RNA splicing analysis using heterogeneous and large RNA-seq datasets

Jorge Vaquero-Garcia[1, 5], Joseph K. Aicher[1, 4, 5], Paul Jewell[1, 5], Matthew R. Gazzara[1, 5], Caleb M. Radens[1, 6], Anupama Jha[2, 6], Christopher J. Green[1], Scott S. Norton[1], Nicholas F. Lahens[3], Gregory R. Grant[1], and Yoseph Barash[1, 2, 7]

[1]Department of Genetics, University of Pennsylvania

[2]Department of Computer and Information Science, University of Pennsylvania

[3]Institute for Translational Medicine and Therapeutics, University of Pennsylvania

[4]Division of Human Genetics, Children's Hospital of Philadelphia

[5]Equal contribution as co-first authors

[6]Equal contribution

[7]To whom correspondence should be addressed

Abstract

The ubiquity of RNA-seq has led to many methods that use RNA-seq data to analyze variations in RNA splicing. However, available methods are not well suited for handling heterogeneous and large datasets. Such datasets scale to thousands of samples across dozens of experimental conditions, exhibit increased variability compared to biological replicates, and involve thousands of unannotated splice variants resulting in increased transcriptome complexity. We describe here a suite of algorithms and tools implemented in the MAJIQ v2 package to address challenges in detection, quantification, and visualization of splicing variations from such datasets. Using both large scale synthetic data and GTEx v8 as benchmark datasets, we demonstrate that the approaches in MAJIQ v2 outperform existing methods. We then apply MAJIQ v2 package to analyze differential splicing across 2,335 samples from 13 brain subregions, demonstrating its ability to offer new insights into brain subregion-specific splicing regulation.

Introduction

The usage of RNA sequencing (RNA-seq) has become ubiquitous in biomedical research. While some studies utilize RNA-seq only to investigate the overall expression level of genes, an increasing number of studies analyze changes in the relative abundance of gene isoforms. Changes in gene isoforms can occur through multiple mechanisms, including alternative promoter usage, alternative polyadenylation, and alternative splicing (AS). The production of different gene isoforms can in turn lead to diverse functional consequences, including changes to the translated protein domains, to degradation rates, and to localization. Previous studies showed that the majority of human genes are alternatively spliced with over a third of them shown to change their major isoform across 16 human tissues[1]. These observations, combined with the association of splicing defects with both monogenic and complex disease, serve to motivate the study of splicing variations across diverse experimental conditions. Consequently, independent labs as well as large consortia produce vast amounts of RNA-seq data. Datasets may involve anywhere from just a few to many thousands of samples each, and are typically heterogeneous as they often do not represent biological or technical replicates. The consequent increased splicing variability, illustrated in Fig 1A,B, can be the result of a multitude of factors, both experimental (e.g. difference in sequencing machine), and biological (e.g. gender, age). While some confounding factors may be corrected with appropriate methods[2], fully removing the observed variability in such data is unlikely and may also over-constrain the data, thus leading to a loss of true biological signal. Thus, there is a general need for methods that can effectively detect, quantify, and visualize splicing variations from large and heterogeneous RNA-seq datasets.

Broadly, the quantification of changes in gene isoform usage can be divided between methods that aim to quantify whole isoforms and those that quantify more local AS "events" within a gene. While quantifying all gene isoforms accurately across diverse conditions can be regarded as the grand challenge of transcriptomics, achieving this goal remains open due to several limiting factors. In the case of long reads technology, these factors include high error rate and high costs which do not allow researchers to capture enough reads from all isoforms. In the case of the more commonly used short reads technology, these limiting factors include the sparsity of reads, their positional bias, and the fact that reads usually cannot be assigned to a unique isoform. In addition, the composition of isoforms in a sample is typically unknown, requiring further inference of the existing isoforms or making

2

54 simplifying assumptions such as a known transcriptome. These issues have led many researchers to

55 focus on local AS "events" which can be more easily and accurately quantified from RNA-seq. AS

56 events are quantified in terms of percent spliced in (PSI, denoted by Ψ), which is the relative ratio of

57 isoforms including a specific splicing junction or retained intron. Traditionally, AS events have been

58 studied only for a restricted set of the most common "types" (e.g. cassette exons). In a previous

59 study, we extended this set of AS event types using the formulation of local splicing variations (LSVs)

60 and introduced MAJIQ as a software package for studying such LSVs. LSVs, which can be defined as

61 splits in a gene splicegraph coming into or from a reference exon, allow researchers to capture not only

62 previously defined AS types but also much more complex variations involving more than two

63 alternative junctions (see examples in Figure 1C for illustration). Furthermore, the LSV formulation,

64 and similar definitions of local AS events suggested in subsequent works, also help incorporate and

65 quantify unannotated (*de novo*) splice junctions. Previous work comparing splicing across mouse

66 tissues has shown that accounting for complex and *de novo* variations results in over 30% increase of

67 detected differentially spliced events while maintaining the same level of reproducibility [3].

68 Importantly, capturing such unannotated splice variations is of particular importance for the study of

69 disease such as cancer and neurodegeneration which often involve aberrant splicing.

70 Despite previous demonstrations of MAJIQ's utility for analyzing AS[3, 4], we found it as well

71 as many other commonly used methods for AS events quantification not to be well-suited for handling

72 heterogeneous and large RNA-seq datasets. Such datasets pose several algorithmic, computational,

73 and visualization challenges. First, the assumption of a shared PSI per LSV junction in a group, used

74 by methods such as MAJIQ and LeafCutter, is violated in such data even when handling only a small

75 dataset with few samples, leading to a potential increase in false positives and loss of power. Second,

76 algorithms need to not only scale to thousands of samples efficiently but also to allow incrementally

77 adding new samples as more data is acquired, and to support multiple group comparisons (e.g.

78 multiple tissue comparisons across GTEx). Third, the increased complexity of the data requires

79 efficient representation. Such efficient representation would allow users to capture the many

80 unannotated splicing variations in the data, while at the same time simplifying its representation and

81 quantification. Such simplification will allow to filter lowly used splice junctions while also detecting

82 possibly new sub-types of significant variations. Finally, efficient and user-friendly visualization is

83 required to probe possibly multiple sample groups as well as individual samples.

3

84   To address the above challenges, we developed an array of tools and algorithms included in the

85   MAJIQ v2 package. These include nonparametric statistical tests for differential splicing (MAJIQ

86   HET), an incremental splicegraph builder, a new algorithm for quantifying intron retention, a method

87   to detect high-confidence negative (non-changing) splicing events, and an algorithm to parse all LSVs

88   across genes into modules which can then be classified into subtypes (Modulizer). These algorithms

89   and tools are coupled with a new visualization package (VOILA v2) which allows users to compare

90   multiple sample groups, simplify splicegraphs, and probe individual data points (e.g. LSV in an

91   individual sample) while representing hundreds or thousands of samples. In addition, to support

92   reproducibility, we develop a package for comparative evaluation of different methods for RNA splicing

93   analysis and use it to demonstrate that the new version of MAJIQ compares favorably with the current

94   state of the art using both synthetic (simulated) and real (GTEx) data. Finally, we apply the MAJIQ

95   v2 toolset to 2,335 RNA-seq samples from 374 donors across 13 brain subregions. We use VOILA v2

96   to visualize the result and highlight several key findings in brain subregions specific variations in

97   cerebellar tissue groups compared to the remaining brain regions.

Results

The MAJIQ v2 splicing analysis pipeline

100   To support RNA splicing analysis using large RNA-seq datasets we implemented the set of tools and

101   algorithms illustrated in Figure 1C. In the first step, the MAJIQ builder combines transcript

102   annotations and coverage from aligned RNA-seq experiments in order to build an updated splicegraph

103   for each gene which includes *de novo* (unannotated) elements such as junctions, retained introns, and

104   exons). Several user-defined filters can be applied at this stage to exclude junctions or retained introns

105   which have low coverage or are not detected in enough samples in user-defined sample groups.

106   Notably, per-experiment coverage is saved separately so that it can be used in subsequent analyses

107   without reprocessing aligned reads a second time (aka incremental build). This feature is highly

108   relevant for large studies with incremental releases such as ENCODE and GTEx and also for individual

109   lab projects where datasets or samples are added as the project evolves.

110   In the second step of the pipeline, the MAJIQ quantifier is executed. As in the original MAJIQ

111   framework, splicing quantification is performed in units of LSVs. Briefly, an LSV corresponds to a split

4

112 in gene splicegraphs coming into or out of a reference exon. Each LSV edge, corresponding to a splice

113 junction or intron retention, is quantified in terms of its relative inclusion (PSI, $\Psi \in [0, 1]$) or changes

114 in its relative inclusion between two conditions (dPSI, $\Delta\Psi \in [-1, 1]$). Given the junction spanning

115 reads observed in each LSV, MAJIQ's Bayesian model results in a posterior distributions over the

116 (unknown) inclusion level ($\mathbb{P}(\Psi)$), or the changes in inclusion levels between conditions ($\mathbb{P}(\Delta\Psi)$).

117 This model accounts not only for the total number of reads but also for factors such as read

118 distribution across genomic locations and read stacks. Given its Bayesian framework, the model can

119 also output the confidence in inclusion change of at least C ($\mathbb{P}(|\Delta\Psi| > C)$), or the expectation over

120 the computed posterior distributions ($\mathbb{E}[\Psi]$, $\mathbb{E}[\Delta\Psi]$). In this work, we introduce two new algorithms

121 within the MAJIQ quantifier. The first involves how intron retention is quantified, allowing for much

122 faster execution with higher accuracy (see Methods). The second addition is the implementation of

123 additional test statistics, termed MAJIQ HET (heterogeneous). Conceptually, the original MAJIQ

124 model assumes a shared (hidden) PSI value for a given group of samples and accumulates evidence

125 (reads) across these samples to infer PSI. In contrast, MAJIQ HET quantifies PSI for each sample

126 separately and then applies robust rank-based test statistics (TNOM, InfoScore, or Mann-Whitney U).

127 As we demonstrate below, the new HET test statistics allow MAJIQ to increase reproducibility in small

128 heterogeneous datasets, and gain power in large heterogeneous datasets.

129   A new optional analysis step introduced here is the VOILA Modulizer, an algorithm which

130 organizes all identified LSVs into AS modules and then groups these modules by type. Briefly, AS

131 modules represent distinct segments of a gene splicegraph involving overlapping LSVs which are

132 contained between a single source and single target exon. However, unlike DiffSplice's AS modules[5],

133 we do not use a recursive definition of these modules and instead classify all identified modules by

134 their substructures into types. The module's substructures are in turn defined by the basic units of

135 alternative splicing, namely intron retention, exon skipping and 3' or 5' splice variations. As we

136 demonstrate below, the automatic AS module classification greatly facilitates a wide range of

137 downstream analysis tasks.

138   The next step of the pipeline involves visualization of the quantified PSI and dPSI using VOILA

139 v2. This new package runs as an app (on macOS, Windows, Linux) which supports the visualization of

140 thousands of samples per LSV as violin beeswarm plots with multi group comparisons and advanced

141 user filters. Users can perform searches by gene name or junction, and simplify the visualization by

142 filtering out lowly included junctions. This option is highly relevant for large heterogeneous datasets

143 where many junctions might be captured but may not be relevant for specific comparisons/samples.

144 Notably, unlike the builder filters described above, the VOILA v2 filters do not affect the underlying

145 splicegraphs but only help declutter the visualization to aid in subsequent analysis. VOILA v2 has the

146 option to run as a server to share results with collaborators while all of the pipeline's results can also

147 be exported into other pipelines as tab-delimited files and for automated primer design for validation

148 using MAJIQ-SPEL[6].

Performance evaluation

150 In order to assess MAJIQ HET, our new method for detecting differential splicing, we performed a

151 comprehensive comparison to an array of commonly used algorithms using both synthetic and real

152 data. We considered only algorithms capable of analyzing large datasets, including the original MAJIQ

153 algorithm (upgraded with the v2 code-base to enable efficient data processing), rMATS turbo,

154 LeafCutter, SUPPA2, and Whippet. Figure 2A shows processing time and memory when performing a

155 multi-group, multi-sample comparison, typical for such datasets. In this case, we perform all pairwise

156 comparisons between 10 tissue groups, and the number of samples in each group grows from 1 (10

157 total samples) to 6 (60 total samples). All algorithms are able to process such large datasets using

158 only 0.5-4 GB of memory, an amount readily available on modern laptops. However, large differences

159 exist in terms of running time, with SUPPA2 (55 hours) and Whippet (50 hours) taking substantially

160 longer to analyze the larger dataset (6 samples per group, 60 total samples) compared to

161 approximately 6 hours by rMATS, LeafCutter and MAJIQ v2.

162   Next, we assessed the accuracy of all algorithms using a large-scale synthetic dataset for

163 comparing two tissue groups. This synthetic dataset, by far the largest of its kind to the best of our

164 knowledge, was constructed to be "realistic" such that each synthetic sample was generated to mimic

165 a real GTEx sample from either cerebellum or smooth muscle tissues (see Methods). All methods were

166 required to report changing AS events which pass the method's statistical significance test and

167 inferred to exhibit a substantial splicing change of at least 20% (see Methods). However, we note that

168 since the various algorithms use significantly different definitions of AS events it is hard to compare

169 those directly. For example, LeafCutter defines AS events as clusters of overlapping introns which may

170 involve multiple 3'/5' alternative splice sites and skipped exons, while rMATS is limited to only

6

171   classical AS events with two alternative junctions. Thus, to facilitate a comparative analysis, we

172   resorted to comparing the various algorithms output at the gene rather than event level using the

173   synthetic dataset shown in Fig. 2B. A more refined analysis at the AS event level for each method can

174   be found in Fig. S1 and follows the same trends discussed here at the gene level. First, we found

175   SUPPA2 consistently reported over 6,000 differentially spliced genes, thousands more than any other

176   method, while Whippet reported roughly 785 genes, significantly less than the other methods which

177   reported over 2,000 changing genes (Fig. 2B top bar chart). Whippet, followed by rMATS, reported

178   significantly more non-changing events. SUPPA2, rMATS, and Whippet all exhibited high FDR

179   ranging around 15-30%, with the former two also exhibiting high FNR over 40%. Both MAJIQ and

180   MAJIQ HET consistently maintained a lower false discovery rate compared to other algorithms (0.3%)

181   and a low level of false negative rate which was similar to that of LeafCutter. On small sets, for

182   example when using 5 samples per group, LeafCutter had a slightly lower FNR (2.5% vs 5.5% for

183   HET), but MAJIQ exhibited lower FDR (0.03% vs 0.8%) while still reporting overall 34% more genes

184   as changing (2,337 vs 1,739) and 6% more as non-changing (7,110 vs 6,713). It is also worth noting

185   that the actual difference in the number of changing AS events reported by MAJIQ and LeafCutter is

186   significantly higher, with 4,267 reported by MAJIQ vs. 2,169 by LeafCutter. This increased difference

187   is mainly due to the increased resolution of event definition by MAJIQ. Specifically, MAJIQ uses the

188   local splice variations formulation described above, while LeafCutter uses a definition of overlapping

189   intronic regions which give rise to coarser event definition and can be sensitive to the coverage

190   threshold used.

191   The significant differences between the methods described above raises the question how the

192   reported sets of differentially spliced genes overlap. Fig. 2C illustrates the result of such analysis when

193   using 10 samples per group. Here, we looked at the intersection between different methods at the

194   gene level and when a set was unique to a method (i.e. the underlying events are well defined) we also

195   estimated the associated FPR. We found SUPPA2 reports a significantly higher number of unique

196   genes (1,777) as differentially spliced but over a quarter of those are false positives. The next set sizes

197   are those for LeafCutter (333), HET and SUPPA2 (288), HET (230), and MAJIQ HET and PSI (214)

198   with a FPR of 4% for the LeafCutter's unique set and close to 0 FPR for both MAJIQ's algorithms

199   unique sets. rMATS and Whippet report significantly fewer unique genes with a high false positive

200   rate of 62% and 79% respectively.

7

201　　　Next, we turned to assess performance on real GTEx data using several metrics. Here, unlike

202　the synthetic data analysis which focused on comparative evaluation at the gene level, we focus on the

203　actual AS events reported by each method. First, we used the reproducibility ratio (RR) statistic as

204　shown in Figure 2D. The RR plots follow a similar procedure to that of irreproducible discovery rate

205　(IDR) plots, used extensively to evaluate ChIP-seq peak callers[7, 3]. Briefly, RR plots answer the

206　following simple question: given an algorithm A and a dataset D, if we rank all the events that

207　algorithm A identifies as differentially spliced $(1, ..., N_A)$, how many would be reproduced if you repeat

208　this with dataset D', comprised of similar experiments using biological or technical replicates? The

209　$RR(n)$ plot, as shown in Fig. 2D, is the fraction of those events that are reproduced (y-axis) as a

210　function of $n \leq N_A$ (x-axis), with the overall reproducibility of differentially spliced events expressed as

211　$RR(N_A)$ (far right point of each curve in Fig. 2D). In our RR analysis using groups of size 3 to 50

212　GTEx samples each, we found both MAJIQ and MAJIQ HET compared favorably to the other

213　methods, but with the new HET algorithm exhibiting improved detection power resulting in a higher

214　number of AS events at the same reproducibility level.

215　　　The second statistic we used for evaluating performance on real data is the intra-to-inter ratio

216　(IIR) [4], which serves as a proxy for FDR on real data where the labels are unknown. Specifically, IIR

217　computes the ratio between the number of differentially spliced events reported when comparing

218　groups of the same condition (e.g. brain) and the number of events reported for similar group sizes of

219　different conditions (e.g. brain vs liver). In our work, we found IIR to be a lower bound estimate of

220　true FDR, though it lacks theoretical guarantees. In the analysis shown in Fig. 2E, we found IIR to

221　behave similarly to FDR on synthetic data with MAJIQ, MAJIQ HET, and LeafCutter exhibiting low

222　IIR of 2%-6% even for small group sets of 5 samples, while rMATS, SUPPA2, and Whippet had an IIR

223　of 13%, 26% and 46% respectively. However, unlike FDR on synthetic data, IIR dropped much more

224　significantly, hitting practically zero for all methods for large sample groups. This result is to be

225　expected since the IIR statistic compares sample groups of the same type, unlike the synthetic dataset

226　described above where different tissues are compared.

227　　　The last component we included for assessing different methods' accuracy is a comparison to

228　PSI quantifications using triplicates of RT-PCR assays, the gold standard in the RNA field. We

229　previously produced over 100 such experiments from two different mouse tissues and showed MAJIQ

230　compared favorably to SUPPA and rMATS[3, 4]. Here, we extended this analysis to LeafCutter and

8

231   found that MAJIQ's quantifications correlates significantly better with those of RT-PCR (see Fig. S2).

232   We note that this analysis for LeafCutter was possible since all events we tested were simple cassette

233   exon skipping, but it is not clear how to translate LeafCutter's output to actual PSI in the general case.

VOILA v2 enables visualization of thousands of samples

235   To facilitate visualization and downstream analysis of both the new outputs from MAJIQ HET over

236   large, heterogeneous datasets and traditional MAJIQ PSI or MAJIQ dPSI quantification over replicate

237   experiments, we developed VOILA v2 as a server based cross-platform app. Replacing the previous

238   HTML file based visualization with VOILA v2 allows for interactive visualization of all LSVs in all

239   genes, with data ranging from one sample to thousands of samples. After an initial indexing step that

240   is run one time, users can now, on the fly, filter their data by several criteria including dPSI levels

241   between groups, read coverage over junctions, LSV types and complexity, and the statistical test for

242   significance, as opposed to re-running VOILA with the filtering criteria, as was required in the previous

243   version. Another advantage of the new VOILA v2 is the ability to run it as a server so that results can

244   be shared with collaborators without the need to transfer large files.

245   To highlight these new features, we ran MAJIQ HET and VOILA v2 on GTEx v8 brain tissues

246   which are known to exhibit high levels of alternative splicing. Overall, this analysis involved 2,335

247   RNA-seq samples from 374 donors across 13 tissue groups (see Methods). Figure 3A shows the

248   VOILA view for this large dataset for the key splicing factor gene *PTBP1*, including a splicegraph

249   (top) with combined read information from 225 cerebellum RNA-seq samples. Users can easily add

250   and remove splicegraphs for other tissue groups or individual samples of interest. Figure 3A bottom

251   panel shows a VOILA visualization for quantifying a single junction in a single LSV across the 2,335

252   RNA-seq samples. Here, the 13 tissues are displayed as violin beeswarm plots with each point

253   representing a single sample which can be interrogated by hovering the user's cursor over it. Finally,

254   VOILA uses a heatmap (Fig. 3A bottom right) to represent the pairwise differences between the tissue

255   groups for the junction of interest. The upper half of the heatmap represents the difference in medians

256   of $\mathbb{E}[\Psi]$ distributions between the tissue groups, while the bottom half represents the p-values

257   associated with these group differences (see Methods). For the example LSV and junction in *PTBP1*,

258   the cerebellar tissues (cerebellum and cerebellar hemisphere) show a distinct splicing pattern with

259   reduced usage of this junction (lower $\mathbb{E}[\Psi]$ values in the left-most violin plots) which was significant

9

260 according to MAJIQ HET (InfoScore shown) (Figure 3A).

