variable and variable exons in neurons (blue) and non-variable and variable exons in glia (orange) for the **b** hippocampus and **c** frontal cortex. Every point in the boxplot is one RBP. *P*-values are calculated using a two-sided paired Wilcoxon signed-rank test. **d-f** Binding profile of U2AF1 in **d** non-variable exons in both neurons and glia, **e** variable exons in glia, and **f** variable exons in neurons

We then hypothesized that splicing mechanisms regulating variable exons in neurons might differ from the non-variable exons. To test this hypothesis, we compared the RBP binding profiles between variable and non-variable exons in neurons and glia (Fig. 2A). We performed these comparisons for exons with a high (≥ 0.5) and a low Ψ value (< 0.5) separately. The binding profiles between variable and non-variable exons differ significantly more in neurons compared to glia in HPC (Fig. 2B) and FC (Fig. 2C). Non-variable exons with high Ψ values more often have a binding site at the 3' splice site for splicing factors such as U2AF1, U2AF2, and SF3B4 compared to non-variable exons with low Ψ values (Fig. 2D, S18AB). In glia, variable exons show a similar pattern (Fig. 2E, S18AB). However, binding sites for these splicing factors cannot differentiate between exons with high and low Ψ values in neurons (Fig. 2F, S18AB), indicating that these RBP binding sites are likely not used in neurons.

In the hippocampus, PTBP1 differs the most between neurons and glia (Fig. S18C). PTBP1 acts as a splicing repressor by inhibiting U2AF binding at the 3' splice site [41, 42]. Additionally, position-dependent effects have been described in HeLa cells: binding within or upstream of an exon represses splicing while binding downstream activates splicing in HeLa cells [43]. In our RBP binding profiles, we mainly observe the inhibiting effect of PTBP1 binding the 3' splice site for variable exons in glia. Strikingly, the binding profile of PTBP1 in variable exons in neurons is again considerably different: there is no clear difference between exons with a high and low Ψ. This is consistent with high expression of PTBP1 in glia, while PTBP1 is downregulated during neuronal

differentiation as its neuron-specific paralog PTBP2 (also known as nPTB) is upregulated [42, 44–46]. However, our scRNA-seq data does not support differential PTBP1 expression between neurons and glia (Fig. S19), potentially explained by low correlation between mRNA and protein levels for some genes. In the hippocampus, only one RBP, HNRNPC, showed the opposite pattern with larger differences in glia compared to neurons (Fig. S18D).

Interpretation of LR models reveals cell-type-specific splicing mechanisms

To further pinpoint the factors underlying differences in splicing between glia and neurons, we analyzed the coefficients of the logistic regression models. These coefficients reflect the importance of each RBP binding position in regulating cell-type-specific splicing. We compared the coefficients of four models for the hippocampus (two cell types, and two training sets) and focused on features present in at least 50 exons and with a coefficient > 0.05 in at least one model (191 out of 732 features). The model coefficients first cluster based on which exons are used during training (all vs. variable) (Fig. 3A). This clustering indicates that the mechanisms for non-variable and variable exons, represented by the LR_{all} and $LR_{var0.1}$, differ more than the cell-type-specific mechanisms. The RBPs cluster into two groups: features with positive and features with negative coefficients (Fig. 3A). As expected, splicing repressors, which are part of the

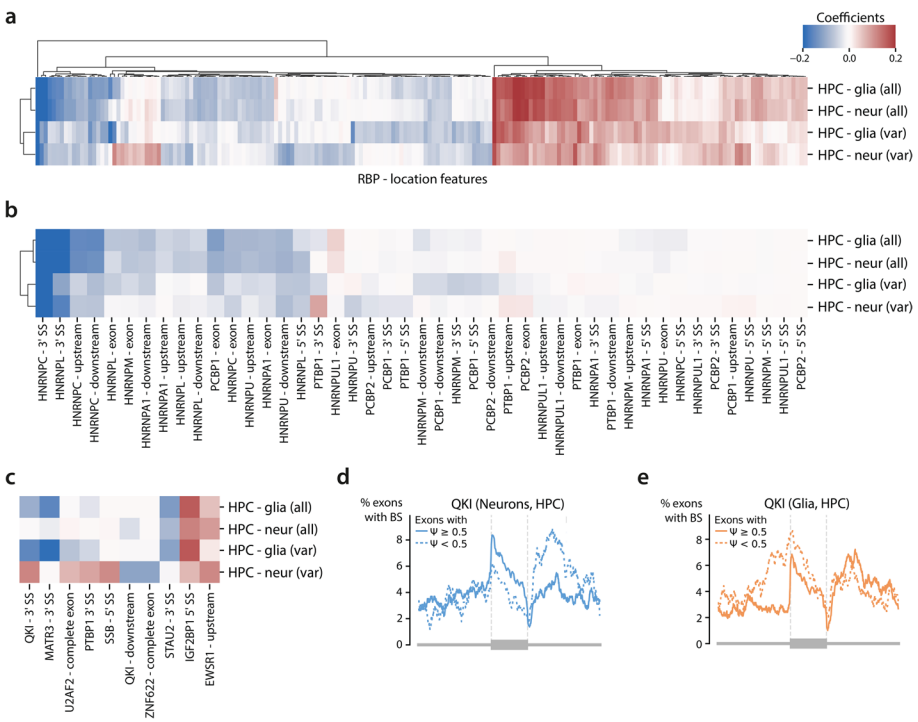

Fig. 3 Interpretation of the logistic regression models. **a** Heatmap showing the coefficients for the RBP-location features in the different logistic regression models. The input features are filtered using a minimum of 50 RBP sites and a value of at least 0.05 in one of the models. The values are clipped between −0.2 and 0.2. **b** Heatmap showing coefficients of hnRNPs in the different models. **c** Heatmap showing the top 10 cell-type-specific input features with the biggest difference between HPC-glia (var) and HPC-neur (var). **d**, **e** Binding profiles of QKI in variable exons in neurons and glia

heterogeneous nuclear ribonucleoproteins (hnRNP) family [47], have a largely negative weight in all models (Fig. 3B). PTBP1, for which we saw a difference between the non-variable and variable exons in the hippocampus, is a member of the hnRNP family and has a potential position-dependent effect in glia based on the RBP binding profiles (Fig. S18C). The $LR_{var0.1\text{-}glia\text{-}HPC}$ model correctly learned this effect: PTBP1 binding at the 3' splice site and within the exon have coefficients of −0.05 and 0.01 respectively. The model coefficient for PTBP1 binding at the 3' splice site is among the ten features that differ the most between glia and neurons (Fig. 3C, $LR_{var0.1\text{-}glia\text{-}HPC}$ vs $LR_{var0.1\text{-}neur\text{-}HPC}$) which indicates a potential cell-type-specific effect corresponding to the established switch between PTBP1 and PTBP2 [42, 44–46]. In glia, PTBP1 is expressed and acts as a splicing repressor by inhibiting the binding of U2AF at the 3' splice site [41, 42], while in neurons PTBP1 is not expressed explaining the opposite learned effect. Furthermore, PTBP1 and MATR3 co-regulate several target genes [48, 49]. Our models indicate that MATR3 binding at the 3' splice site represses exon inclusion in glia, while there is no predicted effect in neurons. This aligns with the known co-regulation of PTBP1 and MATR3. In glia, where PTBP1 is expressed, MATR3 has a regulatory effect, while in neurons, where PTBP1 is absent or lowly expressed, this effect is not observed.

QKI binding at the 3' splice site has the strongest cell-type-specific effect in the hippocampus (model coefficient = −0.15 vs. 0.12 for glia and neurons respectively), which reflects differences in the RBP binding profiles (Fig. 3D-E). In glia, a binding site that overlaps the 3' splice site leads to lower inclusion rates, while the opposite happens in neurons. Besides the 3' splice site, QKI binding downstream of an exon is also in the top 10 cell-type-specific features. QKI's downstream effect contrasts with that of binding at the 3' splice site, indicating a position-dependent effect. Position-dependent regulation has been shown in lung cancer, mouse cardiomyocytes and neural stem cells [50–52]: binding the 3' splice site suppresses exon inclusion, while downstream binding promotes inclusion. This aligns with the learned patterns in glia. Supporting this, QKI is significantly higher expressed in glia than neurons in the hippocampus scRNA-seq data (Wilcoxon rank sum test, adj. p-value < 2.2e-16) (Fig. S20). The opposite effect in neurons can thus be explained by the lack of expression of QKI. In mice, QKI is important during myelination and oligodendrocyte differentiation [53, 54]. Its role in the human brain is less studied, but a role in oligodendrocyte formation and Schizophrenia has been suggested [55, 56]. Interestingly, variable exons are enriched for QKI binding sites compared to non-variable exons (Fisher's exact test, adj. p-value = 1.6e-13).

In contrast to QKI, most of the cell-type-specific RBPs identified using our LR models are neither differentially expressed nor differentially spliced. Exceptions are STAU2, which is upregulated in neurons (Wilcoxon rank sum test, adj. p-value < 3.39e-16), and EWSR1, which is differentially spliced (Table S4). The latter could indicate that distinct isoforms of EWSR1 influence RNA splicing in different ways.