VOILA Modulizer defines alternative splicing modules to facilitate downstream analysis

263 The LSV and junction showcased in the above example are of biological importance. PTBP1 is a

264 widely expressed splicing factor that binds CU-rich sequences, but it is downregulated during

265 neurogenesis which contributes to neuronal splicing patterns[8, 9, 10]. Decreased activity of PTBP1 in

266 neuronal tissues is attributed to numerous mechanisms, some of which involve splicing regulation of

267 two cassette exons in the region highlighted in the *PTBP1* splicegraph (Figure 3A,B boxed

268 regions)[11, 3], making differences between brain subregions of potential interest. Mammalian-specific,

269 neuronal skipping of an alternative cassette exon in the linker region between the second and third

270 RNA recognition motifs (RRMs) of *PTBP1* (exon 12 in the splicegraph) results in a protein isoform of

271 PTBP1 with reduced repressive activity leading to altered splicing patterns during neuronal

272 differentiation[11]. Additionally, in mouse brain we previously described inclusion of a unannotated,

273 premature termination codon (PTC) containing, cassette exon with conserved splice sites in humans

274 that shows increased inclusion in mouse cerebellum (compared to brainstem and hypothalamus) and is

275 developmentally regulated through murine cortex development[3]. While LeafCutter analysis of *PTBP1*

276 on all of GTEx failed to detect this event in human tissues, we find evidence of *de novo* splice junction

277 reads corresponding to both the conserved 3' and 5' splice sites of this unannotated exon that we

278 validated previously in mouse (Figure 3B), suggesting this exon is also included in human brain tissues.

279   This region of the splicegraph is complex, however, and is defined by overlapping LSVs each

280 with multiple splice junctions and intron retention detected (Figure 3B: exon 11 source LSV, left; exon

281 13 target LSV, right). While the LSV formulation has several benefits, including accurate PSI

282 quantification of complex splicing patterns involving more than two splice junctions[3], it is difficult for

283 users to know which junction quantifications and combinations of junctions from different LSVs should

284 be combined to define common alternative splicing (AS) events, like the cassette exons described

285 above in *PTBP1*. Moreover, while certain annotated and *de novo* junctions may have sufficient read

286 coverage for detection and quantification by MAJIQ, they can be very lowly included in a user's

287 condition(s) of interest. For example, several hundred reads across GTEx brain samples support the

10

288 existence of the annotated, intron distal alternative 3'ss of exon 12 of *PTBP1*, but source LSV

289 quantification of the relative usage of this junction is low across all samples (Figure 3B, left. Blue

290 junction median PSI across samples of $< 5\%$ in all tissue groups). Such junctions add additional

291 complexity to the splicegraph and may hinder definition of common AS event types across the

292 transcriptome.

293 To overcome these limitations and to facilitate downstream, transcriptome wide analysis of

294 common AS event types we developed the VOILA Modulizer (Figure 3C). First, users have the option

295 to simplify the splicegraph to remove junctions that do not meet a threshold for raw read coverage,

296 low inclusion levels across the input samples ($\mathbb{E}\left[\Psi\right]$), and/or low relative splicing changes between

297 input comparisons between sample groups ($\mathbb{E}\left[\Delta\Psi\right]$) (Figure 3Ci). This helps remove junctions that do

298 not meet a user's desired threshold for biological significance and facilitate downstream event

299 definitions, like the alternative 3'ss of exon 12 of *PTBP1* discussed above with low inclusion levels

300 across all sample groups (blue junction in Figure 3B, left). Next the simplified splicegraph is traversed

301 to define single entry, single exit regions of the splicegraph that we call alternative splicing modules

302 (AS modules or ASMs), as shown for part of *PTBP1* (Figure 3Cii). Within each AS module, pattern

303 matching is performed between the remaining exon and junction structure of the simplified splicegraph

304 to each of 14 basic AS event types (Figure S3A). This process is illustrated in Figure 3Ciii for two AS

305 modules within *PTBP1*. We note that this step can lead to some redundant event information (e.g.

306 intron retention events sharing the same junction and intron coordinates, as in Figure 3C). Because

307 these events are quantified from both sides through a source and a target LSV, the quantification in

308 terms of PSI or dPSI between conditions may not agree and thus both are provided. Nonetheless,

309 downstream filtering can ensure agreement when counting event types and defining changing events.

310 Running the VOILA Modulizer produces a number of files based on event types with a uniform

311 structure containing coordinate and quantification for each sample group to facilitate downstream

312 analysis on AS modules and AS event types of interest (Figure S3A). Some AS event definitions

313 identified by the Modulizer are analogous to those defined by other splicing quantification algorithms

314 that only handle binary, classical splicing events (e.g. MISO[12] or rMATS[13]). However, the MAJIQ

315 + VOILA Modulizer approach adds a number of benefits compared to other available algorithms.

316 First, our approach allows for *de novo* splice junction and intron retention detection, which is crucial in

317 the context of GTEx brain subregions. Using a simplification threshold of median $\mathbb{E}\left[\Psi\right]$ over brain

11

318 tissue groups of $\geq 5\%$ to be included in the simplified splicegraph, we defined 32,435 AS modules

319 where 70.6% contain at least one unannotated splice junction and/or intron retention (Figure 3D,

320 Figure S3B). The AS module formulation also allows for definition of common splicing patterns across

321 brain subregions beyond binary splicing events, which made up 59.2% of all AS modules. The

322 remaining 40% of AS modules contained multiple AS events which, in many cases, involved mixing of

323 a classical event type with intron retention (Figure S3B). Both at the AS event level (Figure 3D) and

324 at the AS module level (Figure S3B), intron retention was particularly common using our

325 simplification threshold of median $\mathbb{E}[\Psi]$ of greater than 5% in any one brain tissue group. This is

326 consistent with previous studies that have found neuronal tissues to have very high levels of intron

327 retention compared to other contexts[14].

328   Initial analysis of the most common AS module types led us to add additional splicing event

329 patterns to our definitions, beyond those that are classically defined in other tools (intron retention,

330 cassette exon, alternative 3' and 5'ss, alternative first and last exons, tandem cassette exons, and

331 mutually exclusive exons[12, 13]). These included putative alternative first and last exons, where at

332 least one alternative exon is created from a *de novo* junction that does not belong to any nearby exon,

333 and putative alternative 3' or 5'ss, where a cassette exon has an inclusion junction removed during

334 simplification (low inclusion) with sufficiently high intron retention levels (see Figure S3A for full

335 details). These new splicing event types participated in the make up of 13.1% of AS modules or 11.6%

336 of AS events overall detected in the brain (Figure 3D, event types marked with asterisks). Importantly,

337 the Modulizer outputs all of these event types in a format amenable to downstream regulatory

338 analysis, which will facilitate the future characterization of these splicing patterns (Figure S3A).

339 Analysis of unique cerebellar splicing patterns highlights known and novel regulatory

340 programs

341 Finally we wished to use MAJIQ + VOILA Modulizer to analyze differential splicing patterns between

342 brain subregions. Previous studies focused on splicing quantitative trait loci within GTEx brain tissues

343 found the cerebellar tissues cluster separately from other brain subregions based on splicing[15]. Our

344 analysis of *PTBP1* (Figure 3B) and pairwise analysis of the number of significant LSVs according to

345 MAJIQ HET further supports distinct splicing patterns in cerebellar tissues (Figure S4A). For these

346 reasons we sought to identify AS modules and events with unique splicing patterns in the cerebellum.

347 Using the above AS module definitions from all junctions and introns with group level median

348 $\mathbb{E}[\Psi] > 5\%$, we next searched for consistent splicing changes between the two cerebellar tissues

349 (cerebellum and cerebellar hemisphere) and other brain subregions using MAJIQ HET. We required an

350 absolute difference in median $\mathbb{E}[\Psi]$ values of 20% or more when comparing both cerebellar tissue

351 groups to the same other brain region tissue group in addition to having a Wilcoxon rank-sum

352 $p < 0.05$ (Figure 4A, see Methods).

353 From these comparisons we found 3,995 unique, changing AS modules (Figure 4B) comprising

354 over 7,500 changing AS events (Figure S4B). At the changing AS module and AS event levels, intron

355 retention was prevalent, followed by cassette exons and other mixtures of binary AS event types with

356 intron retention (Figure 4B). As with the analysis based on inclusion levels alone (Figure 3), most

357 changing AS modules (53.3%) consisted of multiple, binary AS event types (Figure 4B), highlighting

358 the prevalence of complex splicing changes and the power of MAJIQ + VOILA Modulizer approach.

359 Alternative splicing regulation of cassette exons in neuronal tissues is very well studied with a

360 number of expression changes associated with splicing factors (e.g. expression of the RBFOX family,

361 down regulation of PTB proteins, expression of NOVA proteins, etc.)[9, 10]. For this reason we wished

362 to analyze the regulatory signature around the cassette exons defined from our MAJIQ HET + VOILA

363 Modulizer analysis to see if we could capture known, and potentially novel, regulatory motifs around

364 cerebellar cassette exons.

365 Our initial analysis focused on all changing cassette exon (CE) events. This mirrors the CE

366 landscape that would be identified by other splicing quantification algorithms and consists of a

367 combination of CEs which come from modules consisting of only a single CE event in addition to those

368 from complex modules with multiple event types (Figure 4B, arrowhead, Figure 4C, top). Because

369 RNA binding proteins bind short motifs and splicing factor binding that results in alternative splicing

370 regulation typically occurs proximal to the splice sites of an alternative exon[16], we performed a

371 Z-score analysis for hexamer occurrence within 300 nucleotides upstream or downstream of cerebellar

372 changing cassette exons versus those alternative exons that did not change between brain subregions

373 (see Methods). Moreover, because splicing factors typically act in position-specific manners (e.g.

374 binding downstream of a cassette exon enhances exon inclusion while binding downstream represses

375 inclusion)[17, 18], we further separated cassette events into those with increased exon inclusion in

13

376 cerebellar tissues (Figure 4C, blue) and those with increased exon exclusion in cerebellar tissues

377 (Figure 4C, red) when compared to other brain subregions.

378   Supporting the validity of our approach, this analysis uncovered a number of motifs either

379 upstream or downstream of cerebellar cassette exons with known links to neuronal splicing regulation.

380 For example, for cerebellar inclusion cassettes we found a number of CU-rich and UGC containing

381 hexamers upstream and the RBFOX-binding-motif, UGCAUG[19], enriched downstream (Figure 4C,

382 blue). SRRS6/nSR100 is known to bind UGC-containing sequences upstream of neuronal microexons

383 to enhance their inclusion with the aid of SRSF11 that binds CU-repeat sequences[20]. Accordingly,

384 motif maps across our different cerebellar exon classes based on hexamers shown to bind SRRS6[21]

385 and SRSF11[20] by iCLIP show clear enrichment of these motifs just upstream of cerebellar inclusion

386 cassette exons. This result is consistent with increased expression of these two genes in cerebellar

387 tissues leading to enhanced intronic splicing enhancer (ISE) activity around these events (Figure

388 S5A-E).

389   In addition to SRRS6 and SRSF11, the RBFOX family is highly expressed in neuronal tissues

390 and is known to enhance exon inclusion when it binds downstream of the 5'ss[22, 23] (Figure S5G).

391 Indeed, we find a strong enrichment of the known UGCAUG-binding site just downstream of cerebellar

392 inclusion events (Figure 4C, blue). This result is consistent with increased expression of these genes

393 and increased ISE activity in the cerebellum versus other brain subregions (Figure S5F-H).

394   Interestingly, we found hexamers containing motifs known to bind QKI (e.g. ACUAA

395 containing[24]) were enriched around both cerebellar inclusion (upstream) and exclusion events

396 (downstream) (Figure 4C). QKI is known to act as a splicing enhancer when it binds downstream of

397 cassette exons and represses exonic inclusion when it binds upstream[24]. We generated a motif map

398 of the QKI hexamer (ACUAAY[25]) around these exon classes and found clear positional enrichment

399 proximal to the regulated splice sites in both exon sets (Figure 4D, top). Moreover, we generated RNA

400 maps of in vivo binding events (determined by CLIP peaks) of QKI across multiple cell types and

401 found enriched binding consistent with the motif maps (Figure 4D, bottom, Figure S5I). Compared to

402 other brain subregions, the two cerebellar tissues exhibited lowest expression of QKI (Figure 4E). This

403 result points to a regulatory mechanism by which decreased expression of QKI in cerebellum may

404 contribute to both cerebellar exon exclusion events (loss of enhancing activity downstream leading to

405 exon skipping) and cerebellar exon inclusion events (loss of repressive activity upstream leading to

14

406   inclusion) (Figure 4F).

407   Given that many regulated cassette exons occur within AS modules containing other AS event

408   types (Figure 4B), we next wished to explore if regulatory motifs differed between these subsets.

409   Because AS modules containing cassette exons and those containing both cassette exon and intron

410   retention events are common (Figure 4B, Figure S3B, S4B), we chose to stratify the set of all cassette

411   exons into those that contained a regulated intron retention event and those that occurred in AS

412   modules in which intron retention was not detected. We calculated Z-scores for hexamers from these

413   exon subsets by comparing them against the set of exons that were not changing in cerebellar

414   comparisons and compared the results of the two analyses. Figure 4G shows an example of this

415   analysis for hexamers located downstream of cerebellar exclusion cassette exon subsets. The top two

416   hexamers that match QKI binding motifs (ACUAAC and CUAACG) found when analyzing all CE

417   events (Figure 4C) also had the highest Z-scores in the intron retention regulated and no intron

418   retention CE subsets (Figure 4G, green). On the other hand, several of the G- and C-rich motifs that

419   were enriched downstream of all CE cerebellar exclusion events (Figure 4C) were biased towards higher

420   Z-scores solely in the CE subset that contained regulated intron retention (Figure 4G, orange). This is

421   consistent with observations from previous studies analyzing intron retention events that found

422   retained introns tended to be more G/C-rich when compared to non-retained introns[14]. Motif maps

423   across the different cerebellar exclusion CE sets supported the Z-score analysis and highlight that the

424   enrichment of G-rich sequences (Figure 4H, top) and C-rich sequences (Figure S6A,B) around all

425   cerebellar exclusion CEs is driven mostly by the subset of CEs containing a regulated intron retention

426   event (compare dashed orange and dashed fuchsia lines). The QKI hexamer showed similar positional

427   enrichment downstream of both CE subsets (Figure 4H, bottom).

428   Similar results were seen when comparing Z-scores for upstream and downstream hexamers

429   identified in the all cassette exon analysis (Figure 4C) of cerebellar inclusion and exclusion CE subsets

430   stratified by intron status (Figure S6A). While some of the motifs found in the complete CE analysis

431   scored similarly in subsets stratified by intron retention status (e.g. the RBFOX hexamer or SRRS6

432   hexamers around cerebellar inclusion exons), others showed biased enrichment in CEs with regulated

433   intron retention compared those with no intron retention (e.g. CU-repeat hexamers) (Figure S6).

434   Overall, this analysis highlights some shared and distinct regulatory features of cerebellar cassette

435   exons with and without evidence of intron retention.

Discussion

The work presented here represents the culmination of continuous development of MAJIQ since its original release in 2016[3]. The original MAJIQ, like many other algorithms, was designed for comparing relatively small groups of RNA-seq from biological replicates. However, as we demonstrate here using GTEx v8, datasets nowadays can easily grow to hundreds and thousands of non-replicate samples. The sheer size and heterogeneous nature of such data poses challenges that go beyond just algorithm efficiency. Additional challenges include the ability to capture but also simplify *de novo* and complex splicing variations, the ability to define subtypes over such complex splicing events, and the ability to visualize and process such events and subtypes for downstream analysis. MAJIQ v2 is the only algorithm, to the best of our knowledge, that supports such features through efficient implementation of several algorithmic innovations we introduced here: The simplifier, the modulizer, incremental build options, and the VOILA v2 visualization package. In addition, we perform extensive comparison of MAJIQ v2 to other algorithms, create a resource for reproducible algorithm comparison in the form of both data and software package, and demonstrate the utility of the new splicing analysis features by performing a detailed analysis of differential splicing between more than 2,300 samples from GTEx v8 brain subregions.

The algorithmic contributions in this work include a new method to quantify *de novo* intron retention, an incremental build, addition of the MAJIQ HET statistics which do not assume a shared PSI between samples in a group, and the modulizer in VOILA. The resulting new features enhance splicing analysis, especially on larger datasets. For example, MAJIQ's incremental build saves much of the processing needed when adding new samples to existing repositories. Labs or centers can thus process data such as GTEx once, then efficiently add more relevant samples as needed. Notably, our performance evaluations discussed below show performance for the first analysis, but subsequent analyses can be expected to be even faster. Furthermore, as these datasets get larger, we also expect to see more *de novo* junctions. These junctions increase the complexity of the splicegraph and the size of splicing events considered. The MAJIQ simplifier enables users to more finely control how this complexity enters the analysis.

The new features of MAJIQ v2 are accompanied by matching ones in VOILA v2 visualization and analysis package. The VOILA Modulizer provides a new view of splicing changes on the

16

465 splicegraph in terms of splicing modules. These complex units can be broken down into classical

466 splicing events that may share similar splicing regulation, but allow for a more refined classification

467 than traditional approaches, as demonstrated in our analysis of brain subregions. In contrast, tools

468 that only list classical events (e.g. rMATS) quantify those solely based on the reads within these event

469 definitions. Consequently, reads outside these event definitions, which can greatly alter the splicing

470 quantification, are ignored. In addition, the VOILA viewer now allows for interactive visual analysis

471 and supports a server version, allowing large analyses performed on cluster or cloud environments to

472 be viewed without downloading large datasets locally.

473 In terms of performance we showed MAJIQ v2 compares favorably to available methods. In

474 terms of efficiency, we showed MAJIQ v2 is as fast and memory efficient as the two most efficient

475 tools, rMATS-turbo and LeafCutter. This is a notable achievement given that MAJIQ is the only tool

476 amongst those that offers detection and quantification of *de novo* intron retention. Accounting for IR

477 in splicing analysis is computationally expensive but nonetheless important in many settings as we

478 discuss below.

479 On synthetic data, MAJIQ and LeafCutter were the only two tools that simultaneously

480 demonstrated both low FDR and FNR when identifying genes with differential splicing. We note that

481 our usage of LeafCutter included additional filtering for $\Delta\Psi > 20\%$ beyond the default p-value based

482 filtering as we found that the default settings performed much worse[26]. Whippet was the only other

483 tool that also exhibited low FNR, but it demonstrated FDRs over 20%. Our results suggest that many

484 genes called as differentially spliced by Whippet, rMATS, and SUPPA are false discoveries.

485 Furthermore, it suggests that rMATS and SUPPA miss a substantial fraction ($> 40\%$) of the genes

486 that it should call as differentially spliced.

487 On real RNA-seq data from GTEx we found MAJIQ outperformed the other tools. Specifically,

488 MAJIQ's reproducibility, measured using the reproducibility ratio, was consistently higher than all

489 other tools. The difference between MAJIQ and other tools was particularly striking when comparing a

490 small number of samples but persisted even when comparing 50 vs 50 samples. Comparing MAJIQ

491 HET introduced here to MAJIQ dPSI from[3], we found both to have similar reproducibility, but HET

492 offered a significant increase in detection power. While LeafCutter was comparable to MAJIQ on the

493 synthetic dataset, we found that its reproducibility on real data was not, exhibiting reproducibility

494 lower than rMATS and comparable to SUPPA2. When using intra-to-inter ratio (IIR) to assess false

17

495 discovery, we found IIR approached 0 when considering larger numbers of samples for all tools.

496 However, for very small sample numbers of 3 vs 3, only MAJIQ and LeafCutter achieved IIR below

497 10%.