Important RBPs with known functions in regulating splicing in the brain, such as NOVA [11, 34, 35], are missing in the ENCODE data. To learn how NOVA1 and NOVA2 regulate alternative splicing in the brain, we added binding profiles of NOVA1 and NOVA2 based on eCLIP data measured in iPSC-derived motor neurons of three control individuals [57] to our model. However, including these RBPs did not improve model performance (Fig. S21A), potentially due to differences in cell line origin. Nevertheless,

several NOVA-related input features were assigned high weights (Fig. S21B) in line with literature [58]: in neurons, NOVA binding upstream of the exon or the 3' splice site represses exon inclusion, while binding the 5' splice site or downstream of the exon promotes inclusion. However, our model predicts exonic binding of NOVA to promote inclusion as well, whereas literature reports a repressive effect. In glia, the weights are generally smaller or have the opposite effect compared to neurons, supporting the idea that NOVA primarily regulates splicing in neurons but not glia.

The sequence close to the splice sites is most important for predicting exon inclusion

Given that the RBP-binding-site data is not brain-specific and lacked measurements from some key RBPs, we set out to identify sequence features that influence Ψ predictions in the deep learning models. We used *in-silico* saturation mutagenesis (ISM, Methods) to systematically predict how nucleotide substitutions in the input sequence affect the predicted Ψ value [59–62]. Since $DL_{var0.1}$ performed considerably worse than DL_{all} (Fig. 1D), we focused on interpreting DL_{all} for glia in the hippocampus, which had higher prediction accuracy than neurons, instead of looking for cell-type-specific effects.

Since ISM is computationally expensive, we mutated the input sequence of the 9,929 exons with $|\Delta\Psi_{(glia-neur)}| > 0.1$ instead of all exons. The ISM score indicates how much a mutation increases or decreases the predicted Ψ value compared to the average prediction at that position for that sequence (Methods). As expected, mutations around the splice sites, which has to be recognized by the splicing machinery, and within the exon strongly affect the predicted Ψ value (Fig. 4A). Looking at the maximum absolute ISM score, only mutations within a range of 50 bp upstream of the 3' splice site and 150 bp downstream of the 5' splice site have a value > 0.1 (Fig. S22). However, smaller values of > 0.05 could be observed across almost the whole input sequence. Although distant splicing regulators have been reported [63], potential variability in distant motifs and/or their position may prevent their detection by our model.

Besides the region around the exon of interest, we observed higher-than-average ISM scores within nearby exons and their flanking region (Fig. S23). The enrichment of RBP binding sites in these regions could explain the higher scores. Alternatively, our model potentially recognized coordinated events between exons. To test this, we selected the top 10 exons with the highest absolute ISM scores within their neighboring exons and visualized the single-cell long reads from our data that span both exons (see Methods). These reads can inform whether the two exons pair non-randomly (thus in coordination [21, 23, 64, 65]) or randomly. Exon 24 in *XRN2* appeared twice in the top 10 list with two neighboring exons (exons 21 and 22) strongly influencing its Ψ value (Table S5). All three exons (21, 22, and 24) have a Ψ value of around 0.8 and the exons are either all included or all excluded in the single-cell long-read data, suggesting these exons are mutually associated (Fig. 4BC). Mutations affecting the inclusion of one of these exons will most likely affect the other exons as well. In the top 10 scores, four other cases could pinpoint exon coordination events (Fig. S24-27). In the remaining four cases, the exons pair randomly, so there is no evidence of exon-exon coordination (Fig. S28-31).

To further interpret sequences with a high ISM score, we used TF-MoDISco [66] to identify motifs in sequences with large effects on exon inclusion. Since the region around the splice site had the highest ISM scores, many of the top motifs identified by

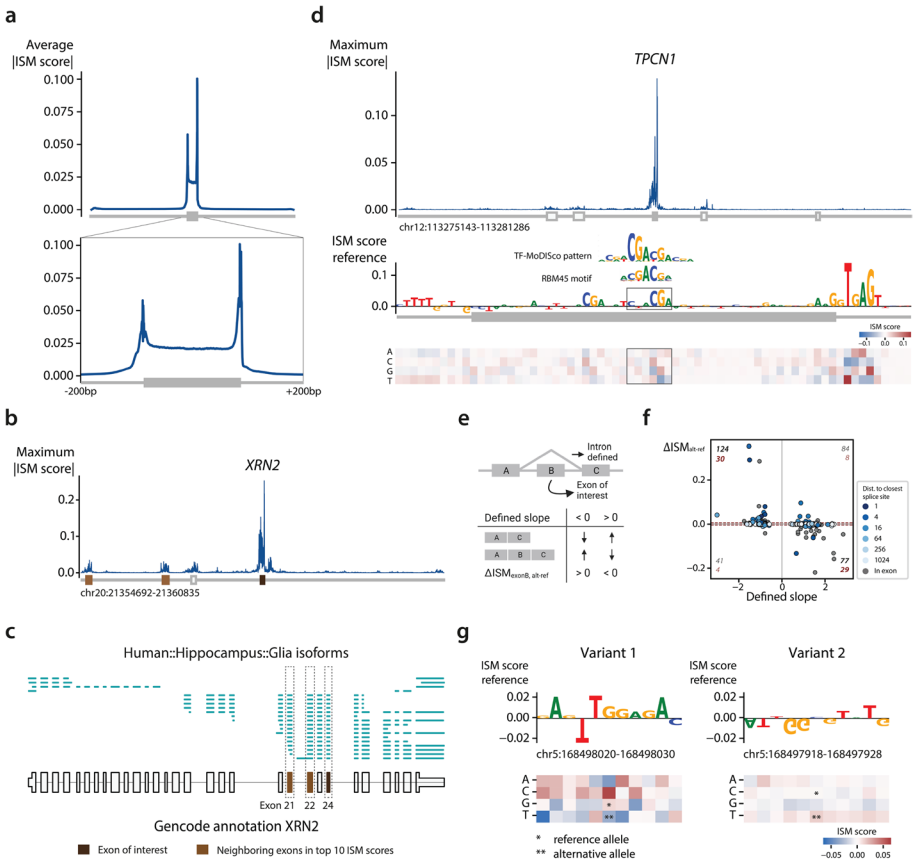

Fig. 4 Interpretation of the deep learning model for glia in the hippocampus. **a** Average absolute ISM score across the 9,929 exons. The mutations within the exons are binned in 300 bins. The zoomed-in plot ranges from 200 bp upstream of the 3′ splice site to 200 bp downstream of the 5′ splice site. **b** Mutation profile for an exon in *XRN2*. The colors of the exons below the profile indicate the exon of interest and the neighboring exons which have an ISM score in the top 10. **c** Single-cell long reads for *XRN2*. Each line is a single cDNA molecule. The bottom black track shows the Gencode annotation. **d** Mutation profile for an exon in *TPCN1*. In the exon, a motif corresponding to RBM45 is found. **e** Schematic overview of the sQTL analysis. **f** Scatterplot showing the predicted effect for each variant. The color of the points indicates the distance to the closest splice site. A grey dot means that a variant falls within the exon of interest. The numbers in black and red indicate the number of predictions in a quadrant when no threshold and a threshold of 0.005 are used respectively. **g** ISM scores for two variants related to the same exon of *RARS1*. A negative effect, corresponding to the positive slope, is predicted for the first variant. A smaller, but positive effect is predicted for the second variant

TF-MoDISco correspond to the consensus splice sites and associated motifs, including the well-known AG acceptor dinucleotide, the poly-pyrimidine tract (PPT) upstream of the exon, and the extended splice donor motif with the GU dinucleotide (Additional file 1, Fig. S32). To assess the consistency between our LR and DL models, we compared RBP motifs identified by TF-MoDISco to learned RBP weights in the LR_{all} and $LR_{var0.1}$ models. In general, TF-MoDISco's predicted direction of effect (positive or negative) agrees with the learned weights in the LR models (Fig. S33). TF-MoDISco assigned both a positive and negative effect to some RBPs, such as PABPC4. This is consistent with the $LR_{var0.1}$ model, where PABPC4 receives a negative weight when binding upstream of the exon but a positive weight when binding downstream. We also found motifs that match known RBP binding motifs, which were not in our input data for the LR model, and

hence could not be tested for cell-type-specific effects. For example, we found a motif corresponding to RBM45 in exon 12 of *TPCN1* (Fig. 4D, Table S5), which seems to promote exon inclusion. RBM45 regulates constitutive splicing and can probably activate or repress the inclusion of an exon, but the exact mechanisms are currently unknown [67]. Taken together, characterizing important sequence features from DL models can identify splicing regulators beyond those we can identify based on available RBP measurements.