498 The extensive evaluations we performed here serve not just to assess the specific tools we

499 included, but as a service for the community. First, we created the largest synthetic RNA-seq dataset

500 to date, with over 300 samples. In contrast to many other works, the data generated here was based

501 on real life GTEx samples. It also does not reflect MAJIQ's model and was based instead on

502 transcript-based quantifications by other algorithms (RSEM). As such, we would expect it to benefit

503 tools that are built around a similar model (e.g. SUPPA). A second contribution is the evaluation

504 package we created, validations-tools. This package allows users to not only reproduce our results but

505 also to easily add future tools and repeat the analysis for future developers or for anyone who wants to

506 assess performance on their own unique dataset. We highly recommend researchers and cores to take

507 advantage of this as it is possible that on a dataset with other characteristics the various algorithms

508 would perform differently. Finally, we note that the efforts to create reproducible results in genomics

509 and specifically for tool development are constantly ongoing. We previously documented in detail

510 issues we identified with using using outdated software, software misuse, and lack of reproducibility for

511 analysis scripts and data that severely affected software assessment, including MAJIQ[26]. We hope

512 the reproducibility tools we included here will help avoid such issues and make it easier for future

513 developers to achieve at least the "bronze" level of reproducibility as was recently proposed[27].

514 Finally, applying our improved pipelines to GTEx brain subregions allowed us to define the

515 complex alternative splicing patterns observed across over 2,300 heterogeneous human neuronal tissue

516 samples from 374 donors and 13 tissue groups. Our approach and subsequent analysis offers several

517 advances compared to previous efforts. For example [15] also analyzed differential splicing in brain

518 subregion splicing but included only annotated, classical splicing events identified by rMATS. Several

519 other GTEx analyses use LeafCutter's framework and focus on detecting sQTLs. Our work advances

520 these efforts through improved quantification accuracy (described above) and our LSV based

521 approach, which is the only method able to capture *de novo* and complex splicing events as well as

522 retained introns (IR). Furthermore, as we illustrated here for cerebellum specific regulation, our newly

523 introduced definition of AS modules and AS event types greatly facilitate downstream regulatory

524 analysis.

525   Applying MAJIQ HET and AS subtypes from the VOILA modulizer allowed us to discover

526 additional, novel complexity within transcripts for the crucial splicing regulator, PTBP1, including a *de*

527 *novo*, premature stop codon containing exon in human that we previously validated in mouse brain

528 subregions[3]. This exon was preferentially included in cerebellar tissues, leading us to focus on the

529 cerebellar specific splicing program. Our regulatory analysis on cerebellum specific cassette exons

530 highlighted many known splicing regulators previously shown to be essential in neuronal splicing

531 programs (i.e. the RBFOX family, SRRS6 with SRSF11, PTBP1, and QKI[9, 10, 20]), highlighting the

532 validity of our approach and definitions of cassette exons based on complex LSVs. Crucially, the

533 MAJIQ + VOILA Modulizer approach allowed us to stratify this superset of cassette exon events into

534 different subsets based on the presence or absence of other AS event types within the module (e.g.

535 CEs with or without intron retention). While some motifs are shared and similarly enriched around

536 CEs with and without regulated intron retention (e.g. RBFOX and QKI), other motifs were specifically

537 enriched in the intron containing subset only. In the case of cerebellar exon exclusion events, the signal

538 for G/C rich motifs observed on the superset of all CEs was driven entirely by the subset of CEs

539 containing intron retention events. We anticipate the new ability we introduced here to interrogate AS

540 modules made up of combinations of AS event types will facilitate future regulatory discoveries in

541 other datasets from additional biological contexts.

542   We note that there are key limitations to the regulatory analysis we performed for

543 cerebellar-specific splicing, which was based solely on bulk tissue RNA-seq experiments from GTEx.

544 Previous work leveraging single cell data to deconvolute bulk GTEx tissues into their relative cell type

545 compositions suggests that cerebellar tissues contain relatively larger proportions of neurons compared

546 to other brain subregions[28]. This fact can confound the interpretation of our results in terms of

547 neurobiology as neurons are known to express certain splicing factors (e.g. RBFOX3/NeuN, SRRS6),

548 which may explain the cerebellar splicing pattern we observed here. Thus, future directions for

549 improving MAJIQ involve accounting for cell type heterogeneity as well as combining long reads for

550 isoform specific deconvolution. Other promising directions for future exploration include analysis of

551 RNA sequencing for clinical diagnostics and exploiting MAJIQ's advantages for improved sQTL

552 analysis.

553   In summary, we introduced here a significant update to the original MAJIQ package. MAJIQ

554 v2 empowers fast, detailed, and accurate analyses of large heterogeneous RNA-seq datasets and is

19

555 already supporting a highly active user group spanning hundreds of labs, centers and companies across

556 the world. Our analysis of brain subregions provides a compelling example of such analysis on over

557 2,300 human neuronal tissue samples leading to several novel findings related to cerebellum specific

558 splicing regulation. We hope the analysis we performed, along with the tool, data, and evaluation

559 package we supply will inspire many more researchers to delve into splicing regulatory analysis in their

560 own data and make exciting new discoveries.

Methods and Materials

MAJIQ builder

563 In this subsection, we review how the MAJIQ builder prepares the structure and observations per

564 experiment that are used for downstream splicing quantification as part of a scalable and principled

565 approach to splicing analysis of large numbers of experiments. We describe the MAJIQ builder's new

566 approach for estimating intron read rates, which allows junction and intron coverage to be calculated

567 once and reused efficiently for multiple analyses, unlike other methods that quantify intron retention.

568 We also describe the MAJIQ simplifier, which reduces the complexity of the structural models of

569 splicing used in quantification that especially arises from the analysis of large and heterogeneous

570 datasets.

571 MAJIQ encodes the set of all possible splicing changes for a gene in terms of a splicegraph. A

572 splicegraph is a graph-theoretic representation of a gene's splicing decisions from one exon to another,

573 with exons as vertices and junctions and retained introns as distinct edges connecting exons. The

574 exons of each gene are non-overlapping genomic intervals. Each junction has a source and target exon

575 with a position within each exon, indicating the positions that are spliced together when the junction

576 is used. Retained introns are between adjacent exons and indicate that intron retention between the

577 exons is possible.

578 MAJIQ first constructs each gene's splicegraph by parsing transcript annotations from a GFF3

579 file. Exon boundaries and junctions from each transcript for a gene are combined in order to produce

580 the minimal splicegraph that includes each transcript's annotated exons and junctions, splitting exons

581 by retained introns to ensure that each junction starts and ends in different exons. MAJIQ then

582 updates the splicegraph with *de novo* junctions and introns found from processing input RNA-seq

20

583 experiments' junction and intron coverage.

584   MAJIQ processes aligned input RNA-seq experiments to per-position junction and intron

585 coverage in the following way. First, MAJIQ identifies reads with split alignments. The genomic

586 coordinates of each split corresponds to a potential junction. Meanwhile, the coordinate of the split on

587 the aligned read is the junction's "position" on the read. MAJIQ counts the number of reads for each

588 junction from each possible position. Afterwards, MAJIQ identifies reads that contiguously intersect

589 known or potential introns (i.e. reads that intersect the genomic coordinates between adjacent exons

590 without splits within the intron boundaries). If the intron start is contained in the aligned read, the

591 intron "position" is defined as for junctions (treating the exon/intron boundary as a junction with zero

592 length). For aligned reads intersecting the intron but not the start, additional positions are defined by

593 the genomic distances of the first positions of the aligned reads to the intron start. These additional

594 positions per intron increase the number of ways aligned reads can intersect introns in comparison to

595 junctions. To adjust for this and model intron read coverage similarly to junction read counts, MAJIQ

596 aggregates together adjacent intron positions to the equivalent number of possible positions per

597 junction, taking the mean number of reads per reduced positions.

598   MAJIQ uses the obtained junction and intron coverage to update the splicegraph in the

599 following way. Each potential junction is mapped to matching genes by prioritizing (1) genes that

600 already contain the junction (i.e. annotated junctions) over (2) genes where both junction coordinates

601 are within 400bp of an exon, which are prioritized over (3) genes where the junction is contained

602 within the gene boundaries. The input experiments are divided into user-defined build groups. MAJIQ

603 adds a *de novo* junction to the splicegraph if there is sufficient evidence for its inclusion in one of the

604 build groups. This happens when the total number of reads and total number of positions with at

605 least one read exceeds the user-defined minimum number of reads and positions in at least a minimum

606 number of experiments. MAJIQ adds new *de novo* exons or adjusts existing exon boundaries to

607 accommodate the added *de novo* junctions as previously described. Potential introns are added to the

608 splicegraph under similar criteria, and their boundaries are adjusted or split to accommodate the

609 adjusted or *de novo* exon boundaries.

610   Since processed intron coverage is averaged over the entire original intronic region, we can

611 carry over the same coverage as an estimate for all resulting splicegraph introns, which are contained

612 in the original intron's boundaries. In contrast, MAJIQ's previous approach, which is also used by most

21

613 other tools that quantify intron retention, quantified intron coverage using local counts of unsplit

614 reads sharing the position of known junctions. These local counts must be calculated using

615 information from all processed experiments (for all *de novo* junctions), which requires samples to be

616 reprocessed each time an analysis with different samples are performed. MAJIQ's new approach allows

617 intron coverage to be processed once and used for multiple builds with potentially different intron

618 boundaries. This enables MAJIQ's new incremental build feature, which saves intermediate files with

619 junction and intron coverage that can be calculated once and reused instead of BAM files for multiple

620 builds. This reduces storage and time processing experiments that are part of multiple analyses.

621 While MAJIQ uses raw totals of read rates and number of nonzero positions for adding

622 junctions and introns to the splicegraph, the MAJIQ builder performs additional modeling of

623 per-position read rates for use in quantification. First, we mask positions with zero coverage and with

624 outlier coverage. Outlier coverage is assessed under the observation that per-position read rates

625 generally follow a Poisson distribution. For each junction/position, we use all other positions with

626 nonzero coverage for that junction to estimate the Poisson rate parameter. Then, MAJIQ calls any

627 position with an extreme right-tailed p-value (default 10^{-7}) under this model an outlier and ignores its

628 contribution to coverage for quantification. Second, we perform bootstrap sampling of the total read

629 rate over unmasked positions in order to model measurement error of true read rates. Under the

630 assumption that each unmasked position is identically distributed, MAJIQ performs nonparametric

631 sampling with replacement to draw from a distribution with identical mean and variance as the

632 observed positions (see supplementary note). Since we assume that our read rates are generally

633 overdispersed relative to the Poisson distribution, MAJIQ replaces nonparametric sampling with

634 Poisson sampling when the nonparametric estimate of variance is less than the mean (i.e.

635 underdispersed).

636 MAJIQ performs quantification of splicing events modeled as LSVs, which are defined by a

637 splicegraph. A source (target) LSV is defined for an exon as a choice over the incoming (outgoing)

638 edges to (from) that exon from (to) a different exon. In general, only LSVs with at least two edges are

639 considered. MAJIQ builder prepares output files with raw and bootstrapped coverage for each

640 junction/intron in each LSV for quick use by downstream quantifiers.

641 We observed that builds from many build groups or with high coverage tend to have

642 increasingly complex splicegraphs and LSVs with many junctions. Many of these junctions are often

22

643 lowly used in all the samples but were included in the splicegraph because they had enough raw reads

644 and positions (noisy *de novo*) or are part of an unused annotated transcript. This motivated the

645 MAJIQ simplifier, which allows junctions and introns to be masked from the final splicegraph used for

646 quantification. After the splicegraph is constructed using all input build groups, MAJIQ calculates the

647 ratio of the raw read rate for each junction/intron relative to the other junctions/introns in each LSV.

648 If a junction has consistently low coverage in each of the build groups relative to the other choices in

649 the two LSVs it can belong to, it is "simplified" and removed from the final splicegraph. This reduces

650 the complexity of the final splicegraph and quantified LSVs, making output files smaller and

651 downstream quantification more efficient.

652 In summary, the MAJIQ builder combines transcript annotations and input RNA-seq

653 experiments in order to build a splicegraph encoding all possible splicing events consistent with both

654 annotations and data and to prepare read coverage for quantification in terms of LSVs. The MAJIQ

655 builder's new approach for estimating intron read rates allows junction and intron coverage to be

656 calculated once and reused as part of an incremental build for multiple analyses, unlike other methods

657 that quantify intron retention. The MAJIQ builder also introduces an approach for simplifying the

658 complexity that arises in splicing events when processing large numbers of experiments. Overall, this

659 allows the MAJIQ builder to produce structural models of possible splicing events and read coverage

660 for downstream quantification that scale to the setting of large numbers of RNA-seq experiments.

MAJIQ quantifiers

662 MAJIQ provides three methods for quantifying RNA-seq experiments. MAJIQ PSI, MAJIQ dPSI, and

663 MAJIQ HET, which we introduce in this paper. MAJIQ PSI and dPSI, which were previously

664 described in [3], quantify groups of experiments that are assumed to be replicates with a shared true

665 value of PSI per group. MAJIQ PSI estimates a posterior distribution of PSI (Ψ) for a single group,

666 while MAJIQ dPSI compares these distributions for two groups in order to estimate a posterior

667 distribution for dPSI ($\Delta\Psi$). MAJIQ HET compares two groups of samples but drops the replicate

668 experiments assumption, enabling analysis of more heterogeneous samples. Instead, experiments are

669 quantified individually and groups are compared under the assumption that the true values of PSI are

670 identically distributed between the two groups.

All three pipelines share the same underlying machinery for inferring posterior distributions for

23

Ψ. Formally, Ψ for a junction in an LSV is defined as the fraction of expressed isoforms using the junction out of all expressed isoforms containing the LSV. This fraction is not directly observable. Instead, we observe the number of reads aligned r_j to each junction j in the LSV. We model each r_j as a realization of a binomial distribution over the isoforms with probability Ψ_j:

$$r_j \sim \text{Binomial}\left(\sum_{j \in \text{LSV}} r_j, \Psi_j\right). \tag{1}$$

We take a Bayesian approach to integrate prior knowledge of Ψ, allowing for improved estimation when there is low read coverage. This requires a prior distribution on Ψ. We previously observed that most values of Ψ are nearly zero or one, which can be modeled using a generalization of the Jeffrey's prior for an LSV with J junctions:

$$\Psi_j \sim \text{Beta}\left(\frac{1}{J}, 1 - \frac{1}{J}\right). \tag{2}$$

This prior is conjugate to the binomial likelihood, allowing for efficient closed-form estimation of the posterior distribution of Ψ_j given the observed number of reads:

$$\Psi_j \big| \{r'_j : j' \in \text{LSV}\} \sim \text{Beta}\left(\frac{1}{J} + r_j, 1 - \frac{1}{J} + \sum_{j' \neq j} r'_j\right). \tag{3}$$

671 Since MAJIQ build obtains bootstrap replicates of observed read rates, we perform this posterior

672 inference on each set of bootstrap replicate read rates to obtain an ensemble of posterior distributions.

673   For MAJIQ PSI, we obtain this ensemble of posteriors for replicate experiments by adding the

674 observed read rates from the experiments that pass more stringent reads and position thresholds than

675 the builder. MAJIQ PSI treats the average of the posterior distributions as a final distribution over Ψ.

676 It reports point estimates of Ψ as the mean of this distribution ($\mathbb{E}[\Psi]$) and saves a discretized version

677 of the distribution for visualization in VOILA.

    MAJIQ dPSI takes this a step further by using the posterior distributions on Ψ_1, Ψ_2 for two groups in order to compute $\Delta\Psi = \Psi_2 - \Psi_1$ between the two groups. We start by computing the distribution of $\Delta\Psi$ under the assumption of independence of Ψ_1 and Ψ_2 by marginalizing the product

of their distributions:

$$\mathbb{P}_{\text{ind}}\left(\Delta\Psi\right) = \sum_{\Psi_2-\Psi_1=\Delta\Psi} \mathbb{P}\left(\Psi\right)_1 \mathbb{P}\left(\Psi\right)_2. \tag{4}$$

We know that Ψ_1 and Ψ_2 are not independent, so we integrate our knowledge that $\Delta\Psi$ is usually close to zero as a prior on $\Delta\Psi$. Following our previous work, we formulate our prior $\mathbb{P}_{\text{prior}}\left(\Delta\Psi\right)$ as a mixture of three components: (1) a spike around $\Delta\Psi = 0$, (2) a broader centered distribution around $\Delta\Psi = 0$, and (3) a uniform slab. We determine our final posterior distribution on $\Delta\Psi$ by adjusting $\mathbb{P}_{\text{ind}}\left(\Delta\Psi\right)$ by the prior and renormalizing:

$$\mathbb{P}\left(\Delta\Psi\right) \propto \mathbb{P}_{\text{ind}}\left(\Delta\Psi\right) \mathbb{P}_{\text{prior}}\left(\Delta\Psi\right). \tag{5}$$

MAJIQ dPSI computes point estimates of $\Delta\Psi$ using the posterior mean of the distribution ($\mathbb{E}\left[\Delta\Psi\right]$) and identifies confidence of measured changes in inclusion as posterior probabilities $\mathbb{P}\left(|\Delta\Psi| > C\right)$.

MAJIQ HET takes a different approach for comparing inclusion between two groups of experiments. MAJIQ HET drops the assumption of replicate experiments to consider heterogeneity in Ψ between experiments within a group. Instead, MAJIQ HET assumes that the values of Ψ per experiment in each of the groups come from the same distribution. We evaluate this assumption using null hypothesis significance testing. Null hypothesis significance testing is performed using one (or more) of four tests: (1) Welch's two-sample t-test, (2) Mann-Whitney U test, (3) Total Number of Mistakes (TNOM) test, and (4) InfoScore test. Welch's two-sample t test and Mann-Whitney U test are well-documented elsewhere[29, 30]. Our implementation of Mann-Whitney U test computes exact p-values when there are at most 64 experiments and computes asymptotic p-values using normal approximation with tie and continuity correction for larger samples. Meanwhile, the InfoScore and TNOM tests are adapted from ScoreGenes[31]. The TNOM test evaluates how well a single threshold on PSI can discriminate between the observed values in the two groups. The Total Number of Mistakes is the minimum number of misclassified observations under the best possible thresholds. The distribution on TNOM when the distributions are equal are calculated using the closed-form formula in [32] to obtain p-values. Similarly, the InfoScore test evaluates how well a single threshold discriminates between groups, but, instead of measuring misclassifications directly, it identifies the threshold with the highest mutual information between the threshold and the true group labels. MAJIQ HET uses the

697   dynamic programming algorithm in [32] to evaluate the distribution of InfoScore under the null

698   hypothesis in order to obtain p-values. All four tests require observed values of Ψ per experiment,

699   which is not directly observed. MAJIQ HET accounts for variable uncertainty per experiment in our

700   estimations of Ψ by repeated sampling of Ψ from the posterior distributions of quantified samples.

701   MAJIQ HET computes the p-value for each repeated sample of Ψ over quantified experiments and

702   reports the 95th-percentile over the resulting p-values. These p-value quantiles are not calibrated, so

703   MAJIQ HET also computes p-values with the posterior means of Ψ. MAJIQ HET also reports the

704   median of the observed posterior means of Ψ for each group. These p-values and the difference

705   between the median observed posterior means are used together downstream in VOILA for the

706   identification of high-confidence differentially spliced LSVs.

VOILA

708   VOILA provides a suite of post-processing and visualization tools designed to allow researchers to

709   make use of MAJIQ quantifications directly, or easily format and filter the output for passing to other

710   post-processing tools.

711   The VOILA viewer acts as a complete visualization tool for interactive analysis of output from

712   MAJIQ PSI, dPSI, or HET. It includes search and filter mode for all discovered LSVs, as well as an

713   in-depth viewer for the full splicegraph of a gene and all of the LSVs found within it. When using the

714   VOILA viewer with output from MAJIQ HET, VOILA will also automatically generate heatmaps for

715   each LSV with the to quickly indicate the discovered $\Delta\Psi$ and statistical results from each group

716   comparison. The viewer frontend runs completely within a web browser interface, so it is able to

717   function with similar results on any modern operating system without installation of special

718   frameworks or system libraries. The viewer can also be configured to run as a standalone web server

719   such that the interactive results can be easily shared with collaborators. Tutorials and parameters are

720   made available to integrate VOILA with a wide range of common web server production software.

721   VOILA also has a number of modes for filtering and rearranging data into a number of human

722   and machine-readable files. Determining confidently non-changing (background) and confidently

723   changing events is one of the primary use cases. We define highly-confident non-changing events from

724   MAJIQ HET as being (1) above a nominal p-value threshold, (2) within-group variance is sufficiently

725   low as measured by IQR, and (3) between-group $\Delta\Psi$ is sufficiently low as measured by difference in

726 medians. We accept that the between-group $\Delta\Psi$ threshold may be redundant in combination with the

727 other two thresholds. We define confident changing events from MAJIQ HET as being (1) below a

728 p-value threshold and (2) between-group $\Delta\Psi$ is sufficiently high as measured by difference in medians.

729   In addition to the basic text output modes, there is a separate comprehensive output mode

730 dedicated to finding specific event types/patterns called the VOILA Modulizer. The VOILA Modulizer

731 searches for a large number of relevant patterns, both common and complex.

732   Each set of events is delimited on the basis of AS "modules" found by MAJIQ in each analyzed

733 gene. Modules refer to areas of the splicegraph between single entry (one junction path, diverges to

734 two or more) and single exit (all junction paths converge back to one).

735   Inside each of the AS "modules" detected by the modulizer, smaller AS "events" (sub patterns

736 matching specific known organizations of junctions or introns) are then categorized. Currently, the list

737 of potential patterns we match to find an event is fixed to a specific set, which can be found in Figure

738 S3. All events which do not match any known splicing pattern are dumped to an "other" category

739 which may be of possible interest in rare cases.

740   Modulizer supports any number or combination of MAJIQ experiments as input, in the form of

741 PSI, dPSI, and/or HET VOILA files. These are used for narrowing modules to form around junctions /

742 introns we find relevant, as well as to verify which AS modules and AS events are changing or

743 non-changing, based on coverage, Ψ, and differences in Ψ ($\Delta\Psi$). All filters may be disabled or

744 adjusted.

745   At a high level, Modulizer uses a sequential pipeline for filtering and assembling output. First,

746 all junctions and introns are read, and any which do not pass the reads, PSI, and/or dPSI thresholds

747 are immediately removed from consideration. Then, using the remaining introns, and junctions,

748 Modulizer identifies AS modules by looking for genomic locations with single-entry / single-exit as

749 previously described. Then, Modulizer filters and removes modules which do not pass criteria such as

750 not being sufficiently changing, lack of LSVs, or being constitutive. After filtering, Modulizer performs

751 pattern matching for each AS event type on each AS module to identify all component AS events.

752 Finally, Modulizer scans the input VOILA files for relevant quantifications in order to produce output

753 TSV files for each individual AS event type, a high-level summary of all events found in each

754 discovered module, and a summary of quantifications per module suitable for generating a heatmap

755 according to the user's filtering criteria (e.g. the shortest discovered junction within the AS module to

756 represent the inclusive AS product, the most changing junction in the AS module from HET and/or

757 dPSI inputs, etc.).

Sample selection from GTEx

759 We selected from GTEx in the following way. We required all samples to have a RIN score of greater

760 than 6. For performance evaluation we chose to evaluate a comparison between cerebellum and

761 skeletal muscle. We randomly selected 150 samples from both tissues, excluding the same donor from

762 being selected in both tissues. For the brain subregions analysis, we selected all samples in GTEx v8

763 associated with brain tissue (not including pituitary gland). We also performed another analysis with

764 all tissues in GTEx v8 using 30 or less samples per tissue. Samples were downloaded as FASTQ or as

765 BAM and converted to FASTQ depending on when they were released. Samples that were part of v7

766 are available on SRA, so they were downloaded using SRA Tools (v2.9.6) as FASTQ files. New

767 samples from the v8 release are only available as BAMs on the cloud, so they were downloaded and

768 converted to FASTQ using samtools (v1.9).

Simulated RNA-seq as ground truth

770 We used the expression quantification data from the GTEx v8 release as the basis for our simulations.

771 Briefly, we downloaded the transcript quantification table

772 (`GTEx_Analysis_2017-06-05_v8_RSEMv1.3.0_transcript_tpm.gct.gz`) and the gene-level

773 quantification table (`GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_reads.gct.gz`) from

774 the GTEx portal (`https://www.gtexportal.org/home/datasets`). To match how the GTEx

775 consortium performed these analyses, we downloaded the GRCh38 build of the reference genome

776 sequence and gene models from v26 of the GENCODE annotation.

777 We selected 300 samples from GTEx to serve as the basis for 300 simulated samples, each real

778 sample providing the expression distribution underlying one simulated sample. To run BEERS, we first

779 need to prepare four configuration files that are customized for the desired dataset: geneinfo, geneseq,

780 intronseq, and feature quants. The geneinfo, genenseq, and intronseq files define the structure and

781 sequence information for each simulated transcript. As a result, these three files are determined solely

782 by the choice of reference genome build and annotation. The feature quant files are specific to each

28

783 individual sample and define a distribution of transcript-level expression. First, we used the genome

784 sequence and gene models to create the geneinfo, geneseq, and intronseq files. Since the genome is

785 fixed across all simulated samples. We used the same set of these files to simulate all GTEx-derived

786 samples. Next, we extracted TPM values for each sample from the GTEx transcript quantification

787 table and used these distributions of TPM values to generate separate BEERS feature quant config

788 files for each simulated sample. Lastly, to determine the total number of reads to simulate for each

789 sample, we used the gene-level quantification file to count the total number of gene-mapping reads in

790 each GTEx sample.

791 To simulated strand-specific reads with uniform coverage across no errors, substitutions, or

792 intron retention events, we ran the BEERS simulator using the following command-line options:

793 `-strandspecific -outputfq -error 0 -subfreq 0 -indelfreq 0 -intronfreq 0 -palt 0`

794 `-fraglength 100,250,500.`

795 We transformed ground-truth transcript abundances into ground-truth splicing quantifications

796 for each splicing quantification tool, taking into account the tools' differing definitions of splicing

797 events. First, we defined ground-truth abundances for each exon or junction by adding the abundances

798 of all transcripts including the exon or junction. Then, for each tool, we adopted their splicing event

799 definitions, mapping the exon/junction abundances to compute their splicing quantifications.

800 **MAJIQ**

801 MAJIQ reports splicing quantifications with respect to LSVs. Therefore, ground-truth values for PSI

802 were calculated by dividing the ground-truth abundance of each junction by the sum of the

803 ground-truth abundances for all junctions in each LSV.

804 **rMATS**

805 rMATS reports a different format file per event type. But since all of them are classical binary event

806 types, all can be reduced to two paths events, inclusion and exclusion. Each file contains the exon that

807 defines each of the ways, so we calculate the Ψ_{gt} as inclusion/(inclusion + exclusion) using the exon

808 transcript combination to get the exons ground-truth abundances for all junctions in each LSV.

29

LeafCutter

LeafCutter reports splicing quantifications with respect to intron clusters composed of several junctions. Ground-truth values for LeafCutter's splicing ratios were calculated using ground-truth junction abundances, similar to MAJIQ.

SUPPA2

SUPPA2 reports classical events similarly to rMATS. So the approach we use here is similar to that tool. The main difference is that SUPPA2 reports the junctions coordinate in each one of the paths, so we use those junctions ground truth quantification to obtain the Ψ_{gt} as inclusion / (inclusion + exclusion).

Whippet

Whippet outputs a psi.gz that contains the psi quantification of an event. That PSI is their formulation of the quantification from inclusion and exclusion paths. Differently to SUPPA2 or rMATS, Whippet combines a set of junctions to define a path, emulating in that way a transcript (or a portion of it). So, in order to find Ψ_{gt} of those paths, we look for those transcripts that include all the junctions (and virtual junctions). We combine the expression of those transcripts to find the Ψ_{gt} of each path.

RNA-seq sample preprocessing before splicing analysis

We aligned RNA-seq reads from real and simulated GTEx samples to the human genome for splicing analysis with MAJIQ and other tools using the following procedure. Simulated GTEx samples were generated as pairs of FASTQ files. We performed quality and adapter trimming on each sample using TrimGalore (v0.4.5). Some tools require reads aligned to the genome. For these tools, we used STAR (v2.5.3a) to perform a two-step gapped alignment of the trimmed reads to the GRCh38 primary assembly with annotations from Ensembl release 94. Other tools required transcript quantifications relative to annotated transcripts. For these tools, we used Salmon (v0.14.0) using the trimmed samples to estimate transcript abundances.

30

Performance evaluations

We wrote a package of evaluation scripts, called validations-tools, in order to compare MAJIQ in terms of speed, memory footprint, accuracy, and reproducibility for each one of the following tools: rMATS, LeafCutter, SUPPA2, and Whippet. This package was written to allow future users to not only reproduce our results but to easily add future tools and repeat these kinds of analyses with different datasets.

We adjusted the tools parameters following recommendations by each tool's authors. Specific parameters are listed in Table S2. For these comparisons, we evaluated the methods' computational efficiency and ability to identify splicing differences.

First, we evaluated computational efficiency of the different methods. We evaluated computational efficiency in terms of runtime and peak memory usage. Not all tools provide an extensive log of their execution, so, in order to measure wall time and memory usage, we used the output of '/usr/bin/time -v'. We ran each method for all pairs comparisons between 10 groups with increasing sample sizes on an Ubuntu Linux environment with 32 cores (Intel Xeon 2.7GHz and 64GB RAM).

Second, we evaluated the different methods' performance in quantifying splicing differences on simulated and real datasets. On the simulated datasets, where we know ground-truth differences in splicing between transcripts, we calculated true and false positive rates for the identification of splicing differences by each method. However, on real datasets, where no ground-truth is available, it is not possible to calculate true or false positive rates. Instead, we evaluated two metrics, reproducibility ratio (RR) and intra-to-inter ratio (IIR), on real (and simulated for comparison) data. The first metric, RR, measures the internal consistency of differential splicing tools. This internal consistency is reflected in the assumption that each tool should identify roughly the same events when repeating a comparison between two groups using different samples. We quantify this by performing two such comparisons and computing the fraction of the top n differentially-spliced events in the first comparison that are also in the top n events of the second comparison. This produces a "reproducibility-ratio" curve, $RR(n)$ for the method as a function of the number of top events. If the first comparison yields N "significant" events, RR(N) is called the reproducibility ratio. For the specific case of MAJIQ, we note that in order to comparisons of LSV-type events more comparable to

31

862   classic AS events such as used by rMATS, we filtered out overlapping LSVs (i.e. those that share

863   junctions) in order to avoid double-counting classic AS events. For example, a classic exon-skipping

864   event would have matching source and target LSVs that overlap. However, we note that this filtering

865   only reduces N_A but does not affect the reproducibility curves (apart from extending to a different

866   value of N_A) (Fig. S7). Although reproducibility of a method on real data is a scientifically important

867   goal, it is not a sufficient goal because highly biased methods can be highly reproducible. To address

868   this limitation, the second metric, IIR, is based on the principle that comparisons between (inter-) two

869   groups should have many more significant events than comparisons within (intra-) a group.

870   Furthermore, significant events within the group are likely false positives. This is quantified by

871   computing the ratio of the number of significant events from an intra-group comparison to the number

872   of significant events from an inter-group comparison. We evaluated these metrics for each tool with

873   varying sample sizes to identify which methods outperformed each other in different settings.

874   **Event-level evaluations**

875   In these evaluations we check reproducibility and accuracy of reported differentially spliced events by

876   the various tools shown in Figure 2. As we describe in the main text, each tool defines alternative

877   splicing events differently so that direct comparison of the events or their number between tools is not

878   possible. Thus, when using real data each method was assessed by its own set of reported events to

879   compute reproducibility ratios (RR) and intra to inter ratio (IIR) as in Figure 2D,E.

880   In contrast, when using GTEx based simulated data we do have the "ground truth" (denoted

881   "gt" below) for the abundance of each transcript. We thus use these values to summarize Ψ and $\Delta\Psi$

882   observed in each method reported AS events and assess accuracy using the following definitions:

883   - True Positive: $\max \Delta\Psi_{\text{tool}} \geq 20\%$ and $\text{pvalue}_{\text{tool}} \leq 0.05$ and $\max \Delta\Psi_{\text{gt}} \geq 20\%$

884   - True Negative: $\max \Delta\Psi_{\text{tool}} < 5\%$ and $\text{pvalue}_{\text{tool}} > 0.05$ and $\max \Delta\Psi_{\text{gt}} < 5\%$

885   - False Positive: $\max \Delta\Psi_{\text{tool}} \geq 20\%$ and $\text{pvalue}_{\text{tool}} \leq 0.05$ and $\max \Delta\Psi_{\text{gt}} < 5\%$

886   - False Negative: $\max \Delta\Psi_{\text{tool}} < 5\%$ and $\text{pvalue}_{\text{tool}} > 0.05$ and $\max \Delta\Psi_{\text{gt}} \geq 20\%$

887   - Ambiguous: all other cases (when either $\Delta\Psi \in [5\%, 20\%)$ or when $\Delta\Psi$ and pvalue reported by

888   the tool conflict),

32

889 where max is taken over all the junctions that belong to the AS events.

890 The above definitions were used to assess accuracy at the event level for each method, as

891 shown in Figure S1, and also served as the base for gene level evaluations described below.

Gene-level evaluations

893 To facilitate more direct comparison between the different methods shown in Figure 2 we aggregated

894 each tool AS events and their respective annotation as TP, TN, FP, and FN as given above to assess

895 gene level performance. Naturally, gene level labels of TP, TN, FP and FN are defined based on the

896 events they contain. The gene level labels are easy to define as positive or negative when all AS events

897 embedded in it are considered positive or negative respectively. The problem arises when a gene has

898 some of its events as false positives and false negatives. In that case, we prioritize the labels according

899 to the following order: FP, FN, TP, TN. This means for example that an occurrence of a false positive

900 event in a gene (according to the method's specific event definition) would be counted as a false

901 positive gene even if some other events were correctly labeled as true negative or even true positives.

902 The rationale for this prioritization is that (a) positive events are expected to be rare and (b) we care

903 the most about trying to validate or follow up on wrong hits (false positives) followed by missing true

904 changes (false negatives).

GTEx brain subregion analysis

MAJIQ HET and VOILA Modulizer on brain subregions

907 MAJIQ HET was run on all 13 choose 2 (78) pairwise comparisons from GTEx v8 brain tissue groups

908 and the results were visualized with VOILA. Significant LSVs were those considered to be those

909 containing at least one junction or intron with an absolute difference in group median $\mathbb{E}[\Psi]$ values of

910 20% or more between the two tissue groups and all four HET statistics (Wilcoxon, InfoScore, TNOM,

911 and t-test) with $p < 0.05$.

912 The VOILA Modulizer was run on the resulting outputs with the following options:

913 `--decomplexify-psi-threshold 0.05` to remove all junctions and introns from the splicegraph

914 that had tissue group median E(PSI) of less than 5% across samples for every group; `--show-all` to

915 include all AS modules and AS events in the output, not just those meeting the changing criteria.

Default values were used for other options that flag changing AS modules and AS events in the output. For changing: a minimum absolute median difference in PSI between groups of 20% or more for the primary threshold and a p-value of less than 0.05 across all four MAJIQ HET statistics (Wilcoxon, InfoScore, TNOM, and t-test). For non-changing: a maximum absolute median difference in PSI of 5% or less between groups; a maximum interquartile range in PSI within a group of 10% or less; and a p-value of 0.05 or greater across the MAJIQ HET statistics).

PSI based AS module and AS counts across the brain

Counting of AS modules based on the initial PSI simplification across the 13 brain tissue groups was done by parsing the resultant VOILA Modulizer summary file. This file is organized by AS module and lists the number of each of the 14 AS event types, outlined in Figure S3A, contained in each. AS modules were classified and counted based on the presence or absence of each of the 14 AS event types. Certain AS event type definitions overlap. Specifically, every tandem cassette exon containing AS module will also contain a multi exon skipping AS event and every putative 5' or putative 3'ss AS module will also contain an intron retention event. In these cases, the additional, partially redundant AS event type was added to the AS module classification if and only if their count within the module was larger than the count of the AS event they overlap with. For example, for an AS module to be classified as containing both tandem cassette exon (TCE) and multi exon skipping events (MES), the number of MES events within the module must be greater than the number of TCE events.

Cerebellar AS module and AS event definitions

Given the large number of LSV-based splicing differences between the two GTEx cerebellar tissues (cerebellum and cerebellar hemisphere) and the other brain subregions according to MAJIQ HET comparisons (Figure S4A), we wished to define AS modules and AS events based on these comparisons. These two cerebellar tissues were derived from sampling in duplicate, with cerebellar hemisphere sampled during initial tissue collection (frozen) and cerebellum sampled after the brain was received at the brain bank (PAXgene)[33]. Therefore, we focused our analysis on AS modules and AS events that displayed changes between both cerebellar tissues and one of the other subregions. For example, a cassette exon AS event would have to be labeled as changing according to the VOILA Modulizer filters (minimum absolute median difference in PSI between groups of 20% or more for the

34

944 primary threshold and a p-value of less than 0.05) in both cerebellum versus cortex and cerebellar

945 hemisphere versus cortex to be counted. We defined all such consistent, changing cerebellar AS events

946 from the 14 AS event files output by the VOILA Modulizer and used these to count the number of

947 modules containing each AS event type or combination of types.

948 **Cerebellar cassette exon regulatory analysis**

949 To perform regulatory analysis around exons with differential cerebellar inclusion patterns we first

950 defined a high confidence set of cassette exons (CEs) by applying additional filters to those described

951 above. In addition to the primary filter of an absolute median difference in PSI of 20% or more

952 between a cerebellar tissue and another brain subregion for one junction in the CE event, a secondary

953 threshold of an absolute median difference in PSI of 10% or more was enforced for all four junction

954 quantifications of the CE (i.e. the inclusion source LSV junction quantification, the inclusion target

955 LSV junction quantification, and the shared exclusion junction quantified in both the source and target

956 LSV). Next we enforced that the direction of change between the two exclusion junction

957 quantifications and the two inclusion junction quantifications agreed in their direction of change in

958 cerebellar versus other tissues. If both inclusion junction qualifications increased in cerebellar tissues

959 and both exclusion junction quantification decreased, this was considered a cerebellar inclusion CE

960 events. The opposite directions were considered cerebellar exclusion CE events. Non-changing CE

961 events were defined as those flagged as non-changing by the VOILA Modulizer in every comparison of

962 both cerebellar tissues versus the other 12 brain tissues. For CE subset analysis, CE with intron

963 retention (IR) events were those where one or more of the CE junctions was also involved in a

964 changing IR event in cerebellar versus other tissues. CE with no IR events were those CE events that

965 came from modules without any IR events detected.

966 For sequence analysis we extracted GRCh38 sequences for intronic regions 300 nucleotides (nts)

967 upstream and 300 nts downstream of every CE in each set. We calculated Z-scores by comparing the

968 occurrence of each hexamer in the upstream intronic region in each cerebellar set of regulated CEs

969 versus the non-changing set of CEs. This was repeated for the downstream intronic region as well.

970 Motif maps were generated to visualize position specific enrichment of particular hexamers of

971 interest. Each hexamer, or set of hexamers, were searched for over sliding windows of 20 nts in the

972 splice site proximal regions around the CE (i.e. intronic region 300 nt upstream of the 3'ss plus 50 nt

35

downstream and 50 nt upstream of the 5'ss plus the intronic region 300 nt downstream). The

frequency of occurrence was determined in each CE set and plotted using a running mean of 5 nts for

smoothing.

      RNAmaps for CLIP based binding of QKI were plotted in a similar way over the same splice site

proximal regions. BED narrowPeaks files were downloaded for ENCODE eCLIP data [34] from

encodeportal.org for QKI in K562 cells (accession ENCSR366YOG) or QKI in HepG2 cells (accession

ENCSR570WLM) and replicate files were concatenated. BED narrowPeaks for uvCLAP data for QKI-5

in HEK293 cells [35] were downloaded from GEO (accession GSE85155) and lifted over from GRCh37

to GRCh38. These peak coordinates were overlapped with CE splice site proximal regions and the

frequency of occurrence was assessed over the various cerebellar CE event sets at each position

proximal to CE splice sites.

References

[1]   Mar Gonzàlez-Porta et al. "Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene". In: *Genome biology* 14.7 (2013), pp. 1–11.

[2]   Barry Slaff et al. "MOCCASIN: A method for correcting for known and unknown confounders in RNA splicing analysis". In: *Nature communications* 12.1 (2021), pp. 1–9.

[3]   Jorge Vaquero-Garcia et al. "A new view of transcriptome complexity and regulation through the lens of local splicing variations". In: *elife* 5 (2016), e11752.

[4]   Scott S Norton et al. "Outlier detection for improved differential splicing quantification from RNA-Seq experiments with replicates". In: *Bioinformatics* 34.9 (Dec. 2017), pp. 1488–1497. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx790. eprint: https: //academic.oup.com/bioinformatics/article-pdf/34/9/1488/25417002/btx790.pdf. URL: https://doi.org/10.1093/bioinformatics/btx790.

[5]   Yin Hu et al. "DiffSplice: the genome-wide detection of differential splicing events with RNA-seq". In: *Nucleic Acids Research* 41.2 (Nov. 2012), e39–e39. ISSN: 0305-1048. DOI: 10.1093/nar/gks1026. eprint:

999   https://academic.oup.com/nar/article-pdf/41/2/e39/25342067/gks1026.pdf. URL:

1000   https://doi.org/10.1093/nar/gks1026.

1001 [6] Christopher J Green, Matthew R Gazzara, and Yoseph Barash. "MAJIQ-SPEL: web-tool to

1002   interrogate classical and complex splicing variations from RNA-Seq data". In: *Bioinformatics*

1003   34.2 (Sept. 2017), pp. 300–302. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx565.