Prioritizing the effect of splice QTLs using the DL models

So far, we showed how LR and DL model interpretations can be used to reveal the regulatory mechanisms of RBP governing cell-type-specific exon inclusion. Besides this fundamental knowledge, we can use our DL models to predict the effects of genetic variants on splicing. Accurately predicting these effects can help prioritize variants of interest. To test the relevance of our model predictions for genetic variants, we used splicing quantitative trait loci (sQTLs) from the hippocampus data from GTEx v8 [68]. Variants in this dataset are linked to intron excision ratios instead of exon inclusion. We extracted introns and their corresponding variant(s) that span an exon in our data and predicted the effect of the variant(s) on that exon (Fig. 4E). In total, 326 variants are within the input range of our model. These variants correspond to 122 introns and 158 exons. Some introns thus span multiple exons and most introns have multiple variants linked. For every variant, a slope indicates whether the corresponding intron is excised more or less compared to the reference allele. We expect negative slopes to correspond to an increased Ψ value of the exon of interest which would result in $\Delta ISM_{(alt-ref)} > 0$. Conversely a positive slope would result in $\Delta ISM_{(alt-ref)} < 0$ (Fig. 4E). However, more complex scenarios, such as a variant affecting adjacent exons, may arise as well.

Using our model, we predicted an effect ($|\Delta ISM_{(alt-ref)}| > 0.005$) for 71 out of 326 variants which corresponds to 61 of the 122 introns. For 83% (59 out of 71) of these variants, our model predicts the expected effect correctly (Fig. 4F, S34). Most of the variants with an effect are very close to the splice sites: 74.6% are within the exon or a distance of 15 bp to either the 3' or 5' splice site. These cases thus affect most likely exonic splicing enhancers or the binding of U1 and U2 snRNA. For 14 of 61 introns where our model did not predict an effect, all corresponding variants are outside of the intron itself. Here, the splicing of adjacent exons is most likely altered instead of our exon of interest. For 2 of these 14 exons, all variants are even outside of the gene itself.

Three exons have multiple corresponding variants with a predicted effect. For exon 15 in *ZNF880* (Table S5), three variants have a predicted expected effect. The other two exons, however, have two variants with a contradicting predicted effect. In both cases, the variant with the biggest predicted effect is in line with the slope of the sQTL of the intron. For exon 25 in *RARS1* (Table S5), for instance, variant one is located in the exon (168,498,025; G → T) and variant two is located before the exon (168,497,923; C → T). For variant one, our model predicted the expected effect, while our model predicted the opposite for variant two (Fig. 4G). Variant one, the variant with the biggest and correctly predicted effect, is located in a binding site for SRSF1 according to eCLIP data [8]. RNA recognition motif 2 (RRM2) of SRSF1 interacts with the GGA motif. A G → T mutation in the first nucleotide will thus hinder the binding of SRSF1 [69]. Variant two is located in a stretch of G's. At this location, there's a binding site for ELAVL1, a protein regulating

mRNA stability, and hnRNP family member HNRNPK, which tends to repress splicing [8]. Using the DL models, we can thus correctly predict the effect for most sQTLs and prioritize their effects.

Discussion

We used logistic regression and deep learning models to understand cell-type-specific exon inclusion in human brain samples. Since this is the first attempt to leverage long-read single-cell sequencing data for this task, we can use our models to decipher the grammar underlying cell-type specificity of splicing. Using model interpretation, we pinpointed interesting RBPs, such as QKI, that could drive differential splicing between neurons and glia. Furthermore, we show that the location of RBP binding sites differs more between variable and non-variable exons in neurons compared to glia. This indicates that the splicing mechanisms controlling exon inclusion in neurons are more different compared to the general mechanism.

For most RBPs, RBP binding profiles of non-variable exons with high and low Ψ values showed distinct patterns. Considering U2AF1 for example, exons with a high Ψ value are more likely to have a binding site close to the 3' splice site compared to exons with a low Ψ value. These RBPs behave differently in variable exons in neurons, and for most RBPs the difference between exons with a low and high Ψ value is missing. These features are thus not informative for neurons, which explains the low performance of the logistic regression models on neurons. The U2AF heterodimer, composed of U2AF1 and U2AF2, is believed to bind every polypyrimidine tract and AG dinucleotide in 3' splice site regions [70–72]. Binding may not happen on specific sites repressed by other factors. The potential binding sites are still there, but they might be used by a competing RBP in neurons. Interestingly, most RBPs are not differentially expressed or differentially spliced between neurons and glia. For these RBPs, post-translational modifications, such as phosphorylation, might differ between neurons and glia and could change their function [73, 74]. A limitation of the RBP binding site data is that they were measured in non-brain cell lines and might not always be representative of splicing in neurons or glia. Furthermore, RBPs known to play a crucial role in alternative splicing in neurons, such NOVA1 and NOVA2, are missing from the ENCODE eCLIP data. Adding NOVA1 and NOVA2 binding sites measured in iPSC-derived motor neurons, however, did not improve model performance, which is potentially explained by batch effects between the different eCLIP datasets and different cell lines of origin.

The deep learning models, however, also perform poorly on the variable exons in neurons. The model trained on all exons focuses only on learning the general splicing mechanisms, and the model trained on the variable exons might not have enough training data. In glia, the model trained on all exons performs well on the variable exons. Again indicating that the variable exons in glia follow the rules of the general splicing mechanisms more. The worse performance of the $DL_{all\text{-}seq}$ models on neurons, in combination with the distinct RBP binding profiles, supports our conclusion that the splicing mechanisms in variable exons in neurons diverged from the mechanisms in non-variable exons.

A potential explanation, in line with the diverged RBP binding sites, is that splicing in neurons is less sequence-dependent. Other factors, such as chromatin features and

polymerase speed [75–88], RNA methylation [89–91] as well as other modifications, and transcription factor binding sites [92], influence splicing as well. These features might explain the difference between neurons and glia. Altered chromatin accessibility or RNA methylation, could, for instance, explain why certain RBP binding sites are not used in neurons anymore. Furthermore, neuronal genes—by definition more expressed in neurons—are more susceptible to missplicing [93]. While we did not focus on missplicing, this indicates that splicing mechanisms might be different in neurons. Also, the gene expression of human neurons diverged faster from other primates compared to glia [94]. A similar divergence could have occurred with the splicing mechanisms.

For the deep learning model, we tested the effect of different lengths for the input sequence. Even though all lengths showed a very similar performance, we used a relatively long input sequence (6,144 bp) which had the advantage that we could predict the effect of more mutations. When predicting the effect of sQTLs, however, we predict a strong effect mainly for variants close to the exon of interest. The region close to the splice sites still contributes the most to the predictions. This is in contrast to splice site predictions from SpliceAI, for which an input sequence of 10 kb significantly outperforms 400 bp [29]. SPLAM, however, outperforms SpliceAI while only using 400 bp [95]. However, variants in motifs distant from the exon may still influence splicing decisions. Such motifs might be rare which could prevent the model from detecting them. Similar limitations have been observed for models that predict gene expression. Even though the best-performing model uses a long input sequence (196 kb), only one-third of the receptive field is used during predictions and distal enhancers are not captured by the model [61, 96].

Another possible advantage of a longer input sequence is that it would be possible to look at coordinated events. Exons in the human brain are often mutually associated or mutually exclusive [23, 64, 97–99]. Such events can even be cell-type-specific. For instance, two neighboring exons in *WDR49* are perfectly coordinated in astrocytes only [23]. Using our model, the ISM scores within neighboring exons are higher than the ISM scores of the rest of the sequence. For some exons, these higher scores indeed indicate that there is exon-exon coordination. Since exon-exon coordination is so common, predicting such events might be an interesting direction as it could be more beneficial than focusing on individual exons.

Besides including exon-exon coordination into the models, increasing the cell-type resolution or even including the spatial location of a cell could improve performance. Recently, using spatial isoform sequencing at a cellular resolution, we showed that cell-type-specific splicing changes do not always follow predefined borders [100, 101]. Training a multitask model that makes predictions for each cell, similar to what has been done in the gene expression domain [102], might be necessary to capture such events. However, enough sequencing data is needed to enable training such models.

Conclusions

In conclusion, to increase our understanding of (alternative) splicing in the brain, we trained two types of models to predict exon inclusion in neurons and glia of the hippocampus and frontal cortex. Ideally, these models make perfect predictions such that they can be used in the clinic for predicting the effects of variants or for personal splicing predictions. The performance of our models, however, is not optimal yet. Nevertheless,

we show how model interpretation yields important biological discoveries including the different mechanisms in neurons and glia. This demonstrates the potential of using long-read single-cell data for this task.

Methods

Calculating cell-type-specific Ψ values

For the human hippocampus and frontal cortex data, we combined SnISOr-Seq data from 6 individuals for the hippocampus and 2 individuals for the frontal cortex to calculate cell-type-specific Ψ values (Table 1). For the mouse data, we combined ScISOr-Seq2 data from two mice for the hippocampus and two mice for the visual cortex (Table S2). For the ENCODE4 bulk long-read data, we combined data from two replicates (Table S3).

Scisorseqr (v0.1.9) was used to map and align reads to GRCh38 for human to identify splice sites for each dataset separately [24]. We used IsoQuant (v2.3.0) to correct the splice sites [103]. We corrected the splice sites using the following steps:

1. We ran IsoQuant with parameter -splice_correction_strategy default_ont on the aligned reads. This outputs amongst others a BED file with the corrected reads.
2. We ran scisorseqr's InfoPerLongRead function on the aligned reads to create an AllInfo file which contains the exon and intron chain per read.
3. We ran scisorseqr's correctedBed2Allinfo function using the BED file from step 1 and the AllInfo file created in step 2 as input to create the new AllInfo file with the corrected exon and intron chain.