1004   eprint: https://academic.oup.com/bioinformatics/article-

1005   pdf/34/2/300/33331720/btx565.pdf. URL:

1006   https://doi.org/10.1093/bioinformatics/btx565.

1007 [7] Qunhua Li et al. "Measuring reproducibility of high-throughput experiments". In: *The Annals of*

1008   *Applied Statistics* 5.3 (2011), pp. 1752–1779. DOI: 10.1214/11-AOAS466. URL:

1009   https://doi.org/10.1214/11-AOAS466.

1010 [8] Niroshika Keppetipola et al. "Neuronal regulation of pre-mRNA splicing by polypyrimidine tract

1011   binding proteins, PTBP1 and PTBP2". In: *Critical reviews in biochemistry and molecular*

1012   *biology* 47.4 (2012), pp. 360–378.

1013 [9] Bushra Raj and Benjamin J Blencowe. "Alternative splicing in the mammalian nervous system:

1014   recent insights into mechanisms and functional roles". In: *Neuron* 87.1 (2015), pp. 14–27.

1015 [10] Celine K Vuong, Douglas L Black, and Sika Zheng. "The neurogenetics of alternative splicing".

1016   In: *Nature Reviews Neuroscience* 17.5 (2016), pp. 265–281.

1017 [11] Serge Gueroussov et al. "An alternative splicing event amplifies evolutionary differences between

1018   vertebrates". In: *Science* 349.6250 (2015), pp. 868–873.

1019 [12] Yarden Katz et al. "Analysis and design of RNA sequencing experiments for identifying isoform

1020   regulation". In: *Nature methods* 7.12 (2010), pp. 1009–1015.

1021 [13] Shihao Shen et al. "rMATS: robust and flexible detection of differential alternative splicing from

1022   replicate RNA-Seq data". In: *Proceedings of the National Academy of Sciences* 111.51 (2014),

1023   E5593–E5601.

1024 [14] Ulrich Braunschweig et al. "Widespread intron retention in mammals functionally tunes

1025   transcriptomes". In: *Genome research* 24.11 (2014), pp. 1774–1786.

[15] Yida Zhang et al. "Regional variation of splicing QTLs in human brain". In: *The American Journal of Human Genetics* 107.2 (2020), pp. 196–210.

[16] Yoseph Barash et al. "Deciphering the splicing code". In: *Nature* 465.7294 (2010), pp. 53–59.

[17] Xiang-Dong Fu and Manuel Ares. "Context-dependent control of alternative splicing by RNA-binding proteins". In: *Nature Reviews Genetics* 15.10 (2014), pp. 689–701.

[18] A Gregory Matera and Zefeng Wang. "A day in the life of the spliceosome". In: *Nature reviews Molecular cell biology* 15.2 (2014), pp. 108–121.

[19] Jason G Underwood et al. "Homologues of the Caenorhabditis elegans Fox-1 protein are neuronal splicing regulators in mammals". In: *Molecular and cellular biology* 25.22 (2005), pp. 10005–10016.

[20] Thomas Gonatopoulos-Pournatzis et al. "Genome-wide CRISPR-Cas9 interrogation of splicing networks reveals a mechanism for recognition of autism-misregulated neuronal microexons". In: *Molecular cell* 72.3 (2018), pp. 510–524.

[21] Bushra Raj et al. "A global regulatory mechanism for activating an exon network required for neurogenesis". In: *Molecular cell* 56.1 (2014), pp. 90–103.

[22] Gene W Yeo et al. "An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells". In: *Nature structural & molecular biology* 16.2 (2009), pp. 130–137.

[23] Martin Jacko et al. "Rbfox splicing factors promote neuronal maturation and axon initial segment assembly". In: *Neuron* 97.4 (2018), pp. 853–868.

[24] Megan P Hall et al. "Quaking and PTB control overlapping splicing regulatory networks during muscle cell differentiation". In: *Rna* 19.5 (2013), pp. 627–638.

[25] Andre Galarneau and Stephane Richard. "Target RNA motif and target mRNAs of the Quaking STAR protein". In: *Nature structural & molecular biology* 12.8 (2005), pp. 691–698.

[26] Jorge Vaquero-Garcia, Scott Norton, and Yoseph Barash. "LeafCutter vs. MAJIQ and comparing software in the fast moving field of genomics". In: *bioRxiv* (2018). DOI: 10.1101/463927. eprint: https://www.biorxiv.org/content/early/2018/11/08/463927.full.pdf. URL: https://www.biorxiv.org/content/early/2018/11/08/463927.

[27]   Benjamin J Heil et al. "Reproducibility standards for machine learning in the life sciences". In: *Nature Methods* 18.10 (2021), pp. 1132–1135.

[28]   Margaret KR Donovan et al. "Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants". In: *Nature communications* 11.1 (2020), pp. 1–14.

[29]   Bernard L Welch. "The generalization of 'STUDENT'S'problem when several different population varlances are involved". In: *Biometrika* 34.1-2 (1947), pp. 28–35.

[30]   H. B. Mann and D. R. Whitney. "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other". In: *The Annals of Mathematical Statistics* 18.1 (1947), pp. 50–60. DOI: 10.1214/aoms/1177730491. URL: https://doi.org/10.1214/aoms/1177730491.

[31]   Yoseph Barash et al. "Comparative analysis of algorithms for signal quantitation from oligonucleotide microarrays". In: *Bioinformatics* 20.6 (Jan. 2004), pp. 839–846. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btg487. eprint: https://academic.oup.com/bioinformatics/article-pdf/20/6/839/634843/btg487.pdf. URL: https://doi.org/10.1093/bioinformatics/btg487.

[32]   Amir Ben-Dor, Nir Friedman, and Zohar Yakhini. "Overabundance analysis and class discovery in gene expression data". In: *Agilent Laboratories, Palo Alto, Tech. Rep* (2002).

[33]   GTEx Consortium et al. "The GTEx Consortium atlas of genetic regulatory effects across human tissues". In: *Science* 369.6509 (2020), pp. 1318–1330.

[34]   Eric L Van Nostrand et al. "A large-scale binding and functional map of human RNA-binding proteins". In: *Nature* 583.7818 (2020), pp. 711–719.

[35]   Daniel Maticzka et al. "uvCLAP is a fast and non-radioactive method to identify in vivo targets of RNA-binding proteins". In: *Nature communications* 9.1 (2018), pp. 1–13.

[36]   Ray Zhang et al. "A circadian gene expression atlas in mammals: implications for biology and medicine". In: *Proceedings of the National Academy of Sciences* 111.45 (2014), pp. 16219–16224.

[37]   Florian C Oberstrass et al. "Structure of PTB bound to RNA: specific binding and implications for splicing regulation". In: *Science* 309.5743 (2005), pp. 2054–2057.

39

Acknowledgments

Author contributions

YB conceived the project. JVG, JKA, SSN, and YB developed and tested the methodology for MAJIQ HET. JKA conceived and implemented the final methodology for quantification, sampling, and statistical testing in MAJIQ HET. JVG did the initial work porting Python code from MAJIQ v1 to MAJIQ v2. JVG and YB developed the new IR quantification algorithm. JKA formalized the approach for bootstrapping readrates. JVG and JKA implemented the updated MAJIQ Builder and MAJIQ PSI. JVG implemented MAJIQ dPSI. CJG and PJ implemented the VOILA viewer with input from MRG, CMR, and AJ. PJ, CMR, AJ, and YB developed the methodology for VOILA Modulizer. PJ implemented the VOILA Modulizer. NFL and GRG generated the simulated RNA-seq data for the performance comparisons. JVG carried out the performance comparisons vs other tools. MRG and CR performed the modules analysis. MRG conceived and carried out the brain subregions analysis. YB, JKA, and MRG wrote the final manuscript with input from PJ and JVG. All authors read and approved the final manuscript.

Competing interests

The MAJIQ software used in this study is available for licensing for free for academics, for a fee for commercial usage. Some of the licensing revenue goes to Y.B and members of the Barash lab.

Data and materials availability:

The code for MAJIQ and VOILA are available for academic/non-commercial use at https://majiq.biociphers.org/. Licensing information for commercial use can be found at https://majiq.biociphers.org/commercial.php. The code for validations-tools will be made available at https://bitbucket.org/biociphers/validations-tools. All processed data and code to reproduce figures will be deposited in a Zenodo before publication.

40

1111   GTEx data used for the analyses in this manuscript are available in dbGaP under accession
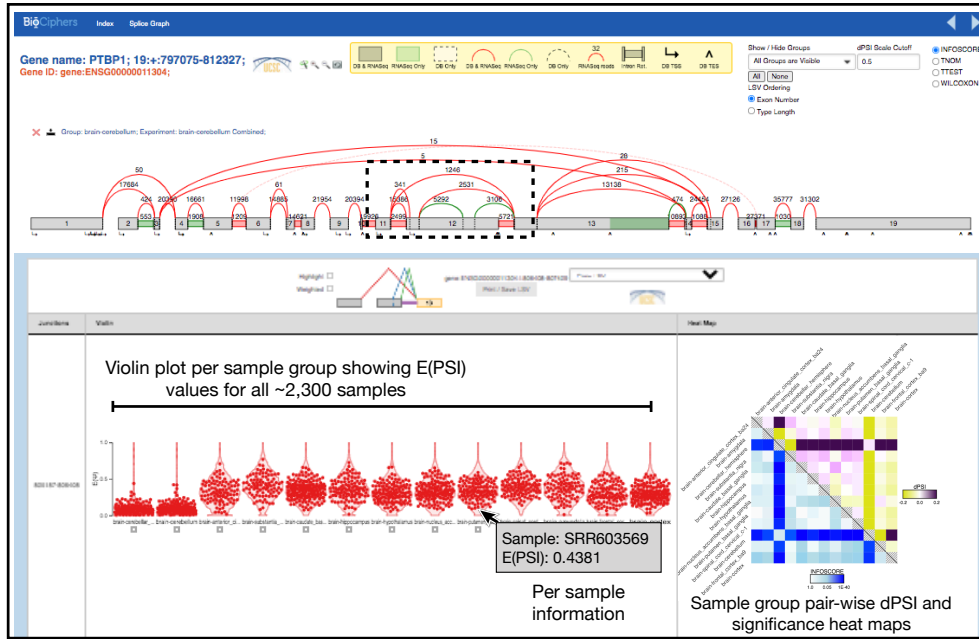
1112   phs000424.

Figures
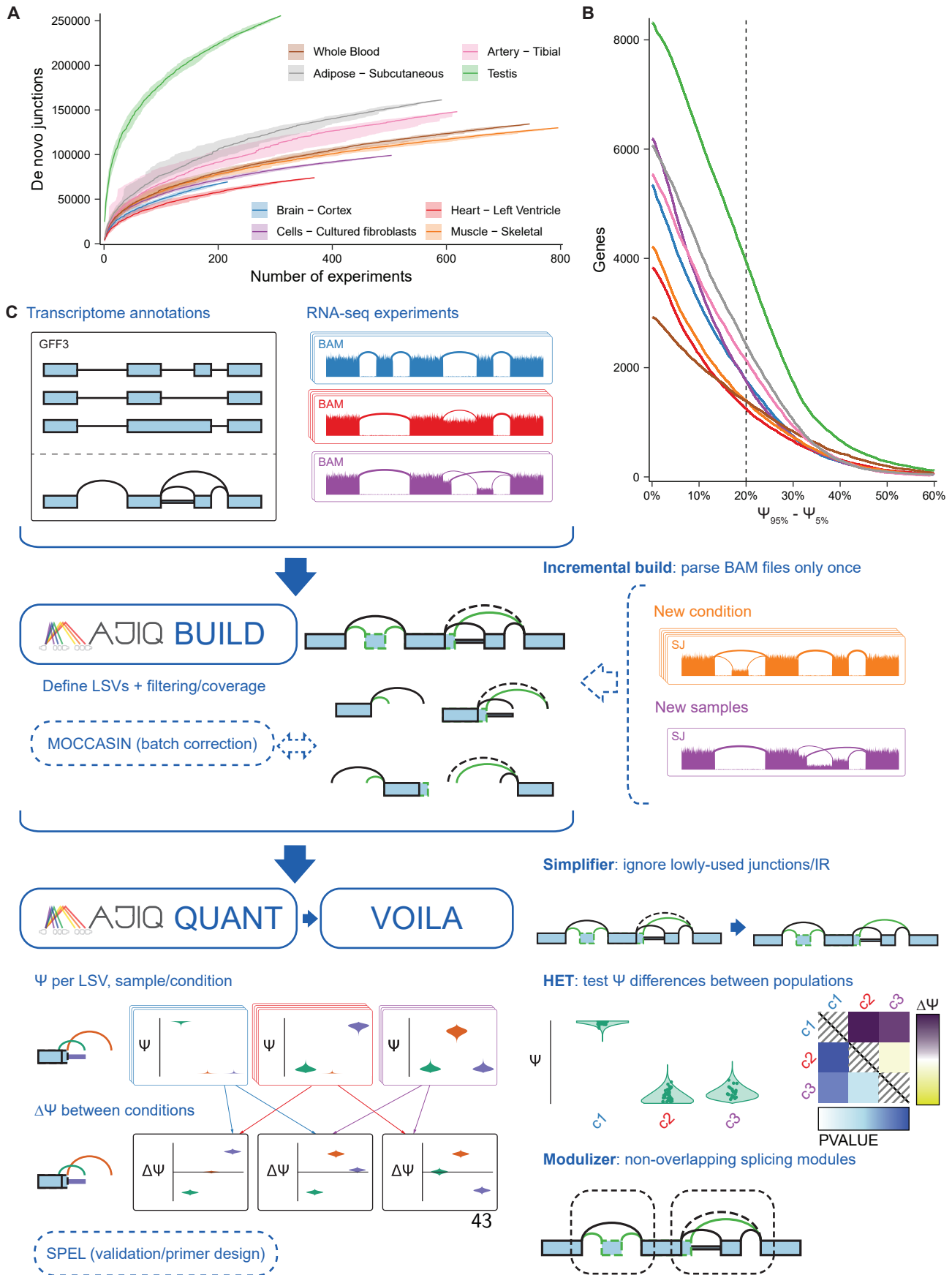
Fig. 1: MAJIQ efficiently and accurately models, quantifies, and visualizes RNA splicing from large and complex RNA-seq datasets. (A) The number of identified distinct unannotated *de novo* junctions increases with larger subsets of different tissues from GTEx. Lines show the median over 30 randomly selected permutations over experiments in each subset, confidence bands show the 5th to 95th percentiles over permutations of samples per tissue. **(B)** The number of genes with at least one junction where the difference between the 95th percentile and 5th percentile of PSI exceeds a given value for different tissues from GTEx (same tissues/colors as in Fig. 1A). Dashed vertical line indicates how many genes have a difference in PSI exceeding 20%. **(C)** MAJIQ combines annotated transcript databases and coverage from input RNA-seq experiments to build a model of each gene as a collection of exons connected by annotated and *de novo* junctions and retained introns (splicegraph). Junctions and retained introns sharing the same source or target exon form local splicing variations (LSVs). MAJIQ quantifies the relative inclusion of junctions and retained introns in each LSV in terms of percent spliced in (PSI, Ψ) and provides VOILA to make interactive visualizations of splicing quantifications with respect to each gene's splicegraph and LSV structures. MAJIQ v2 introduces an incremental build, which allows RNA-seq coverage to be read from BAM files only once to a coverage file (SJ), accelerating subsequent builds with different experiments. MAJIQ v2 introduces a simplifier, which can be used to reduce splicegraph/LSV complexity by ignoring lowly used junctions and retained introns. MAJIQ v2 introduces a new mode for quantification, HET, which compares PSI differences between populations of independent RNA-seq experiments and accounts for variable uncertainty per experiment. MAJIQ v2 introduces the modulizer, which allows performing analysis relative to non-overlapping splicing modules rather than LSVs.
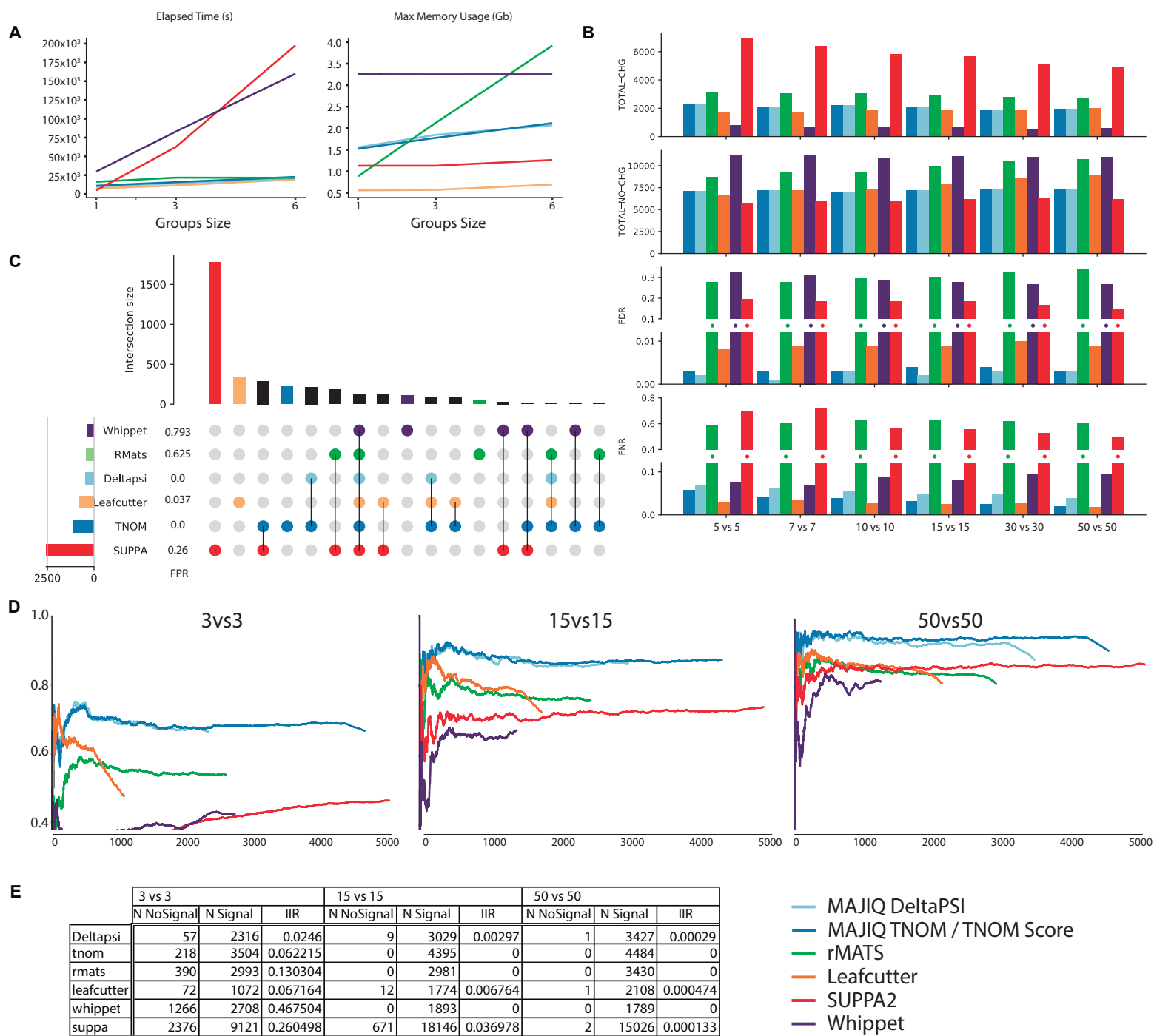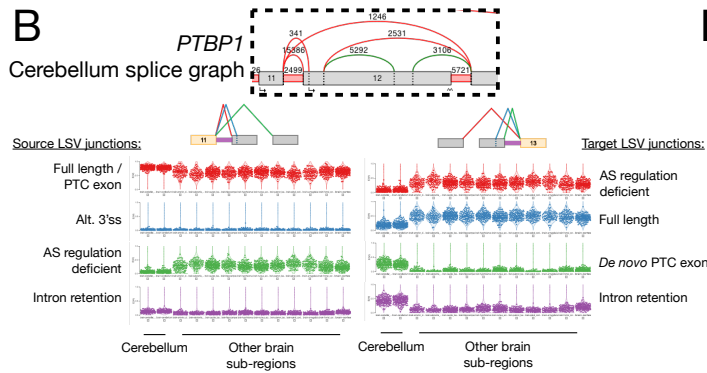
Fig. 2: Performance evaluation using synthetic and real data. (A) Time (left) and memory (right) consumption when analyzing multiple sample groups. Results shown are for running all pairwise differential splicing analysis between 10 tissue groups from GTEx v8 as the number of samples per group increases from 1 to 6 (x-axis). **(B)** Performance evaluation for differential splicing calls using simulated GTEx cerebellum and skeletal muscle samples and aggregated over genes (see main text and Methods). Metrics include the total number of genes reported as changing or non changing by each method, and the associated FDR and FNR. X-axis denotes the size of the groups. **(C)** Upset plot based on the 10vs10 analysis shown in (B). The bars on top represent the overlap between genes reported as differentially spliced by each method indicated below it. The bars and FPR values by each method name on the left refer to genes reported only by that method. **(D)** Reproducibility ratio (RR) plots for real data, using GTEx cerebellum and liver samples. Analysis here is based on each method's reported list of splicing events (not genes) and unique scoring approach. X-axis is the ranked number of events reported by each method and Y-axis is the fraction of those events reproduced within the same number of top-ranking events when repeating the analysis using a different set of samples from the same tissue groups. The length of the line represents the total number of differentially splicing events reported by each method (see Methods for details). RR graphs are shown for comparing group sizes of 3 (left), 15 (middle), and 50 (right). **(E)** Intra-to-Inter Ratio (IIR) results for GTEx samples as in (D). IIR computes the ratio between the number of events reported as significantly changing when comparing two sample groups of the same type (N No Signal column) and the number of events reported as significantly changing when comparing groups of different types (here GTEx liver and cerebellum samples as in (D)).
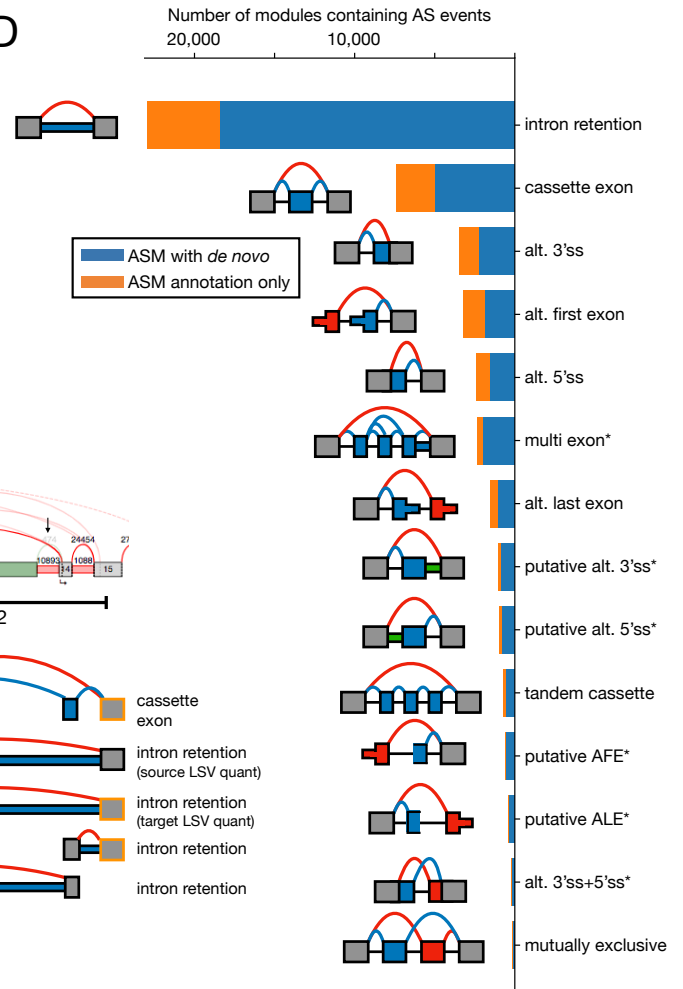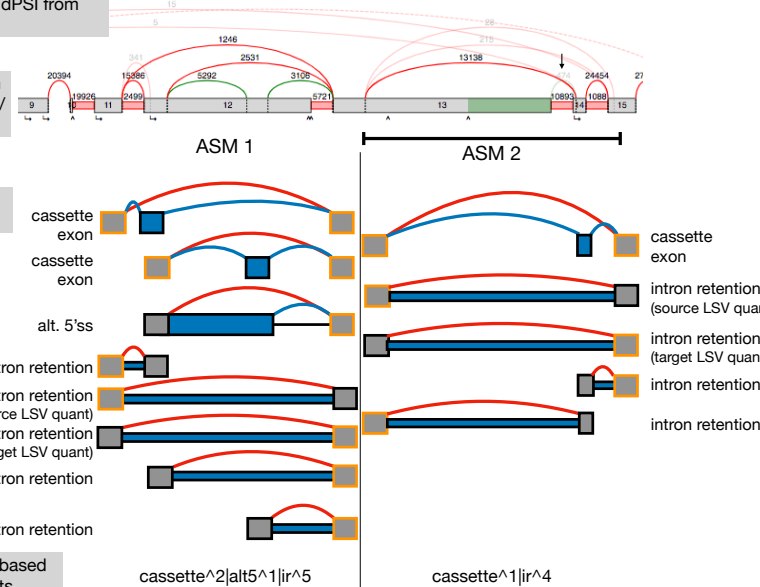
46

A Voila view for interactive analysis

B

PTBP1 Cerebellum splice graph

Source LSV junctions:

- Full length / PTC exon
- Alt. 3'ss
- AS regulation deficient
- Intron retention

Target LSV junctions:

- AS regulation deficient
- Full length
- *De novo* PTC exon
- Intron retention

Cerebellum / Other brain sub-regions

C Voila Modulizer

i) Remove junctions with low read support, E(PSI), and/or dPSI from splice graph

ii) Traverse splice graph and define single entry / exit AS modules

ASM 1 ASM 2

iii) Define AS events within each module

- cassette exon
- cassette exon
- alt. 5'ss
- intron retention
- intron retention (source LSV quant)
- intron retention (target LSV quant)
- intron retention
- intron retention

- cassette exon
- intron retention (source LSV quant)
- intron retention (target LSV quant)
- intron retention
- intron retention

iv) Categorize modules based on AS event components

cassette^2|alt5^1|ir^5   cassette^1|ir^4

D

Number of modules containing AS events

ASM with *de novo*
ASM annotation only

- intron retention
- cassette exon
- alt. 3'ss
- alt. first exon
- alt. 5'ss
- multi exon*
- alt. last exon
- putative alt. 3'ss*
- putative alt. 5'ss*
- tandem cassette
- putative AFE*
- putative ALE*
- alt. 3'ss+5'ss*
- mutually exclusive

47

Fig. 3: Enhanced visualization of large datasets and downstream analysis of alternative splicing modules with VOILA v2. (A) VOILA view of MAJIQ HET output for 13 brain tissue groups from GTEx from 2,335 RNA-seq samples originating from 374 unique donors. Top portion shows gene information and filtering criteria as well as the splicegraph for *PTBP1* showing combined reads from 225 cerebellum samples. Bottom portion displays visualization and PSI quantification for each junction in each LSV for the gene of interest. Here the distribution of $\mathbb{E}[\Psi]$ values across the indicated tissue groups is displayed as a violin beeswarm plot for the red junction for the exon f13 target LSV, represented in the cartoon, for all 2,335 RNA-seq samples. Individual sample information is given by hovering the cursor over individual points that represent each sample (gray box). Bottom right heatmap displays MAJIQ HET quantifications of all group pairwise comparisons across the 13 brain tissue groups to highlight significant splicing changes. Yellow to purple color scale on the top right indicates the expected $\Delta\Psi$ between tissue groups while blue color scale on the bottom left indicates the significance of the difference between group PSI distributions for one of four statistics used by MAJIQ HET (InfoScore displayed). **(B)** Top shows region of human *PTBP1* splicegraph (with reads from combined cerebellum samples) and two LSVs corresponding to a mammalian specific exon skipping event that alters PTBP1 splicing regulatory activity [11] (green junction in exon 11 source LSV, left; red junction in exon 13 target LSV, right) and *de novo* detection of a conserved, PTC-containing exon previously shown to be included in mouse neuronal tissues [3] (green junction in exon 13 target LSV). Bottom shows distribution of PSI across the 13 brain tissue groups as well as annotation of each junction. **(C)** VOILA Modulizer workflow (gray boxes) and an example region of the *PTBP1* splicegraph where junctions that did not meet a median $\mathbb{E}[\Psi]$ value of 5% or more in any of the 13 brain tissue groups were removed (arrows). Two alternative splicing modules (ASMs) were defined as single entry, single exit regions of the splicegraph and within these modules binary, AS events are defined. Gray exons highlighted in yellow indicate reference exons that belonged to LSVs for which MAJIQ quantification exists. Blue junctions and exonic or intronic regions indicate inclusion of the alternative region of the event and red junctions indicate exclusion of the alternative region. **(D)** Stacked bar chart showing the number of binary AS event types that make up AS modules across the 13 brain tissue groups from GTEx. AS event types are represented with a cartoon to the right of the chart and are named to the left of bars. Asterisks indicate non-classical AS event types. Each junction or intron had to have a median of $\mathbb{E}[\Psi]$ values of 5% or more across the samples of at least one tissue

1186 group to contribute to AS module definitions. Blue regions indicate AS events that contained *de novo*

1187 junctions and/or introns not found in the Ensembl transcriptome annotation (VERSION?) while

1188 orange regions indicate AS events containing only annotated junctions and introns.

50
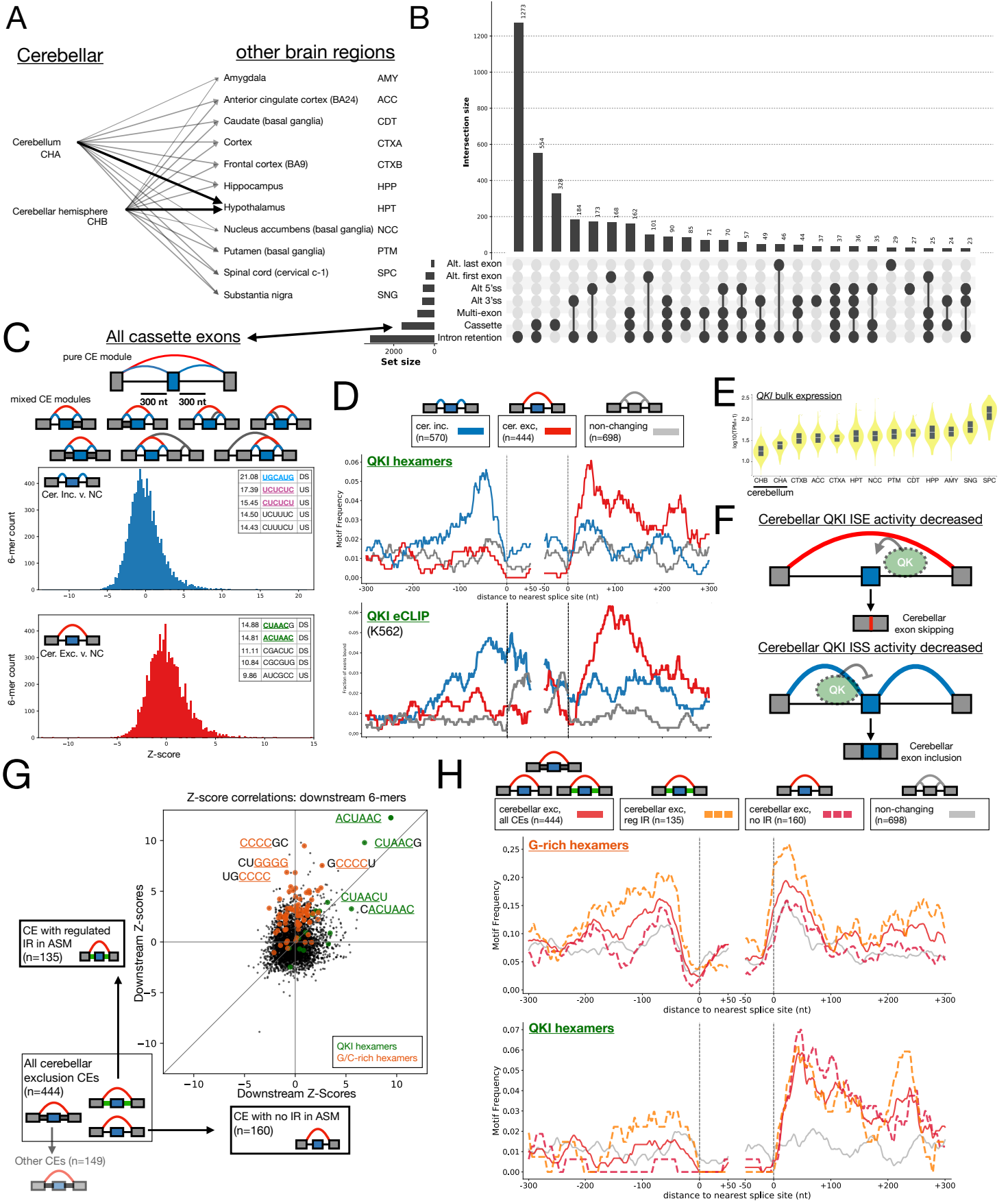
Fig. 4: MAJIQ HET + VOILA Modulizer defines the complex landscape of cerebellar splicing changes and regulation. (A) Pairwise comparisons run through MAJIQ HET to find significant splicing changes between GTEx cerebellar tissues (cerebellum and cerebellar hemisphere) versus the other 11 brain tissue groups. Dark arrows indicate an example of a consistent change, where both cerebellar tissue groups versus the same other brain region, the hypothalamus, shared a significant change. Alternative splicing modules (AS modules) were kept for downstream analysis if at least one such consistent comparison was significant (see Methods). GTEx abbreviations are given to the left of each tissue. **(B)** Upset plot showing the consistent, significantly changing AS event type(s) that make up AS modules. AS events had to have an absolute difference in median $\mathbb{E}[\Psi]$ of 20% or more when comparing both cerebellar tissue groups (cerebellum and cerebellar hemisphere) to the same other brain region tissue group in addition to having a Wilcoxon rank-sum $p < 0.05$ as reported by MAJIQ HET. **(C)** Top shows examples of all cassette exons (CEs) used for motif analysis. All CEs have quantified inclusion junctions (blue junctions) and a shared exclusion junction (red junction), potentially within a mixture of other AS event types (gray junctions and introns). Bottom shows distribution of Z-scores for hexamer motif occurrences within 300 nucleotides upstream or within 300 nucleotides downstream of all CE events when comparing CEs showing significant cerebellar inclusion versus CEs that did not change when compared to other brain regions (see Methods) (middle, blue) or when comparing CEs showing cerebellar exclusion versus non-changing CEs (bottom, red). Top motifs corresponding to putative binding sites of RBPs of interest are highlighted (QKI (green), RBFOX (light blue), SRRS6 (yellow), SRSF11 or PTB (purple)). All motifs and Z-scores are given in Table S1. **(D)** RNAmaps showing the frequency of QKI hexamer motif occurrence (top, ACUAAY frequency over sliding windows of 20 nucleotides with smoothing using a running mean of 5 nucleotides) or in vivo binding of QKI (K562 eCLIP peaks frequency, bottom) around cerebellar inclusion (blue), exclusion (red), or non-changing (gray) CEs. **(E)** *QKI* bulk tissue gene expression ($\log_{10}(1 + \text{TPM})$ for ENSG00000112531.16) sorted by median brain tissue expression. Chart generated using `gtexportal.org`. **(F)** Model for QKI position dependent regulation in GTEx brain tissues. Decreased expression of QKI in cerebellar tissues results in a decrease in downstream intronic splicing enhancer (ISE) activity of QKI, leading to cerebellar exon exclusion (top, red), and a decrease of upstream intronic splicing silencer (ISS) activity of QKI, leading to cerebellar exon inclusion (bottom, blue), when compared to other brain tissue groups. **(G)** Scatter plot showing hexamer Z-score

51

correspondence for two non-overlapping sets of cerebellar CE exclusion events: (y-axis) CE exclusion events which came from AS modules containing changing intron retention (IR) event(s) versus non-changing and (x-axis) CE exclusion events from AS modules without IR event(s) detected. Motifs of interest are highlighted according to colors in the inset. **(H)** RNAmaps of motifs of interest for given sets of cerebellar exclusion cassette exon event sets. Top shows G-rich hexamer motif occurrence (five of six positions are G and contains GGGG) while bottom shows QKI hexamers (ACUAAY). Frequencies were determined over sliding windows of 20 nucleotides with smoothing using a running mean of five. Lines indicate CE exon set according to the legend: red, all cerebellar exclusion CEs; orange dashed, subset of exclusion CEs which also contained a changing IR event; fuchsia dashed, subset of exclusion CEs with no IR event with the AS module; gray, all CEs which were not changing between comparisons.

Supplementary Materials

Supplementary Note: procedure for bootstrapped readrates from per-position reads

MAJIQ's bootstrapping procedure can be defined as follows. Without loss of generality, consider a single junction. For each RNA-seq read aligned with a split for this junction, we define the read's position relative to the junction (or vice-versa) with a position i and count the number of reads associated with each position, which we call S_i.

These raw readrates include biases that we would like to correct for; in particular, we define an explicit procedure for removing stacks by comparing the number of reads at each position against a Poisson model using the observed readrates at all other positions, which results in a set of stack-corrected nonzero readrates R_i for $i \in \{1, \dots, P\}$, where P is the number of nonzero positions after stack removal. These are the observed units for bootstrapping, so to emphasize:

$$R_i \equiv \# \text{ of RNA-seq reads associated with } i\text{-th position,} \qquad \text{(observed readrates)}$$

$$i \in \{1, \dots, P\}. \qquad \text{(nonzero positions after stack removal)}$$

Other methods typically sum directly over positions R_i (really S_i since they generally also ignore read stacks) to produce a total junction readrate for use in quantification:

$$R \equiv \sum_{i=1}^{P} R_i. \qquad \text{(observed total junction readrate)}$$

Since we are unsatisfied with uncertainty/variance accounted for by directly using R, we generate samples from a bootstrap distribution over the P nonzero positions.

If we make the assumption that we are given the number of nonzero positions P and that the underlying readrate for each of these positions is independent and identically distributed with finite mean $\mathbb{E}[R_i] = \mu$ and variance $\mathbb{V}[R_i] = \sigma^2$, we can derive the mean and variance of our observed total readrate:

$$
\begin{aligned}
\mathbb{E}\left[R\right] &= \mathbb{E}\left[\sum_{i=1}^{P} R_i\right] \\
&= \sum_{i=1}^{P} \mathbb{E}\left[R_i\right] \\
&= P\mu, \qquad\qquad\qquad\text{(observed total readrate mean)} \\
\mathbb{V}\left[R\right] &= P\sigma^2. \qquad\qquad\qquad\text{(observed total readrate variance)}
\end{aligned}
$$

If we were able to take two samples for the observed total readrate (i.e. R and R'), their difference has mean 0 and variance $2P\sigma^2$.

We define our bootstrapping procedure over observed nonzero reads $R_1, \ldots, R_P$ to generate bootstrapped total reads $\hat{R}, \hat{R}', \ldots$ such that the variance of the difference between bootstrap samples would be equivalent to that of the difference between two samples from the true distribution (i.e. $2P\sigma^2$). In order to do this, we take $P-1$ samples from $\{R_1, \ldots, R_P\}$ with replacement and scale their sum by $P/(P-1)$.