Using all exons appearing as an internal exon in a read, we calculated:

- The number of long-read molecules containing this exon (both splice sites included): X_{in}
- The number of long-read molecules assigned to the same gene as the exon, which skipped the exon but includes ≥ 50 bases on both sides: X_{out}
- The number of long-read molecules supporting the acceptor of the exon and ending on the exon: X_{accIn}
- The number of long-read molecules supporting the donor of the exon and ending on the exon: X_{donIn}
- The number of long-read molecules overlapping the exon: X_{tot}

Non-annotated exons with one or two annotated splice sites, ≥ 70 bases of non-exonic (in the annotation) bases, were excluded as intron-retention events or alternative acceptors/donors.

We then calculated.

- $$\Psi_{overall} = \frac{X_{in} + X_{accIn} + X_{donIn}}{X_{in} + X_{accIn} + X_{donIn} + X_{out}}$$

- $\Psi_{acceptor} = \frac{X_{in} + X_{accIn}}{X_{in} + X_{accIn} + X_{out}}$
- $\Psi_{donor} = \frac{X_{in} + X_{donIn}}{X_{in} + X_{donIn} + X_{out}}$

If

- $0.02 \leq \Psi_i \leq 0.98$ where $i \in \{overall, acceptor, donor\}$

in the pseudo-bulk data, the exon was kept.

Next, we filtered exons based on the number of reads. We only calculate $\Psi_{overall}$ for a cell type in a certain brain region if at least 10 long-read UMIs are sequenced across the different individuals ($X_{tot} \geq 10$). Since individuals of different datasets were sequenced using a different read depth, we normalized the read counts by dividing it by the total number of reads for an individual before calculating $\Psi_{overall}$. We then calculated $\Psi_{overall}$ for each cell type (Ψ_{neur} and Ψ_{glia}) for the hippocampus and frontal cortex. If there were not enough reads, $\Psi_{overall}$ for that exon and cell type was set to "NA". We used the cell-type labels defined in the original datasets. For neurons, we grouped the inhibitory and excitatory neurons. For glia, we grouped the oligodendrocytes, astrocytes, and oligodendrocyte precursor cells. While other cell types (e.g., vascular cells and immune cells) are present in our data, these cell types contain fewer cells and are therefore more challenging to train a model on. For example, in the hippocampus, we have 21 million and 17 million long reads for neurons and glia respectively. However, the next largest cell type, progenitor cells, only contains 3.9 million reads. For immune and vascular cells, we have even fewer reads: 1.4 million and 0.3 million reads respectively.

Downsampling cell-type-specific Ψ values

In the human data, many exons (30,273 out of 68,215 for the hippocampus, 45,680 out of 56,427 for the frontal cortex, and 24,135 out of 28,978 for the ENCODE4 data) have $\Psi_{neur} > 0.9$, $\Psi_{glia} > 0.9$, and $\Delta\Psi_{(glia-neur)} < 0.03$. We downsampled these to 5,000 to make the distribution less skewed towards one.

In the mouse hippocampus data, 18,351 out of 23,857 exons have $\Psi = 1$ in neurons and glia, so we downsampled these to 5,000 as well. For the visual cortex, 27,073 out of 48,515 exons have $\Psi_{neur} > 0.9$, $\Psi_{glia} > 0.9$ and $\Delta\Psi_{(glia-neur)} < 0.03$. We downsampled these to 5,000.

ENCODE RBP-binding-site data

We downloaded the eCLIP data for 122 RBPs from the ENCODE portal (https://www.encodeproject.org/metadata/?status=released&internal_tags=ENCORE&assay_title=eCLIP&biosample_ontology.term_name=K562&target.investigated_as=RNA+binding+protein&biosample_ontology.term_name=HepG2&assembly=GRCh38&type=Experiment&files.processed=true). From this file list, we used the BED files that store

the peaks per replicate. We merged peaks from different replicates or cell lines to ensure one BED file per RBP.

NOVA RBP-binding-site data

We downloaded NOVA eCLIP data from: https://data.mendeley.com/datasets/v7p6dh5tvc/1. We merged peaks from different control individuals to ensure one BED file per RBP.

Logistic regression models

The logistic regression model is implemented as one fully connected layer between the input features (the RBP binding sites) and the output (the Ψ value) with a sigmoid activation function to scale the output between 0 and 1. The models are single-task models which means that a separate model was trained for each cell type.

When training the model, we use a binary cross entropy loss with L1 and L2 regularization (alpha = 0.001, and L1-ratio = 0.7), a learning rate of 0.005, and a batch size of 256.

As input for the logistic regression models, we counted the number of peaks in the BED files for every RBP and exon at six locations: 1) upstream of the exon (maximum 400 bp away from the splice site), 2) overlapping the 3' splice site, 3) within the exon, 4) spanning the exon, 5) overlapping the 5 splice site, and 6) downstream of the exon (maximum 400 bp away from the splice site). Since we used the eCLIP data of 122 RBPs and there are 6 possible locations, this resulted in 732 input features for every exon (Fig. 1A). If peaks of different replicates were overlapping, we counted those peaks only once.

The logistic regression model is implemented in PyTorch Lightning (v1.8.5) [104, 105].

Deep learning models

We adapted the architecture of the Saluki model [36] by removing one convolutional layer, shortening the maximum sequence from 12,288 to 6,144 bp, and changing the final activation function to a sigmoid activation function (Fig. S6). The exon of interest was centered in the middle of the input sequence. The input channels of the model depend on the input features used (sequence, splice sites, and/or RBP binding sites). For the sequence, we one-hot encoded the sequence which results in four channels. If the splice sites were used as input, this added an extra channel that indicates the start and end of the exon of interest. When adding the RBP binding sites, we add a channel for every RBP which one-hot encodes whether there is a binding site in any of the replicates of the eCLIP data for that RBP based on the BED files.

Similar to the logistic regression models, we trained a model for every cell type separately. Even though we adapted the Saluki model, we retrained all the weights in the model. When adding the mouse data, we adapted the same approach as Saluki and made the model a multi-head model where the weights of the convolutional and recurrent neuronal network layers are shared and the weights of the fully connected layer are species-specific (Fig. S6).

When training the model, we used the same hyperparameters, including the learning rate, batch size, etc., as the original Saluki model (Fig. S6).

For the hippocampus, we tested how input-sequence length and the number of convolutional layers affect the performance. The benefit of a longer input sequence is that the

model can learn how long-distance interactions of regulatory elements affect splicing, but these models contain more parameters and are more difficult to train. The different models performed similarly which indicates that the most important information is close to the splice sites of the exon (Fig. S35). The model using 6,144 bp and five channels performed slightly better for both neurons and glia and therefore we used it during all the experiments.

Evaluation

We evaluated the performance of the models using a tenfold cross-validation. We ensured that the same set of exons was always in the same test fold such that we could compare the performance of the models. Exons from the same gene were always in the same test fold. When training the deep learning models on human and mouse data simultaneously, we ensured that human-mouse homologs were in the same test fold. We used biomart to obtain the human-mouse homologs.

Some exons do not have any binding sites measured for any of the RBPs (5,560 exons in the hippocampus and 3,462 in the frontal cortex). This could for instance happen if certain genes were not expressed in the cell lines when the RBP binding sites were measured. Since the logistic regression model could not predict a Ψ value for these exons, we filtered these from the training set used for the logistic regression model and from all test sets (to enable a fair comparison between the logistic regression and deep learning models). The deep learning models are thus trained on more exons (Table S1). In the test set, there are 1,827 and 1,072 variable exons for the hippocampus and frontal cortex respectively.

We trained all models five times for every fold and averaged the predictions across these five runs. We evaluated the performance by calculating the Spearman correlation between the true and predicted Ψ values.

Evaluating dataset size and model performance

We calculated Ψ values for each subset of the data as explained above. We added individuals sequentially based on their individual ID, meaning the order was effectively random and not determined by read depth or any other quality metric.

RBP binding profiles

We generated RBP binding profiles by calculating the fraction of exons with an RBP binding site at every location (400 bp upstream of the exon - 400 bp downstream of the exon). Since exons have variable lengths, we bin the exons in 50 bins and only include exons that are at least 50 bp long in the analysis. We also filter out exons without RBP binding sites.

We calculate these profiles for four different groups of exons: 1) non-variable exons with $\Psi \geq 0.5$, 2) non-variable exons with $\Psi < 0.5$, 3) variable exons with $\Psi \geq 0.5$, and 4) variable exons with $\Psi < 0.5$.

To define how much the mechanisms in the variable exons diverged from non-variable exons, we calculate the mean-squared error between the RBP binding profiles of

the non-variable and variable exons. We do this for the exons with a high and low Ψ separately.

RBP expression data

We used the 10X scRNA-seq data from the same samples to look at the gene expression of the RBPs that were measured using the eCLIP data. We used Seurat (v4) for the analysis [106]. To create the heatmap in Fig. S20, we normalized the data per dataset using log normalization and a scale factor of 1e6. Next, we averaged the expression over all the cells. We plotted the $log(x+1)$ values.

We used the FindConservedMarkers() function using the default parameters (including Bonferroni multiple testing correction) from Seurat to find differentially expressed RBPs between neurons and glia. This tests for differentially expressed genes per individual and merges the results.

Interpretation of logistic regression model

For the interpretation of the logistic regression models, we looked at the coefficients of the input features. To obtain one value per input feature, we average the coefficients of the 10 folds and 5 runs per fold (so the average across 50 models in total).

We only compared the coefficients across models, if there were at least 50 exons with a binding site for that input feature.

In-silico saturation mutagenesis

We used *in-silico* saturation mutagenesis (ISM) to interpret how nucleotide substitutions in the input sequence affect the predictions. We did this for 9,929 exons using the $DL_{\text{all-seq-m}}$ model trained on glia in the hippocampus. For every exon, we used the fold for which that exon was in the test set. We averaged the predictions across the 5 runs. The ISM score is defined as follows: $ISM_{e,p,n} = \Psi_{pred,e,p,n} - \frac{1}{4}\sum_{i \in A,C,G,T} \Psi_{pred,e,p,i}$ where e is the exon we predict the Ψ value for, and p and n are the position and nucleotide used at that position respectively.

To visualize the ISM scores across the input sequence, we binned the upstream region, exon, and downstream region since they all had varying lengths.

Analysis of neighboring exons

We compared the ISM scores at the exon of interest, the neighboring exons, and the remaining sequence. We extracted the locations of annotated exons from GENCODE v35 [107]. The ISM scores for the exon of interest and the neighboring exons include the flanking sequence of 150 bp upstream and downstream of the exon.

Next, we selected ten exons on the positive strand with the highest absolute ISM scores in a neighboring exon. We visualized the long-reads spanning both exons using ScisorWiz [108].

Motif discovery

We used TF-MoDISco-lite (v2.2.0) [66] to discover motifs using the ISM scores as input. When creating the report, we compare the found motifs to the position weight

matrices from oRNAment which includes motifs found using RNAcompete and RNA-bind-n-seq experiments [8, 109, 110].

TF-MoDISco-lite is designed for DNA instead of RNA and tries both the forward strand and its reverse complement when finding seqlets (parts of the sequence with high ISM scores). We used the results file, to check whether the forward or reverse complement was used to generate a motif. We kept forward motifs if at least for 25 sequences the forward strand was used. We kept the reverse motif if at least for 25 sequences the reverse complement was used.

sQTL analysis

We used the sQTLs defined for the hippocampus in GTEx v8. These variants are linked to introns instead of exons. We predicted the effect for variants that are linked to an intron that spans an exon in our dataset (Fig. 4E). For most introns, there are multiple variants linked to them. We only predicted the effect for the best variants (the variants with the lowest *p*-value for an intron). For most introns, there were still more than two after this filter.

Exon naming

We named exons after their position in the transcript by counting their position in the GTF file. A conversion from exon names to genomic coordinates can be found in Table S5.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-026-04015-z.

Additional file 1. TF-MoDISco results after filtering for strand-specificity.

Additional file 2.

Acknowledgements
Research reported in this work was facilitated by computational resources and support of the Delft AI Cluster (DAIC) at TU Delft (RRID: SCR_025091), but remains the sole responsibility of the authors, not the DAIC team. Figures 1 A, 2 A, 4E, and S6 were created with BioRender.com.

Peer review information
Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
A.M., H.T., and L.M. conceived the project and designed the experiments. L.M. performed the experiments. N.B. and A.J. processed the data. J.H. performed RBP expression analysis. L.M. wrote the manuscript. A.M., H.T. and M.R. supervised the project and revised the manuscript. All authors approved the manuscript.

Funding
This research was supported by an NWO Gravitation project: BRAINSCAPES: A Roadmap from Neurogenetics to Neurobiology (NWO: 024.004.012), EMBO Scientific Exchange Grant 9673, the KNAW Van Leersum Grant/KNAW Medical Sciences Fund, Royal Netherlands Academy of Arts & Sciences, and the Feil Family Foundation.

Data availability
The source code to reproduce the figures and train logistic regression or deep learning models is available in the GitHub repository, at https://github.com/lcmmichielsen/PSI_pred [111], and in the Zenodo repository, at https://doi.org/10.5281/zenodo.3369158 [112]. The source code is released under MIT license. SnISOr-Seq HPC and ScISOr-Seq2 HPC and FC data can be downloaded from the Neuroscience Multi-Omic data archive (identifier dat-717krsa) [113]. SnISOr-Seq FC data can be downloaded from GEO (accession number GSE178175) [114]. Targeted SnISOr-Seq FC data can be downloaded from the SRA (accession number PRJNA1021558) [115]. ENCODE4 PGP1-derived astrocyte data can be downloaded from GEO (accession number GSE219423) [116]. ENCODE4 PGP1-derived excitatory neuron data can be downloaded from GEO (accession number GSE175137) [117]. ENCODE eCLIP data can be downloaded from the ENCODE portal: (https://www.encodeproject.org/metadata/?status=released&internal_tags=ENCORE&assay_title=eCLIP&biosample_ontology.term_name=K562&target.investigated_as=RNA+binding+protein&biosample_ontology.term_name=

HepG2&assembly=GRCh38&type=Experiment&files.processed=true). From this file list, we used the BED files that store the peaks per replicate. NOVA eCLIP data can be downloaded from Mendeley Data (accession number v7p6dh5tvc) [118]. The processed data (Ψ values, predictions, and RBP binding profiles) are available on Zenodo: https://zenodo.org/doi/10.5281/zenodo.10669666 [119].
SnISOr-Seq HPC and ScISOr-Seq2 HPC and FC data can be downloaded from the Neuroscience Multi-Omic data archive (identifier dat-717krsa) [113]. SnISOr-Seq FC data can be downloaded from GEO (accession number GSE178175) [114]. Targeted SnISOr-Seq FC data can be downloaded from the SRA (accession number PRJNA1021558) [115]. ENCODE4 PGP1-derived astrocyte data can be downloaded from GEO (accession number GSE219423) [116]. ENCODE4 PGP1-derived excitatory neuron data can be downloaded from GEO (accession number GSE175137) [117]. ENCODE eCLIP data can be downloaded from the ENCODE portal: (https://www.encodeproject.org/metadata/?status=released&internal_tags=ENCORE&assay_title=eCLIP&biosample_ontology.term_name=K562&target.investigated_as=RNA+binding+protein&biosample_ontology.term_name=HepG2&assembly=GRCh38&type=Experiment&files.processed=true). From this file list, we used the BED files that store the peaks per replicate. NOVA eCLIP data can be downloaded from Mendeley Data (accession number v7p6dh5tvc) [118].
The processed data (Ψ values, predictions, and RBP binding profiles) are available on Zenodo: https://doi.org/10.5281/zenodo.10669666 [119].

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
H.U.T. has presented at user meetings of 10x Genomics, Oxford Nanopore Technologies, and Pacific Biosciences, which in some cases included payment for travel and accommodations. Additionally the Tilgner lab has received discounted reagents from Agilent, Pacific Biosciences, Illumina and Oxford Nanopore. HUT has recently agreed to consult for ISOgenix Ltd., for work unrelated to the present manuscript. The other authors declare no competing interests.

Received: 12 November 2024 Accepted: 16 February 2026
Published online: 28 February 2026