It is straightforward to see that the bootstrapped total readrate has the same mean as the observed total readrate. In order to prove that the variance of the difference between two sample matches, we note that the covariance $\text{Cov}\left(R_{Z_k}, R_{Z_{k'}}\right)$ between any two draws from the observed per-position readrates with $Z_i \sim \text{Uniform}(P)$ is:

$$
\begin{aligned}
\text{Cov}\left(R_{Z_k}, R_{Z_{k'}}\right) &= \mathbb{E}\left[\left(R_{Z_k} - \mu\right)\left(R_{Z_{k'}} - \mu\right)\right] \\
&= \mathbb{E}\left[R_{Z_k} R_{Z_{k'}}\right] - \mu^2.
\end{aligned}
$$

. We note that $\mathbb{E}\left[R_i R_j\right] = \sigma^2 \delta_{ij} + \mu^2$ (where δ_{ij} is the Kroencker delta). When $k = k'$, it follows that $\mathbb{E}\left[R_{Z_k} R_{Z_{k'}}\right] = \sigma^2 + \mu^2$. Otherwise, the law of total expectation gives:

$$
\begin{aligned}
\mathbb{E}\left[R_{Z_k} R_{Z_{k'}}\right] &= \mathbb{E}\left[\mathbb{E}\left[R_{Z_k} R_{Z_{k'}} \big| Z_k, Z_{k'}\right]\right] \qquad\qquad \text{(given } k \neq k') \\
&= \frac{1}{P^2} \sum_{i=1}^{P} \sum_{j=1}^{P} \sigma^2 \delta_{ij} + \mu^2 \\
&= \mu^2 + \frac{1}{P}\sigma^2.
\end{aligned}
$$

54

Combining the two cases, we have:

$$\mathbb{E}\left[R_{Z_k} R_{Z_{k'}}\right] = \delta_{kk'}\left(\mu^2 + \sigma^2\right) + \left(1 - \delta_{kk'}\right)\left(\mu^2 + \frac{1}{P}\sigma^2\right)$$

$$= \mu^2 + \frac{1}{P}\sigma^2 + \delta_{kk'}\frac{P-1}{P}\sigma^2. \qquad \text{(second moment of sampled readrates)}$$

Therefore,

$$\mathrm{Cov}\left(R_{Z_k}, R_{Z_{k'}}\right) = \frac{1}{P}\sigma^2 + \delta_{kk'}\frac{P-1}{P}\sigma^2. \qquad \text{(covariance of sampled readrates)}$$

Thus, the variance of the bootstrapped total readrate is

$$\mathbb{V}\left[\hat{R}\right] = \frac{P^2}{(P-1)^2}\mathbb{V}\left[\sum_{k=1}^{P-1} R_{Z_k}\right]$$

$$= \frac{P^2}{(P-1)^2}\sum_{k=1}^{P-1}\sum_{k'=1}^{P-1}\mathrm{Cov}\left(R_{Z_k}, R_{Z_{k'}}\right)$$

$$= \frac{P^2}{(P-1)^2}\sum_{k=1}^{P-1}\sum_{k'=1}^{P-1}\frac{1}{P}\sigma^2 + \delta_{kk'}\frac{P-1}{P}\sigma^2$$

$$= 2P\sigma^2. \qquad \text{(true bootstrap readrate variance)}$$

But we want the variance of the difference between two samples from the bootstrap procedure. So, we calculate the covariance between two distinct samples:

$$\mathrm{Cov}\left(\hat{R}, \hat{R}'\right) = \frac{P^2}{(P-1)^2}\sum_{k=1}^{P-1}\sum_{k'=1}^{P-1}\frac{1}{P}\sigma^2$$

$$= P\sigma^2. \qquad \text{(covariance between samples of }\hat{\mathcal{P}}\text{)}$$

Therefore, we find that:

$$\mathbb{V}\left[\hat{R} - \hat{R}'\right] = 2\mathbb{V}\left[\hat{R}\right] - 2\,\mathrm{Cov}\left(\hat{R}, \hat{R}'\right)$$

$$= 4P\sigma^2 - 2P\sigma^2$$

$$= 2P\sigma^2. \qquad \text{(bootstrap total readrate variance as difference)}$$

55

1256   In practice, the observed nonzero positions can lead to a bootstrap distribution with variance

1257   less than its mean (underdispersed). We generally expect readrates to follow a Poisson or negative

1258   binomial (overdispersed) distribution, so in these cases, we fall back to parametric bootstrapping with

1259   a Poisson distribution with mean $P\mu$. Otherwise, we use the nonparametric bootstrap sampling

1260   procedure as described above.

Supplementary Figures
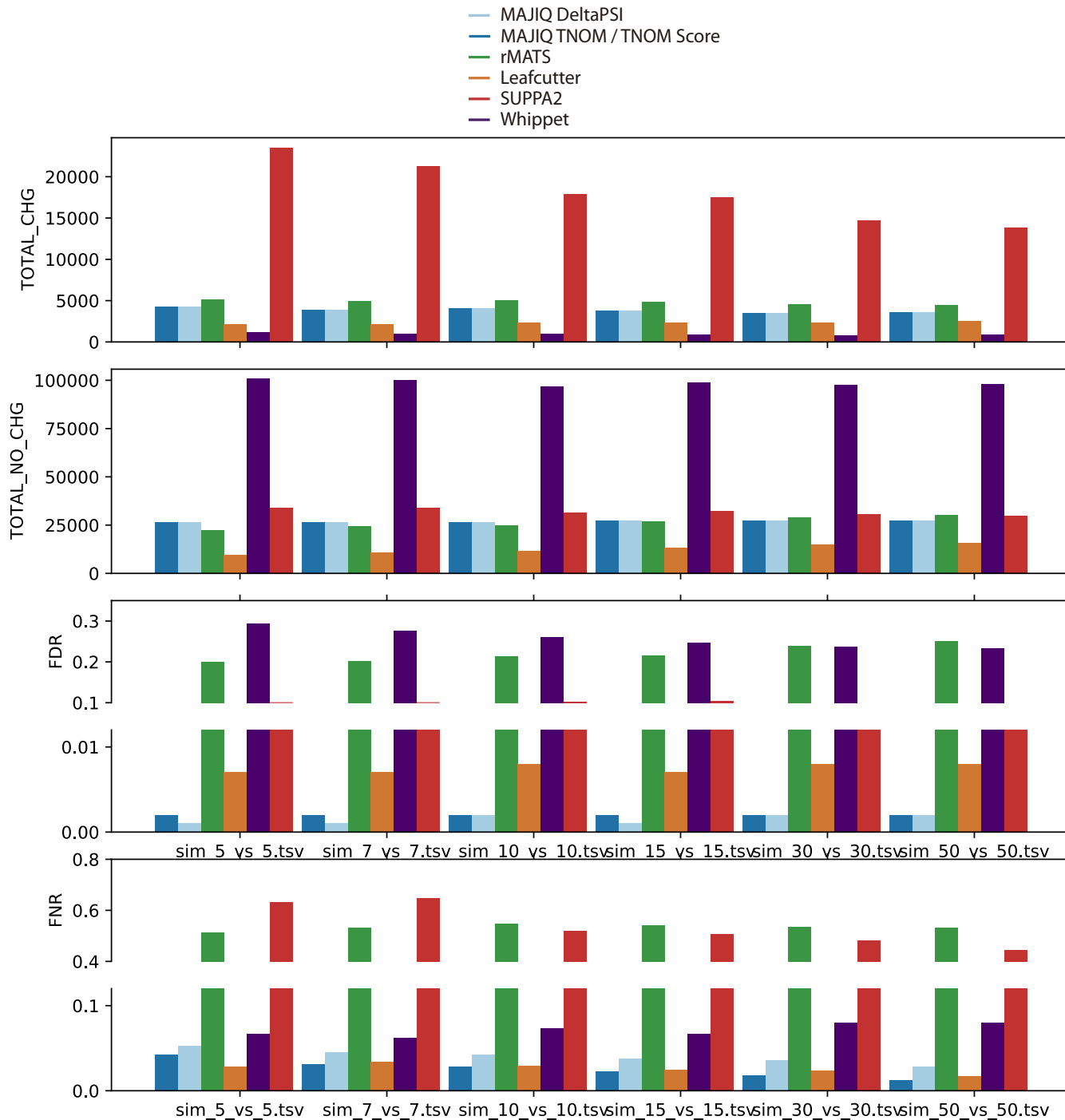
Fig. S1: Performance evaluation using simulated data at event level. This figure is equivalent to Fig. 2B in the main text but displays the results when using each method's unique event definition

1264 rather than aggregated at the gene level. For methods that quantify local AS events such as rMATS

1265 and MAJIQ, the number of changing events is approximately double that of changing genes (2,337 vs

1266 4,267 for MAJIQ HET), while for LeafCutter, which uses a coarser definition for events based on

1267 intron clusters, the number of changing events and genes is similar (1,739 vs 2,169).
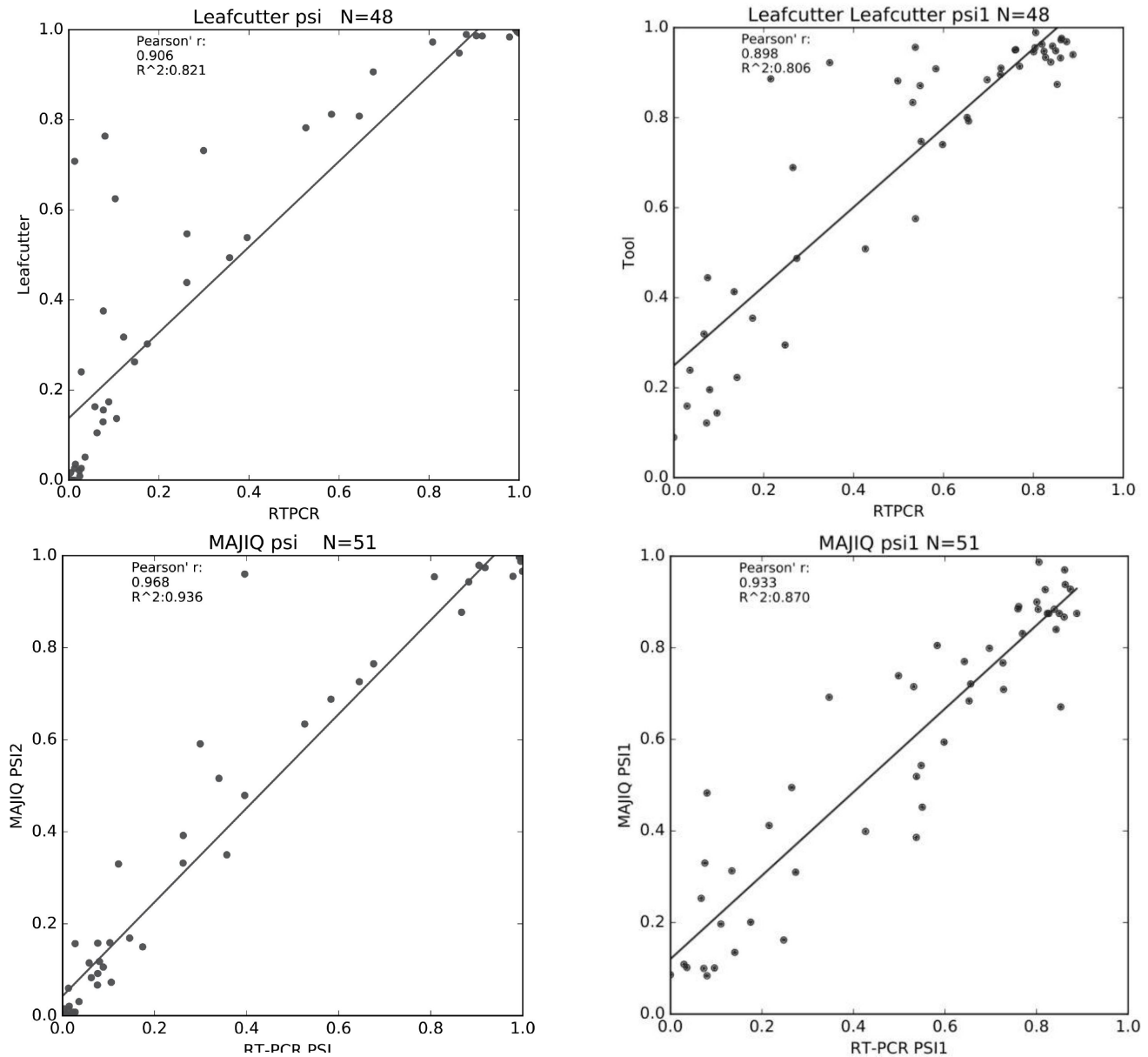
1268 **Fig. S2: Correlation between LeafCutter and MAJIQ quantifications and RT-PCR.** Correlation

1269 between RNA-seq based quantifications by LeafCutter (top row) or MAJIQ (bottom row) and

1270 RT-PCR in liver (left) and cerebellum (right). RT-PCR quantifications are from [3] using RNA used

1271 by [36] to produce the RNA-seq samples. Note that all splicing events shown here were selected by [3]

58

1272 to be binary, annotated, and changing between the two tissues to allow direct comparison to rMATS.