References

1. Cheng J, Zhou T, Liu C, Shapiro JP, Brauer MJ, Kiefer MC, et al. Protection from Fas-mediated apoptosis by a soluble form of the Fas molecule. Science. 1994;263:1759–62. https://doi.org/10.1126/science.7510905.
2. Wright CJ, Smith CWJ, Jiggins CD. Alternative splicing as a source of phenotypic diversity. Nat Rev Genet. 2022;23:697–710. https://doi.org/10.1038/s41576-022-00514-4.
3. Yang X, Coulombe-Huntington J, Kang S, Sheynkman GM, Hao T, Richardson A, et al. Widespread expansion of protein interaction capabilities by alternative splicing. Cell. 2016;164:805–17. https://doi.org/10.1016/j.cell.2016.01.029.
4. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008;40:1413–5. https://doi.org/10.1038/ng.259.
5. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;456:470–6. https://doi.org/10.1038/nature07509.
6. Yeo G, Holste D, Kreiman G, Burge CB. Variation in alternative splicing across human tissues. Genome Biol. 2004;5:R74. https://doi.org/10.1186/gb-2004-5-10-r74.
7. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. Nat Rev Genet. 2014;15:829–45. https://doi.org/10.1038/nrg3813.
8. Van Nostrand EL, Freese P, Pratt GA, Wang X, Wei X, Xiao R, et al. A large-scale binding and functional map of human RNA-binding proteins. Nature. 2020;583:711–9. https://doi.org/10.1038/s41586-020-2077-3.
9. Zhang X, Chen MH, Wu X, Kodani A, Fan J, Doan R, et al. Cell-Type-Specific Alternative Splicing Governs Cell Fate in the Developing Cerebral Cortex. Cell. 2016;166:1147–62.e15. https://doi.org/10.1016/j.cell.2016.07.025
10. Fisher E, Feng J. RNA splicing regulators play critical roles in neurogenesis. WIREs RNA. 2022;13:e1728. https://doi.org/10.1002/wrna.1728.
11. Zhang C, Frias MA, Mele A, Ruggiu M, Eom T, Marney CB, et al. Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. Science. 2010;329:439. https://doi.org/10.1126/science.1191150.
12. Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. Systematic identification and analysis of exonic splicing silencers. Cell. 2004;119:831. https://doi.org/10.1016/j.cell.2004.11.010.
13. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, et al. Deciphering the splicing code. Nature. 2010;465:53. https://doi.org/10.1038/nature09000.
14. Barash Y, Vaquero-Garcia J, González-Vallinas J, Xiong HY, Gao W, Lee LJ, et al. AVISPA: a web tool for the prediction and analysis of alternative splicing. Genome Biol. 2013;14:R114. https://doi.org/10.1186/gb-2013-14-10-r114.
15. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science. 2015;347:1254806. https://doi.org/10.1126/science.1254806
16. Sharon D, Tilgner H, Grubert F, Snyder M. A single-molecule long-read survey of the human transcriptome. Nat Biotechnol. 2013;31:1009. https://doi.org/10.1038/nbt.2705.

17. Au KF, Sebastiano V, Afshar PT, Durruthy JD, Lee L, Williams BA, et al. Characterization of the human ESC transcriptome by hybrid sequencing. Proc Natl Acad Sci U S A. 2013;110:E4821-30. https://doi.org/10.1073/pnas.1320101110.
18. Marx V. Method of the year: long-read sequencing. Nat Methods. 2023;20:6–11. https://doi.org/10.1038/s41592-022-01730-w.
19. Foord C, Hsu J, Jarroux J, Hu W, Belchikov N, Pollard S, et al. The variables on RNA molecules: concert or cacophony? Answers in long-read sequencing. Nat Methods. 2023;20:20. https://doi.org/10.1038/s41592-022-01715-9.
20. Lucas MC, Novoa EM. Long-read sequencing in the era of epigenomics and epitranscriptomics. Nat Methods. 2023;20:25. https://doi.org/10.1038/s41592-022-01724-8.
21. Gupta I, Collier PG, Haase B, Mahfouz A, Joglekar A, Floyd T, et al. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. Nat Biotechnol. 2018;36:1197–202. https://doi.org/10.1038/nbt.4259.
22. Singh M, Al-Eryani G, Carswell S, Ferguson JM, Blackburn J, Barton K, et al. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. Nat Commun. 2019;10:3120. Available from: https://www.nature.com/articles/s41467-019-11049-4
23. Hardwick SA, Hu W, Joglekar A, Fan L, Collier PG, Foord C, et al. Single-nuclei isoform RNA sequencing unlocks barcoded exon connectivity in frozen brain tissue. Nat Biotechnol. 2022. https://doi.org/10.1038/s41587-022-01231-3.
24. Joglekar A, Prjibelski A, Mahfouz A, Collier P, Lin S, Schlusche AK, et al. A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. Nat Commun. 2021;12:463. Available from: http://www.nature.com/articles/s41467-020-20343-5
25. Joglekar A, Hu W, Zhang B, Narykov O, Diekhans M, Marrocco J, et al. Single-cell long-read sequencing-based mapping reveals specialized splicing patterns in developing and adult mouse and human brain. Nat Neurosci. 2024;27:1051–63. https://doi.org/10.1038/s41593-024-01616-4.
26. Irimia M, Weatheritt RJ, Ellis JD, Parikshak NN, Gonatopoulos-Pournatzis T, Babor M, et al. A highly conserved program of neuronal microexons is misregulated in autistic brains. Cell. 2014;159:1511–23. https://doi.org/10.1016/j.cell.2014.11.035.
27. Parikshak NN, Swarup V, Belgard TG, Irimia M, Ramaswami G, Gandal MJ, et al. Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. Nature. 2016;540:423–7. https://doi.org/10.1038/nature20612.
28. Gonatopoulos-Pournatzis T, Blencowe BJ. Microexons: at the nexus of nervous system development, behaviour and autism spectrum disorder. Curr Opin Genet Dev. 2020;65:22–33. https://doi.org/10.1016/j.gde.2020.03.007.
29. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019;176:535–48.e24. https://doi.org/10.1016/j.cell.2018.12.015
30. Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Kelso DJ, Kelso J, editors. Bioinformatics. 2021; Available from: https://academic.oup.com/bioinformatics/advance-article/doi/. https://doi.org/10.1093/bioinformatics/btab083/6128680
31. Zeng T, Li YI. Predicting RNA splicing from DNA sequence using pangolin. Genome Biol. 2022;23:103. https://doi.org/10.1186/s13059-022-02664-4.
32. Cheng J, Çelik MH, Kundaje A, Gagneur J. Mtsplice predicts effects of genetic variants on tissue-specific splicing. Genome Biol. 2021;22:94. https://doi.org/10.1186/s13059-021-02273-7.
33. Kuroyanagi H. Fox-1 family of RNA-binding proteins. Cell Mol Life Sci. 2009;66:3895–907. https://doi.org/10.1007/s00018-009-0120-5.
34. Dredge BK, Stefani G, Engelhard CC, Darnell RB. Nova autoregulation reveals dual functions in neuronal splicing. EMBO J. 2005;24:1608–20. https://doi.org/10.1038/sj.emboj.7600630.
35. Ule J, Ule A, Spencer J, Williams A, Hu J-S, Cline M, et al. Nova regulates brain-specific splicing to shape the synapse. Nat Genet. 2005;37:844–52. https://doi.org/10.1038/ng1610.
36. Agarwal V, Kelley DR. The genetic and biochemical determinants of mRNA degradation rates in mammals. Genome Biol. 2022;23:245. https://doi.org/10.1186/s13059-022-02811-x.
37. Reese F, Williams B, Balderrama-Gutierrez G, Wyman D, Çelik MH, Rebboah E, et al. The ENCODE4 long-read RNA-seq collection reveals distinct classes of transcript structure diversity. bioRxivorg. 2023 [cited 2025 Jul 7]. p. 2023.05.15.540865. Available from: https://www.biorxiv.org/content/. https://doi.org/10.1101/2023.05.15.540865v1.abstract
38. Handel AE, Chintawar S, Lalic T, Whiteley E, Vowles J, Giustacchini A, et al. Assessing similarity to primary tissue and cortical layer identity in induced pluripotent stem cell-derived cortical neurons through single-cell transcriptomics. Hum Mol Genet. 2016;25(5):989–1000. https://doi.org/10.1093/hmg/ddv637.
39. Barral S, Kurian MA. Utility of induced pluripotent stem cells for the study and treatment of genetic diseases: focus on childhood neurological disorders. Front Mol Neurosci. 2016;9:78. https://doi.org/10.3389/fnmol.2016.00078.
40. Hu W, Foord C, Hsu J, Fan L, Corley MJ, Lanjewar SN, et al. Combined single-cell profiling of chromatin-transcriptome and splicing across brain cell types, regions and disease state. Nat Biotechnol. 2025 [cited 2025 Jul 29];1–13. https://doi.org/10.1038/s41587-025-02734-5
41. Oberstrass FC, Auweter SD, Erat M, Hargous Y, Henning A, Wenter P, et al. Structure of PTB bound to RNA: specific binding and implications for splicing regulation. Science. 2005;309:2054–7. https://doi.org/10.1126/science.1114066.
42. Keppetipola N, Sharma S, Li Q, Black DL. Neuronal regulation of pre-mRNA splicing by polypyrimidine tract binding proteins, PTBP1 and PTBP2. Crit Rev Biochem Mol Biol. 2012;47:360–78. https://doi.org/10.3109/10409238.2012.691456.
43. Llorian M, Schwartz S, Clark TA, Hollander D, Tan L-Y, Spellman R, et al. Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. Nat Struct Mol Biol. 2010;17:1114–23. https://doi.org/10.1038/nsmb.1881.
44. Boutz PL, Stoilov P, Li Q, Lin C-H, Chawla G, Ostrow K, et al. A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. Genes Dev. 2007;21:1636–52. https://doi.org/10.1101/gad.1558107.

45. Vuong JK, Lin C-H, Zhang M, Chen L, Black DL, Zheng S. PTBP1 and PTBP2 serve both specific and redundant functions in neuronal pre-mRNA splicing. Cell Rep. 2016;17:2766–75. https://doi.org/10.1016/j.celrep.2016.11.034.
46. Makeyev EV, Zhang J, Carrasco MA, Maniatis T. The microRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. Mol Cell. 2007;27:435–48. https://doi.org/10.1016/j.molcel.2007.07.015.
47. Matlin AJ, Clark F, Smith CWJ. Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol. 2005;6:386–98. https://doi.org/10.1038/nrm1645.
48. Coelho MB, Attig J, Bellora N, König J, Hallegger M, Kayikci M, et al. Nuclear matrix protein Matrin3 regulates alternative splicing and forms overlapping regulatory networks with PTB. EMBO J. 2015;34(5):653–68. https://doi.org/10.15252/embj.201489852.
49. Uemura Y, Oshima T, Yamamoto M, Reyes CJ, Costa Cruz PH, Shibuya T, et al. Matrin3 binds directly to intronic pyrimidine-rich sequences and controls alternative splicing. Genes Cells. 2017;22:785–98. https://doi.org/10.1111/gtc.12512.
50. Wang J-Z, Fu X, Fang Z, Liu H, Zong F-Y, Zhu H, et al. QKI-5 regulates the alternative splicing of cytoskeletal gene ADD3 in lung cancer. J Mol Cell Biol. 2021;13:347–60. https://doi.org/10.1093/jmcb/mjaa063.
51. Hayakawa-Yano Y, Suyama S, Nogami M, Yugami M, Koya I, Furukawa T, et al. An RNA-binding protein, Qki5, regulates embryonic neural stem cells through pre-mRNA processing in cell adhesion signaling. Genes Dev. 2017;31:1910–25. https://doi.org/10.1101/gad.300822.117.
52. Montañés-Agudo P, Aufiero S, Schepers EN, van der Made I, Cócera-Ortega L, Ernault AC, et al. The RNA-binding protein QKI governs a muscle-specific alternative splicing program that shapes the contractile function of cardiomyocytes. Cardiovasc Res. 2023;119(5):1161–74. https://doi.org/10.1093/cvr/cvad007.
53. Darbelli L, Vogel G, Almazan G, Richard S. Quaking regulates neurofascin 155 expression for myelin and axoglial junction maintenance. J Neurosci. 2016;36:4106–20. https://doi.org/10.1523/JNEUROSCI.3529-15.2016.
54. Darbelli L, Choquet K, Richard S, Kleinman CL. Transcriptome profiling of mouse brains with qkI-deficient oligodendrocytes reveals major alternative splicing defects including self-splicing. Sci Rep. 2017;7:7554. https://doi.org/10.1038/s41598-017-06211-1.
55. Haroutunian V, Katsel P, Dracheva S, Davis KL. The Human Homolog of the QKI Gene Affected in the Severe Dysmyelination "Quaking" Mouse Phenotype: Downregulated in Multiple Brain Regions in Schizophrenia. AJP. 2006;163:1834–7. Available from: https://ajp.psychiatryonline.org/doi/abs/. https://doi.org/10.1176/ajp.2006.163.10.1834
56. Åberg K, Saetre P, Jareborg N, Jazin E. Human QKI, a potential regulator of mRNA expression of human oligodendrocyte-related genes involved in schizophrenia. Proceedings of the National Academy of Sciences. 2006;103:7482–7. Available from: https://www.pnas.org/doi/abs/.https://doi.org/10.1073/pnas.0601213103
57. Krach F, Wheeler EC, Regensburger M, Boerstler T, Wend H, Vu AQ, et al. Aberrant NOVA1 function disrupts alternative splicing in early stages of amyotrophic lateral sclerosis. Acta Neuropathol. 2022;144:413–35. https://doi.org/10.1007/s00401-022-02450-3.
58. Ule J, Stefani G, Mele A, Ruggiu M, Wang X, Taneri B, et al. An RNA map predicting Nova-dependent splicing regulation. Nature. 2006;444:580–6. https://doi.org/10.1038/nature05304.
59. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. 2016;26:990–9. https://doi.org/10.1101/gr.200535.115.
60. Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Res. 2018;28:739–50. https://doi.org/10.1101/gr.227819.117.
61. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021;18:1196–203. https://doi.org/10.1038/s41592-021-01252-x.
62. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015;12:931–4. https://doi.org/10.1038/nmeth.3547.
63. Lenasi T, Peterlin BM, Dovc P. Distal regulation of alternative splicing by splicing enhancer in equine β- casein intron 1. RNA. 2006;12:498–507. https://doi.org/10.1261/rna.7261206.
64. Tilgner H, Jahanbani F, Blauwkamp T, Moshrefi A, Jaeger E, Chen F, et al. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. Nat Biotechnol. 2015;33:736–42. https://doi.org/10.1038/nbt.3242.
65. Fededa JP, Petrillo E, Gelfand MS, Neverov AD, Kadener S, Nogués G, et al. A polar mechanism coordinates different regions of alternative splicing within a single gene. Mol Cell. 2005;19:393–404. https://doi.org/10.1016/j.molcel.2005.06.035.
66. Shrikumar A, Tian K, Avsec Ž, Shcherbina A, Banerjee A, Sharmin M, et al. Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5. arXiv [cs.LG]. 2018. Available from: http://arxiv.org/abs/1811.00416
67. Choi SH, Flamand MN, Liu B, Zhu H, Hu M, Wang M, et al. RBM45 is an m6A-binding protein that affects neuronal differentiation and the splicing of a subset of mRNAs. Cell Rep. 2022;40:111293. https://doi.org/10.1016/j.celrep.2022.111293.
68. GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369:1318–30. https://doi.org/10.1126/science.aaz1776.
69. Cléry A, Sinha R, Anczuków O, Corrionero A, Moursy A, Daubner GM, et al. Isolated pseudo-RNA-recognition motifs of SR proteins can regulate splicing using a noncanonical mode of RNA recognition. Proc Natl Acad Sci U S A. 2013;110:E2802-11. https://doi.org/10.1073/pnas.1303445110.
70. Merendino L, Guth S, Bilbao D, Martínez C, Valcárcel J. Inhibition of msl-2 splicing by Sex-lethal reveals interaction between U2AF35 and the 3′ splice site AG. Nature. 1999;402:838–41. https://doi.org/10.1038/45602.
71. Singh R, Valcárcel J, Green MR. Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. Science. 1995;268:1173–6. https://doi.org/10.1126/science.7761834.