1273 The usage of simple binary events allowed us to calibrate LeafCutter's intron cluster quantifications to

1274 PSI, which is not possible in the general case.
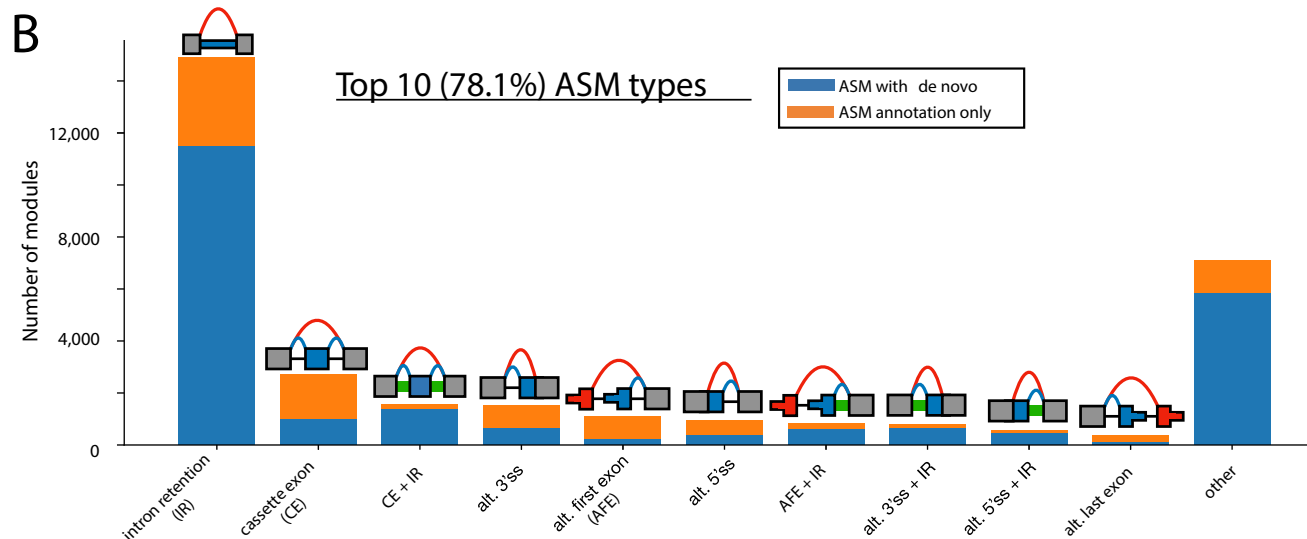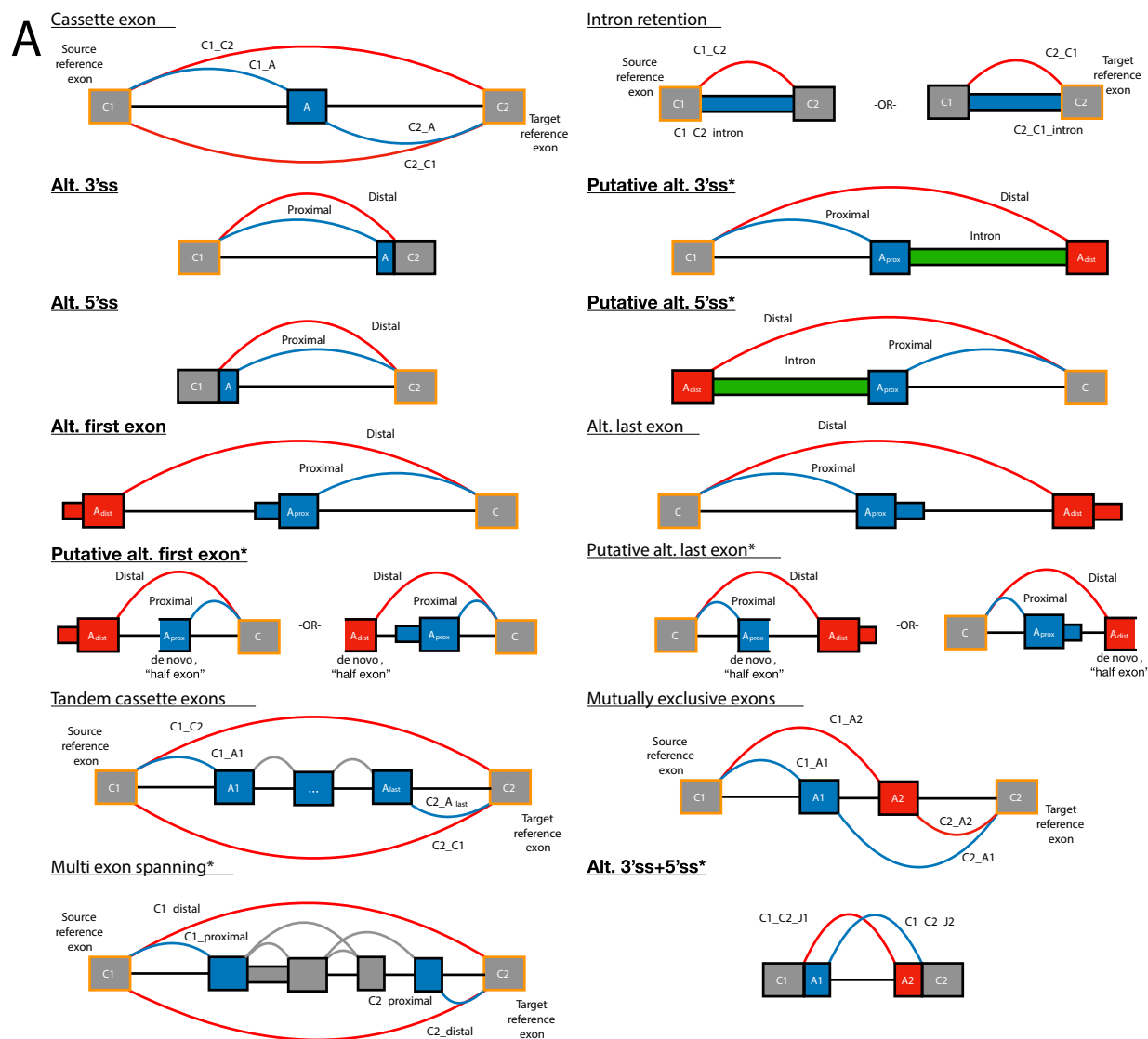
Fig. S3: VOILA Modulizer AS event types. (A) Diagrams outlining the structure of alternative splicing event (AS event) types used in the VOILA Modulizer. Exons and junctions are labeled in a way consistent with the tab separated value text file outputs of the Modulizer. Grey exons outlined in orange indicate the reference exon(s) from local splicing variations (LSVs, source and/or target) used to create the splicing events. Blue junctions, introns, exons, and exonic regions correspond to inclusion products while red corresponds to exclusion products. Grey junctions in tandem cassette exons and multi-exon skipping correspond to other junctions present in the splicegraph after simplification, but are not directly considered or output by the Moduilzer in terms of quantifications. Green introns in putative 5′ and 3′ss events indicate a retained intron that was quantified to high inclusion, but had the corresponding splice junction removed during simplification due to low PSI. This suggests A_{prox}, the intron, and A_{dist} behave as a single exon unit with the red (intron distal) and blue (intron proximal) splice junctions acting as alternative splice sites. Asterisks indicate non-classical AS event types. **(B)** Stacked barchart showing the AS event makeup of the top 10 alternative splicing modules (ASMs) across the 13 GTEx brain tissue groups from the VOILA Modulizer after applying a 5% PSI simplification threshold (e.g. junctions with a group median of less than 5% in all groups are removed). Modules named with a plus sign (e.g. CE + IR) correspond to AS modules made up of more than one AS event type (e.g. CE + IR modules were made up of both cassette exon and intron retention events). Blue bar regions indicate AS modules that had one or more de novo or unannotated junctions, after simplification, while orange regions indicate AS modules consisting of solely annotated junctions and/or retained introns.
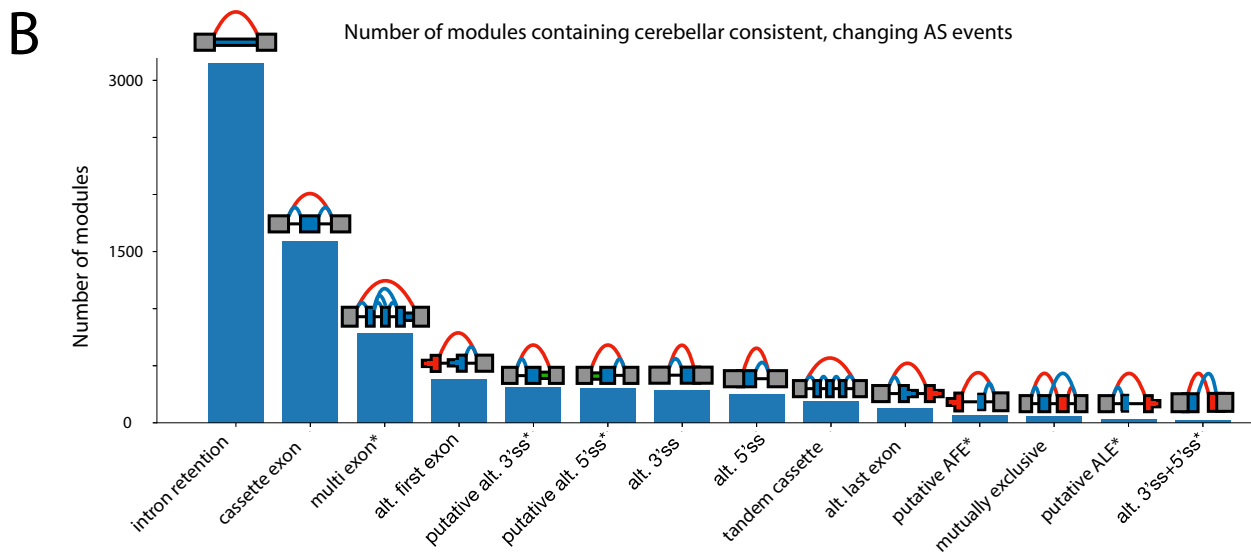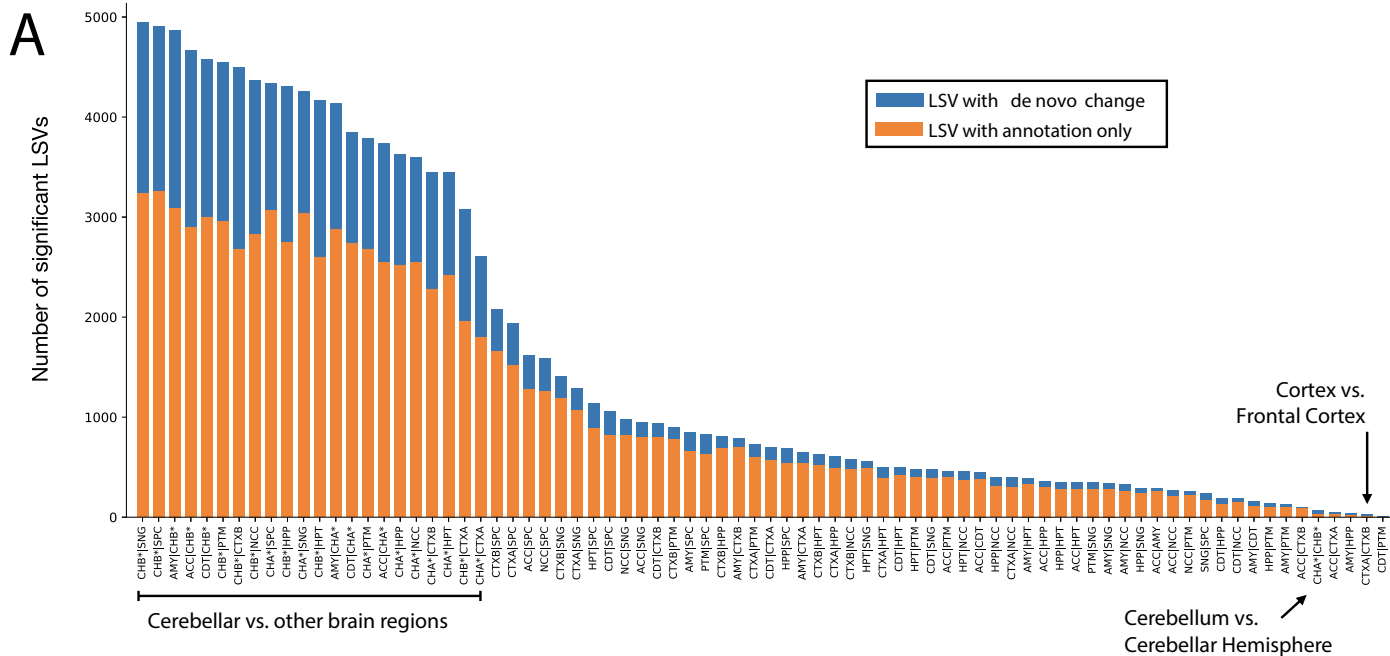
Fig. S4: Cerebellar vs other brain tissues with LSVs and splicing modules. (A) Barchart showing the number of significant LSVs from 78 pairwise MAJIQ HET comparisons between the 13 GTEx brain tissue groups. Significant LSVs were those containing at least one junction or intron with an absolute difference in group median expected PSI values of 20% between two tissue groups and all four HET statistics (Wilcoxon, InfoScore, TNOM, and t-test) with $p < 0.05$. Comparisons that include a cerebellar tissue (Cerebellum, CHA; or Cerebellar Hemisphere, CHB) are highlighted. Blue indicates LSVs containing an unannotated, de novo junction/intron that was changing and orange indicates LSVs with only annotated junctions/introns. **(B)** Barchart showing the number of modules containing at least one of the 14 alternative splicing event types found to be significantly changing in both cerebellar tissues versus one or more other brain subregion tissues in a consistent way (see Methods). Event types are outlined in Figure S3A. Non-classical event types are marked with an asterisk.
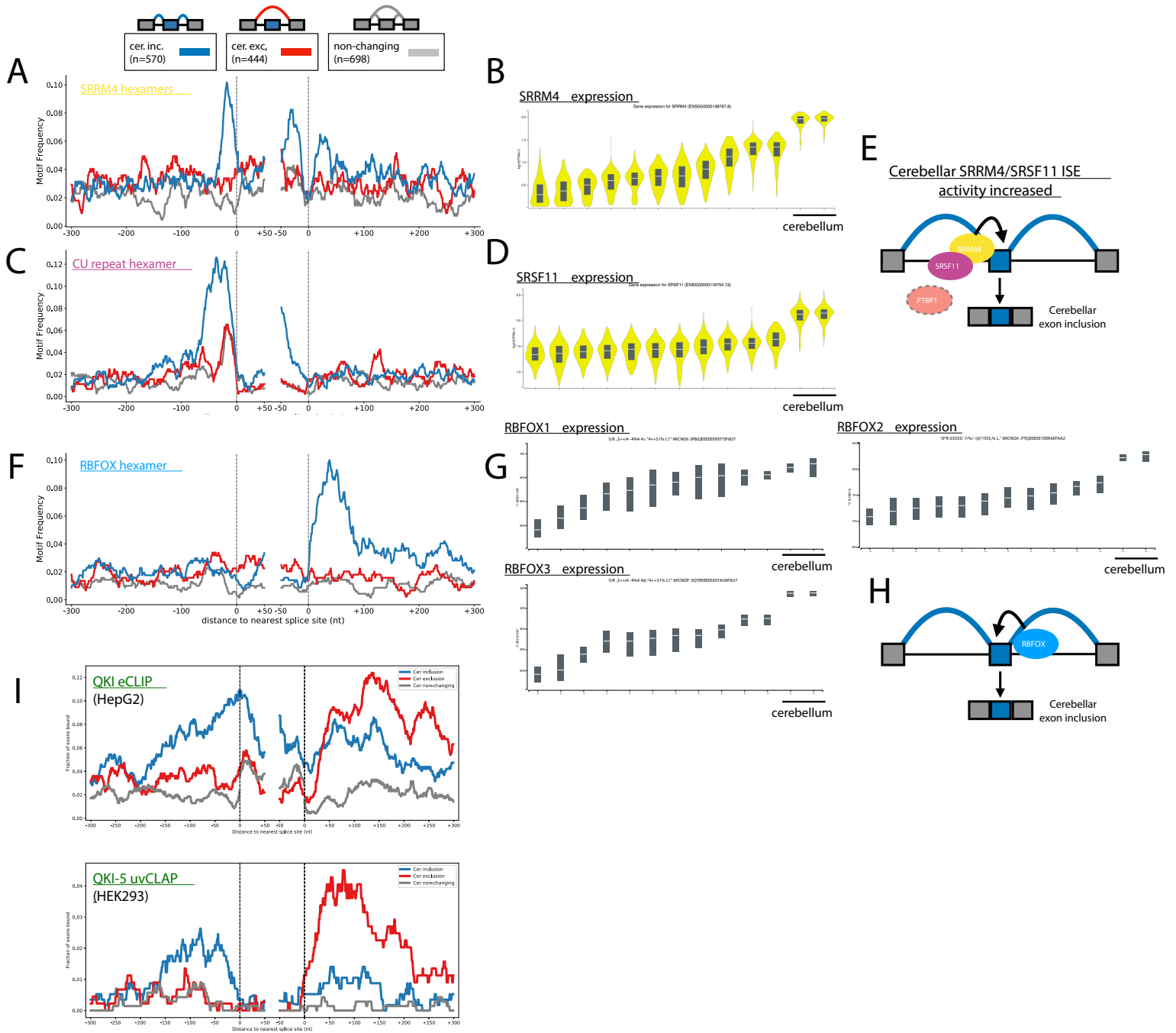
Fig. S5: motif enrichment and RBP expression for changing and non-changing cassette exons. (A) RNAmaps showing the frequency of the top UGC containing SRRS6/nSR100 hexamer motifs, as determined by iCLIP (UGCUGC, CUGCUG, GCUGCC, GCUGCU[21]), around cerebellar inclusion (blue), exclusion (red), or non-changing (gray) CEs. Frequency was determined by searching for motif occurrence over sliding windows of 20 nucleotides with smoothing using a running mean of 5 nucleotides. **(B)** *SRRS6* bulk tissue gene expression ($\log_{10}(1 + \text{TPM})$ for ENSG00000139767.8) sorted by median brain tissue expression. Chart generated using GTExportal.org. **(C)** RNAmaps showing the frequency of the top CU-repeat hexamers that bind SRSF11, as determined by iCLIP (UCUCUC and CUCUCU[20]), around cerebellar inclusion (blue), exclusion (red), or non-changing (gray) CEs. Frequency was determined by searching for motif occurrence over sliding windows of 20 nucleotides with smoothing using a running mean of 5 nucleotides. **(D)** *SRSF11* bulk tissue gene expression ($\log_{10}(1 + \text{TPM})$ for ENSG00000116754.13) sorted by median brain tissue expression. Chart generated using GTExportal.org. **(E)** Model for SRRS6/SRSF11 promotion of exon inclusion in cerebellar tissues. Increased expression of SRRS6 and SRSF11 increases intronic splicing enhancer (ISE) activity by increased binding to CU- and UGC- rich regions just upstream of cerebellar included exons. Decreased PTB expression, which also binds CU repeat elements[37], may also contribute to increased SRSF11 activity. Model is based on previous work showing cooperative binding and splicing enhancement of neuronal microexons by SRSF11 and SRRS6[20]. **(F)** RNAmap showing the frequency of the RBFOX hexamer, UGCAUG, around cerebellar inclusion (blue), exclusion (red), or non-changing (gray) CEs. Frequency was determined by searching for motif occurrence over sliding windows of 20 nucleotides with smoothing using a running mean of 5 nucleotides. **(G)** RBFOX family bulk tissue gene expression ($\log_{10}(1 + \text{TPM})$ for *RBFOX1*: ENSG00000078328.19, *RBFOX2*: ENSG00000100320.22, and *RBFOX3*: ENSG00000167281.18) sorted by median brain tissue expression. Chart generated using GTExportal.org. **(H)** Model for position dependent RBFOX regulation in GTEx brain tissues. Increased expression of RBFOX family members in cerebellar tissues leads to increased intronic splicing enhancer activity (ISE) through increased RBFOX binding downstream of exons, resulting in cerebellar exon inclusion (blue), when compared to other brain tissue groups. **(I)** RNAmaps showing the frequency of QKI CLIP peak occurrence, indicating *in vivo* binding of QKI around cerebellar inclusion (blue), exclusion (red), or non-changing (gray) CEs. Top plot shows the frequency of QKI eCLIP peaks in HepG2 cells [34] while bottom shows uvCLAP peak frequencies for the predominantly nuclear

1336   isoform of QKI (QK-5) in HEK293 cells that is thought to regulate splicing[35].
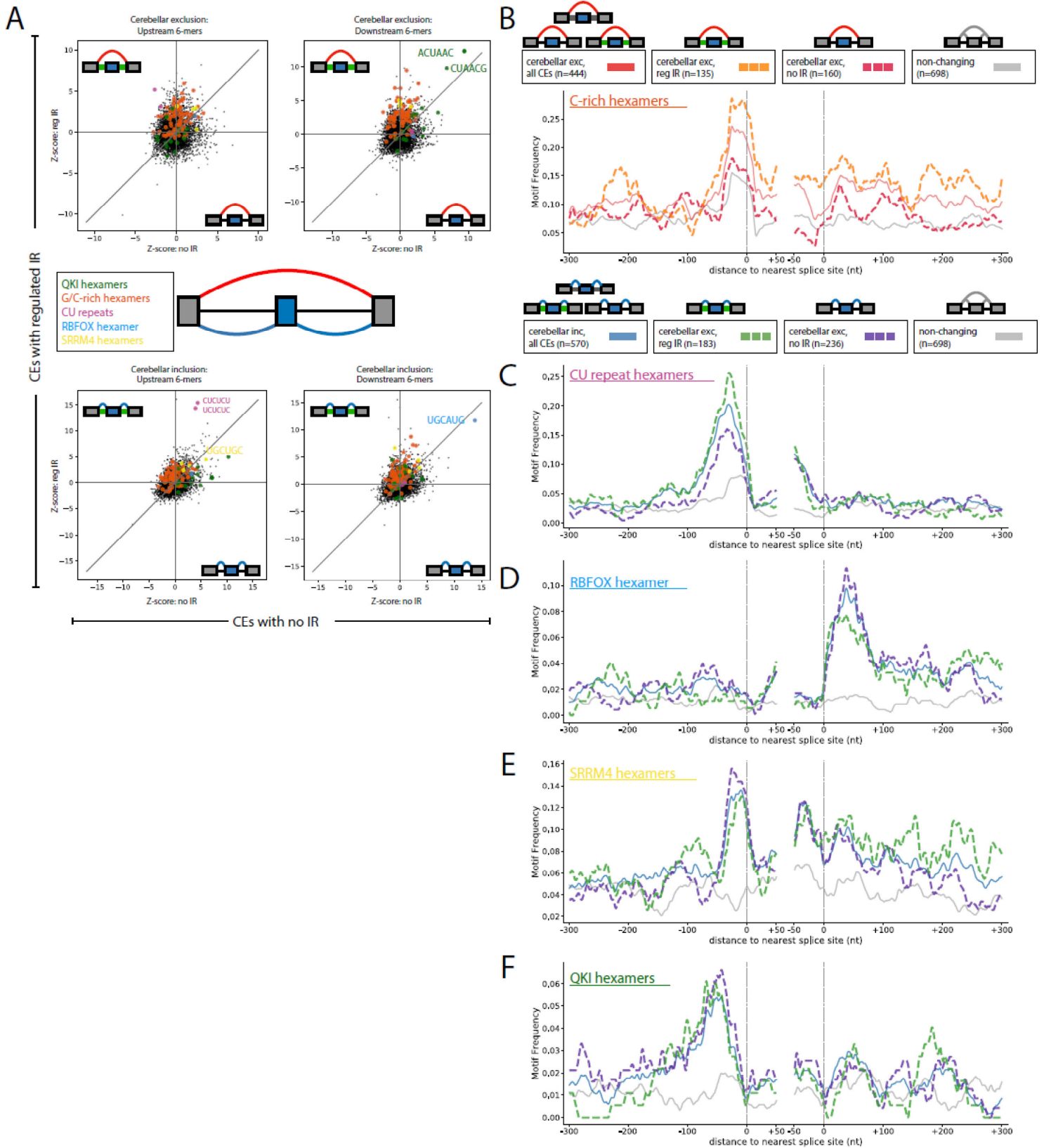
Fig. S6: Cerebellar cassette exons with and without intron retention. (A) Scatter plots showing hexamer Z-score correspondence between non-overlapping sets of cerebellar cassette exon (CE) sets. Each y-axis shows Z-scores from CE events which came from AS modules containing changing intron retention (IR) event(s) versus non-changing. Each x-axis shows Z-scores from CE events coming from AS modules without IR event(s) detected. Motifs of interest are highlighted according to colors in the inset. Top plots show enrichment around cerebellar exclusion event sets while bottom plots show enrichment around cerebellar inclusion event sets. Left plots show Z-scores derived from intronic regions 300 nucleotides upstream of the 3'ss while right plots show Z-scores derived from intronic regions 300 nucleotides downstream of the 5'ss of the cassette exon. All hexamer Z-scores for various CE sets are listed in Table S1. **(B)** RNAmaps for C-rich hexamer motif for given sets of cerebellar exclusion cassette exon event sets. Lines indicate CE set according to the legend: red, all cerebellar exclusion CEs; orange dashed, subset of exclusion CEs which also contained a changing IR event; fuchsia dashed, subset of exclusion CEs with no IR event with the AS module; gray, all CEs which were not changing between comparisons. Frequency of C-rich hexamers (five of six positions are C and contain CCCC) was determined by searching for motif occurrences over sliding windows of 20 nucleotides with smoothing using a running mean of 5 nucleotides. **(C)** RNAmaps for CU-repeat hexamer motifs for given sets of cerebellar inclusion cassette exon event sets. Lines indicate CE set according to the legend: blue, all cerebellar inclusion CEs; green dashed, subset of inclusion CEs which also contained a changing IR event; purple dashed, subset of inclusion CEs with no IR event with the AS module; gray, all CEs which were not changing between comparisons. Frequency of CU-repeat hexamers (CUCUCU, UCUCUC) was determined by searching for motif occurrences over sliding windows of 20 nucleotides with smoothing using a running mean of 5 nucleotides. **(D)** Same as in (C), but shown for RBFOX hexamer (UGCAUG). **(E)** Same as in (C), but shown for SRRS6/nSR100 iCLIP hexamers (UGCUGC, CUGCUG, GCUGCC, GCUGCU [21]). **(F)** Same as in (C), but shown for QKI hexamers (ACUAAY).

68
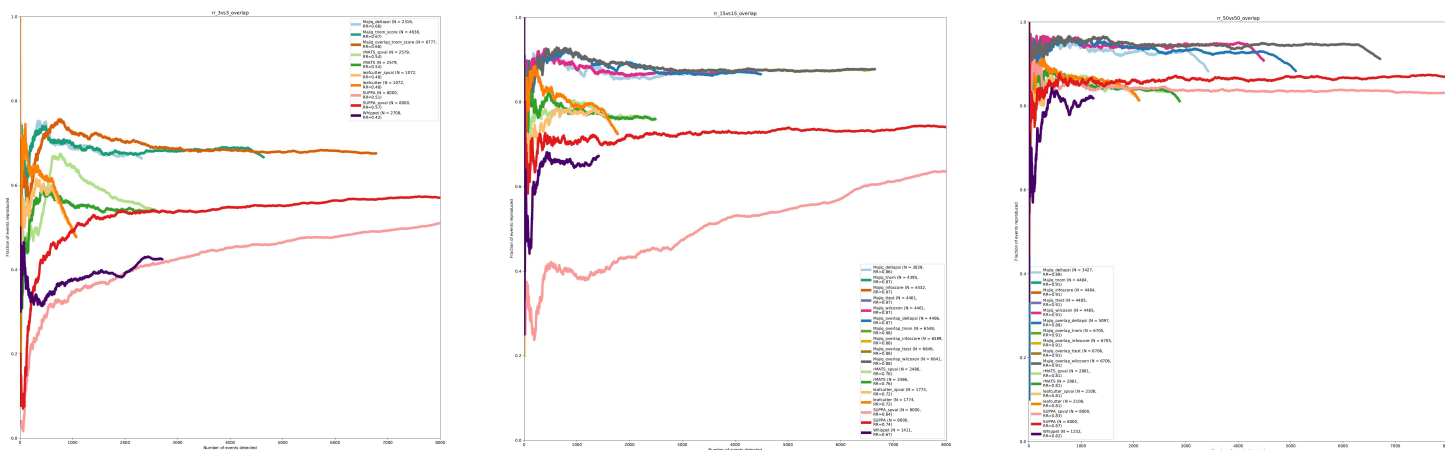
Fig. S7: Reproducibility ratio plots without filtering MAJIQ overlapping LSVs. This figure is equivalent to Fig. 2D for reproducibility ratio plots in the main text but demonstrates the effect of not filtering MAJIQ's list of overlapping LSVs. This filtering, used in Fig. 2D, is done to make the number of LSV events comparable to the number of classical splicing events reported by rMATS (see Methods). Note that removing the filtering step increases the number of reported differentially spliced LSVs by approximately 50% but retains similar reproducibility ratio curves.

Supplementary Tables

The supplementary tables are provided as an Excel workbook with sheets prefixed by "Table" and "README". The "Table" sheets specify the tables themselves, and for each table, there is a corresponding "README" sheet which describes the format/columns of the table.

Table S1: Hexamer Z scores for cerebellar CE sets Z-scores calculated for hexamer occurence within 300 nucleotides upstream or downstream of cerebellar cassette exon sets versus a set of stringent non-changing cassette exons.

Table S2: Tool parameters for performance evaluations The tools, versions, and additional parameters used for the performance evaluations.