72. Wu S, Romfo CM, Nilsen TW, Green MR. Functional recognition of the 3′ splice site AG by the splicing factor U2AF35. Nature. 1999;402:832–5. https://doi.org/10.1038/45590.
73. O'Neill AC, Uzbas F, Antognolli G, Merino F, Draganova K, Jäck A, et al. Spatial centrosome proteome of human neural cells uncovers disease-relevant heterogeneity. Science. 2022;376:eabf9088. https://doi.org/10.1126/science.abf9088.
74. Velázquez-Cruz A, Baños-Jaime B, Díaz-Quintana A, la De Rosa MA, Díaz-Moreno I. Post-translational control of RNA-binding proteins and disease-related dysregulation. Front Mol Biosci. 2021;8:658852. https://doi.org/10.3389/fmolb.2021.658852.
75. Agirre E, Oldfield AJ, Bellora N, Segelle A, Luco RF. Splicing-associated chromatin signatures: a combinatorial and position-dependent role for histone marks in splicing definition. Nat Commun. 2021;12:682. https://doi.org/10.1038/s41467-021-20979-x.
76. Petrova V, Song R, DEEP Consortium, Nordström KJV, Walter J, Wong JJL, et al. Increased chromatin accessibility facilitates intron retention in specific cell differentiation states. Nucleic Acids Res. 2022;50:11563–79. https://doi.org/10.1093/nar/gkac994.
77. Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. Regulation of alternative splicing by histone modifications. Science. 2010;327:996–1000. https://doi.org/10.1126/science.1184208.
78. de la Mata M, Alonso CR, Kadener S, Fededa JP, Blaustein M, Pelisch F, et al. A slow RNA polymerase II affects alternative splicing in vivo. Mol Cell. 2003;12:525–32. https://doi.org/10.1016/j.molcel.2003.08.001.
79. Roberts GC, Gooding C, Mak HY, Proudfoot NJ, Smith CW. Co-transcriptional commitment to alternative splice site selection. Nucleic Acids Res. 1998;26:5568–72. https://doi.org/10.1093/nar/26.24.5568.
80. Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, et al. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature. 2011 [cited 2024 Jan 30];479:74–9. Available from: https://www.nature.com/articles/nature10442
81. Andersson R, Enroth S, Rada-Iglesias A, Wadelius C, Komorowski J. Nucleosomes are well positioned in exons and carry characteristic histone modifications. Genome Res. 2009;19:1732–41. https://doi.org/10.1101/gr.092353.109.
82. Hon G, Wang W, Ren B. Discovery and annotation of functional chromatin signatures in the human genome. PLoS Comput Biol. 2009;5:e1000566. https://doi.org/10.1371/journal.pcbi.1000566.
83. Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J. Differential chromatin marking of introns and expressed exons by H3K36me3. Nat Genet. 2009;41:376–81. https://doi.org/10.1038/ng.322.
84. Nahkuri S, Taft RJ, Mattick JS. Nucleosomes are preferentially positioned at exons in somatic and sperm cells. Cell Cycle. 2009;8:3420–4. https://doi.org/10.4161/cc.8.20.9916.
85. Schwartz S, Meshorer E, Ast G. Chromatin organization marks exon-intron structure. Nat Struct Mol Biol. 2009;16:990–5. https://doi.org/10.1038/nsmb.1659.
86. Spies N, Nielsen CB, Padgett RA, Burge CB. Biased chromatin signatures around polyadenylation sites and exons. Mol Cell. 2009;36:245–54. https://doi.org/10.1016/j.molcel.2009.10.008.
87. Iannone C, Pohl A, Papasaikas P, Soronellas D, Vicent GP, Beato M, et al. Relationship between nucleosome positioning and progesterone-induced alternative splicing in breast cancer cells. RNA. 2015;21:360–74. https://doi.org/10.1261/rna.048843.114.
88. Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcárcel J, et al. Nucleosome positioning as a determinant of exon recognition. Nat Struct Mol Biol. 2009;16:996–1001. https://doi.org/10.1038/nsmb.1658.
89. Mendel M, Delaney K, Pandey RR, Chen K-M, Wenda JM, Vågbø CB, et al. Splice site m6A methylation prevents binding of U2AF35 to inhibit RNA splicing. Cell. 2021;184:3125–42.e25. https://doi.org/10.1016/j.cell.2021.03.062
90. Wang S, Lv W, Li T, Zhang S, Wang H, Li X, et al. Dynamic regulation and functions of mRNA m6A modification. Cancer Cell Int. 2022;22:48. https://doi.org/10.1186/s12935-022-02452-x.
91. Yu B, Yu X, Xiong J, Ma M. Methylation modification, alternative splicing, and noncoding RNA play a role in cancer metastasis through epigenetic regulation. BioMed Res Int. 2021;2021:4061525. https://doi.org/10.1155/2021/4061525.
92. Lin T-C, Tsai C-H, Shiau C-K, Huang J-H, Tsai H-K. Predicting splicing patterns from the transcription factor binding sites in the promoter with deep learning. BMC Genomics. 2024 [cited 2024 Nov 4];25:830. Available from: https://bmcgenomics.biomedcentral.com/articles/. https://doi.org/10.1186/s12864-024-10667-7
93. García-Ruiz S, Zhang D, Gustavsson EK, Rocamora-Perez G, Grant-Peters M, Fairbrother-Browne A, et al. Splicing accuracy varies across human introns, tissues, age and disease. Nat Commun. 2025;16(1):1068. https://doi.org/10.1038/s41467-024-55607-x.
94. Jorstad NL, Song JHT, Exposito-Alonso D, Suresh H, Castro-Pacheco N, Krienen FM, et al. Comparative transcriptomics reveals human-specific cortical features. Science. 2023;382:eade9516. Available from: https://www.science.org/doi/abs/. https://doi.org/10.1126/science.ade9516
95. Chao K-H, Mao A, Salzberg SL, Pertea M. Splam: a deep-learning-based splice site predictor that improves spliced alignments. Genome Biol. 2024 [cited 2024 Nov 4];25:243. Available from: https://genomebiology.biomedcentral.com/articles/. https://doi.org/10.1186/s13059-024-03379-4
96. Karollus A, Mauermeier T, Gagneur J. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. Genome Biol. 2023;24:56. https://doi.org/10.1186/s13059-023-02899-9.
97. Treutlein B, Gokce O, Quake SR, Südhof TC. Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. Proc Natl Acad Sci U S A. 2014;111:E1291-9. https://doi.org/10.1073/pnas.1403244111.
98. Schreiner D, Nguyen T-M, Russo G, Heber S, Patrignani A, Ahrné E, et al. Targeted combinatorial alternative splicing generates brain region-specific repertoires of neurexins. Neuron. 2014;84:386–98. https://doi.org/10.1016/j.neuron.2014.09.011.
99. Tilgner H, Jahanbani F, Gupta I, Collier P, Wei E, Rasmussen M, et al. Microfluidic isoform sequencing shows widespread splicing coordination in the human transcriptome. Genome Res. 2018;28:231–42. https://doi.org/10.1101/gr.230516.117.

100. Foord C, Prjibelski AD, Hu W, Michielsen L, Vandelli A. A spatial long-read approach at near-single-cell resolution reveals developmental regulation of splicing and polyadenylation sites in distinct cortical layers and cell types. bioRxiv. 2025 [cited 2025 Jun 11]. p. 2025.06.10.658877. Available from: https://www.biorxiv.org/content/. https://doi.org/10.1101/2025.06.10.658877v1.abstract
101. Michielsen L, Prjibelski A, Foord C, Hu W, Jarroux J, Hsu J, et al. Spatial isoform sequencing at sub-micrometer single-cell resolution reveals novel patterns of spatial isoform variability in brain cell types. bioRxiv. 2025 [cited 2025 Jul 9]. p. 2025.06.25.661563. Available from: https://www.biorxiv.org/content/. https://doi.org/10.1101/2025.06.25.661563v1.abstract
102. Schwessinger R, Deasy J, Woodruff RT, Young S, Branson KM. Single-cell gene expression prediction from DNA sequence at large contexts. bioRxiv. 2023 [cited 2023 Dec 7]. p. 2023.07.26.550634. Available from: https://www.biorxiv.org/content/. https://doi.org/10.1101/2023.07.26.550634v2
103. Prjibelski AD, Mikheenko A, Joglekar A, Smetanin A, Jarroux J, Lapidus AL, et al. Accurate isoform discovery with IsoQuant using long reads. Nat Biotechnol. 2023;41:915–8. https://doi.org/10.1038/s41587-022-01565-y.
104. Falcon W, The PyTorch Lightning team. PyTorch Lightning. 2019. Available from: https://github.com/PyTorchLightning/pytorch-lightning
105. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach H, Larochelle H, Beygelzimer A, d\textquotesingle Alché-Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems 32. Curran Associates, Inc.; 2019. p. 8024–35. Available from: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
106. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;0. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867421005833
107. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47:D766–73. https://doi.org/10.1093/nar/gky955
108. Stein AN, Joglekar A, Poon C-L, Tilgner HU. Scisorwiz: visualizing differential isoform expression in single-cell long-read data. Bioinformatics. 2022;38:3474–6. https://doi.org/10.1093/bioinformatics/btac340.
109. Benoit Bouvrette LP, Bovaird S, Blanchette M, Lécuyer E. Ornament: a database of putative RNA binding protein target sites in the transcriptomes of model species. Nucleic Acids Res. 2020;48:D166-73. https://doi.org/10.1093/nar/gkz986.
110. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature. 2013;499:172–7. https://doi.org/10.1038/nature12311.
111. Michielsen L, Hsu J, Joglekar A, Belchikov N, Reinders MJT, Tilgner HU, & Mahfouz A. Cell-type specific PSI prediction code. GitHub. 2026. https://github.com/lcmmichielsen/PSI_pred
112. Michielsen L, Hsu J, Joglekar A, Belchikov N, Reinders MJT, Tilgner HU, & Mahfouz A. Cell-type specific PSI prediction code. Zenodo. 2026. https://doi.org/10.5281/zenodo.18642297
113. Joglekar A, Hu W, Zhang B, Narykov O, Diekhans M, Marrocco J, et al. Single-cell long-read sequencing-based mapping reveals specialized splicing patterns in developing and adult mouse and human brain. dat-717krsa. The Neuroscience Multi-omic Data Archive. 2024. https://assets.nemoarchive.org/dat-717krsa
114. Hardwick SA, Hu W, Joglekar A, Fan L, Collier PG, Foord C, et al. Single-nuclei isoform RNA sequencing unlocks barcoded exon connectivity in frozen brain tissue. GSE178175. Gene Expression Omnibus. 2022. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178175
115. Hu W, Foord C, Hsu J, Fan L, Corley MJ, Lanjewar SN, et al. Combined single-cell profiling of chromatin-transcriptome and splicing across brain cell types, regions and disease state. PRJNA1021558. Sequence Read Archive. 2024. https://www.ncbi.nlm.nih.gov/sra/PRJNA1021558
116. Reese F, Williams B, Balderrama-Gutierrez G, Wyman D, Çelik MH, Rebboah E, et al. The ENCODE4 long-read RNA-seq collection reveals distinct classes of transcript structure diversity. GSE219423. Gene Expression Omnibus. 2022. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE219423
117. Reese F, Williams B, Balderrama-Gutierrez G, Wyman D, Çelik MH, Rebboah E, et al. The ENCODE4 long-read RNA-seq collection reveals distinct classes of transcript structure diversity. GSE175137. Gene Expression Omnibus. 2022. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE175137
118. Krach F, Wheeler EC, Regensburger M, Boerstler T, Wend H, Vu AQ, et al. Aberrant NOVA1 function disrupts alternative splicing in early stages of amyotrophic lateral sclerosis. v7p6dh5tvc. Mendeley Data. 2022. https://data.mendeley.com/datasets/v7p6dh5tvc/1
119. Michielsen L, Hsu J, Joglekar A, Belchikov N, Reinders MJT, Tilgner HU, & Mahfouz A. Cell-type specific PSI prediction processed data. Zenodo. 2026. https://doi.org/10.5281/zenodo.18642297

Publishers' Note