

# Supplemental Materials

## betAS: intuitive analysis and visualisation of differential alternative splicing using beta distributions

Mariana Ascensão-Ferreira<sup>1,†</sup>, Rita Martins-Silva<sup>1,†</sup>, Nuno Saraiva-Agostinho<sup>2</sup> and Nuno L. Barbosa-Morais<sup>1,\*</sup>

<sup>1</sup> Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa, Av. Professor Egas Moniz, 1649-028 Lisboa, Portugal

<sup>2</sup> European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, CB10 1SD, Cambridge, UK

<sup>†</sup> These authors contributed equally to this work.

\* To whom correspondence should be addressed. Tel: +351 217 999 411; Email:

[nmorais@medicina.ulisboa.pt](mailto:nmorais@medicina.ulisboa.pt)

<b>SUPPLEMENTAL METHODS .....</b>	<b>2</b>
<b>The beta and the binomial distributions .....</b>	<b>2</b>
<b>Publicly available RNA-seq datasets used to illustrate the betAS differential AS analysis pipeline.....</b>	<b>2</b>
<b>betAS pipeline to model inclusion level quantification and differential AS.....</b>	<b>2</b>
<b>Pipeline to simulate empirically-inspired RNA-seq junction read counts.....</b>	<b>6</b>
<b>Visual comparison of approaches for differential splicing analyses between two conditions ....</b>	<b>9</b>
<b>betAS applied to comparisons between multiple sequential groups .....</b>	<b>11</b>
<b>REFERENCES.....</b>	<b>13</b>

## SUPPLEMENTAL METHODS

### The beta and the binomial distributions

If we consider the set of inc + exc junction reads supporting a particular alternative splicing (AS) event, each single read can be interpreted as the result of a sampling experiment with two possible outcomes: either to support inclusion (considered as “success”, with a given probability) or exclusion (a “failure”). A problem in which there is either success or failure, with a given probability of success, is a binomial problem. When the parameter of a binomial distribution (i.e., the probability of success) is not known, a beta distribution can be used as a general distribution of possible hypotheses for the probability value we want to estimate, as beta distributions are conjugate priors to the binomial likelihood (1). Thus, the *number of successes* inc can be modelled as a binomial distribution with the probability of success PSI as its parameter. Such PSI can be estimated using the beta distribution (Fig. 1A), used to model proportions. The shape of the beta distribution is determined by two parameters,  $\alpha$  and  $\beta$ , with its mean given by the ratio between  $\alpha$  and  $\alpha + \beta$ , which is directly comparable to PSI ratios, if  $\alpha \sim \text{inc}$  and  $\beta \sim \text{exc}$ .

$$\text{mean of the beta distribution with parameters } \alpha \text{ and } \beta = \frac{\alpha}{\alpha + \beta}$$

For a given mean, the higher  $\alpha$  and  $\beta$  are, the narrower the distribution is (Fig. 1, right). Thus, shape parameters  $\alpha$  and  $\beta$  can conveniently incorporate information on the number of reads used in the quantification.

### Publicly available RNA-seq datasets used to illustrate the betAS differential AS analysis pipeline

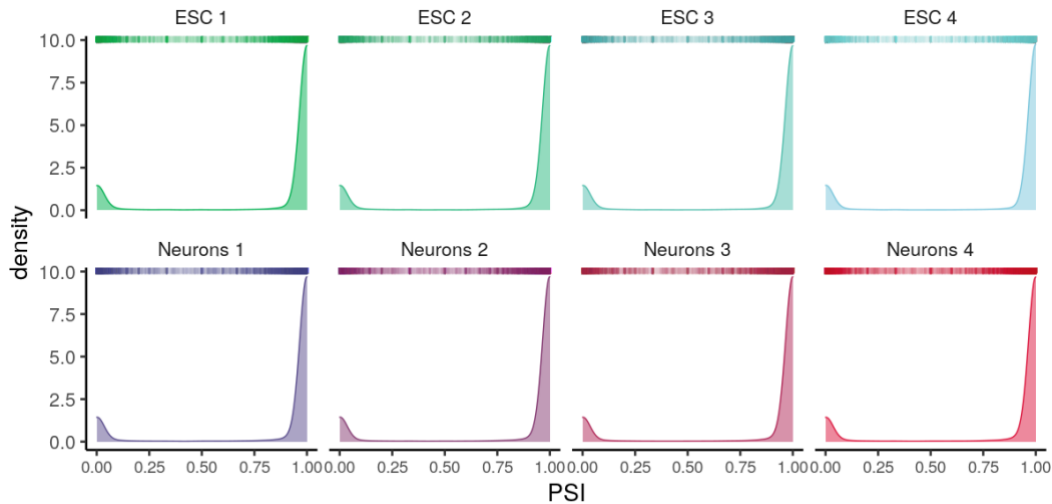
We illustrate the betAS pipeline for the two-group comparison by analysing mouse embryonic stem cells (ESC) and differentiated cortical glutamatergic neurons (Neurons), making use of public RNA sequencing data (SRA project [PRJNA185305](#)) that profile transcriptomic changes during mouse neuronal differentiation (2). For the multiple group comparisons, we made use of the human subset of an RNA-seq time-series experiment on the transcriptomic basis of the development of seven major organs (ArrayExpress accession code [E-MTAB-6814](#)) (3).

### betAS pipeline to model inclusion level quantification and differential AS

betAS requires an association between *samples* and *groups* (i.e., biological conditions), reflecting the experimental design. The pipeline starts by taking the inclusion (inc) and exclusion (exc) read counts associated with each AS event for each *sample* and uses function *rbeta* from the *stats* R package to emit a set of  $N_{\text{ind}}$  values from a beta distribution with shape parameters  $\alpha = \text{inc}$  and  $\beta = \text{exc}$  if both inc

and `exc` are non-zero. `rbeta(n = Nind, shape1 =  $\alpha$ , shape2 =  $\beta$ )` exemplifies R code to generate `n` random values from a beta distribution with shape parameters **shape1** and **shape2**.

Given that, in most biological samples, the majority of annotated alternative sequences are actually either nearly always included ( $\text{PSI} \approx 1$ ) or nearly always excluded ( $\text{PSI} \approx 0$ ), empirical PSI distributions are typically zero-one-inflated:



**Figure SM1: Empirical PSI distributions are zero-one-inflated:** most observed PSIs are close to 0 or 1. Density and rugplots (top) for the distribution of PSI values per sample for mouse embryonic stem cells (ESC) and differentiated cortical glutamatergic neurons (Neurons) that are part of the public RNA sequencing dataset PRJNA185305 (2).

While `inc` or `exc` can be equal to zero, values obtained with `rbeta` from a beta distribution with  $\alpha = 0$  or  $\beta = 0$  are single-valued as 0 or 1, respectively, and are inappropriate to model inclusion levels. To account for this, different strategies were studied to implement an “increment” factor to be added to values of `inc` or `exc` equal to zero, so that the described beta distribution-based approach is still appropriate to model inclusion levels, i.e., such that the dispersion of the emitted values decreases as coverage increases but their average values do not deviate much from 0 or 1.

### **incr = 1**

Since the minimum unit of evidence for a sequence is one read, the simplest way to increment `inc` or `exc` in order to avoid zeros would be to make  $\alpha = 1$  (when the measured  $\text{PSI} = 0$ ) or  $\beta = 1$  (when the measured  $\text{PSI} = 1$ ). Incrementing in one the measured number of reads has an impact on the PSI deviation that is inversely proportional to the read coverage and observed to be too high when the coverage is of only a few reads (Fig. SM2 below).

### **incr = cov/99**

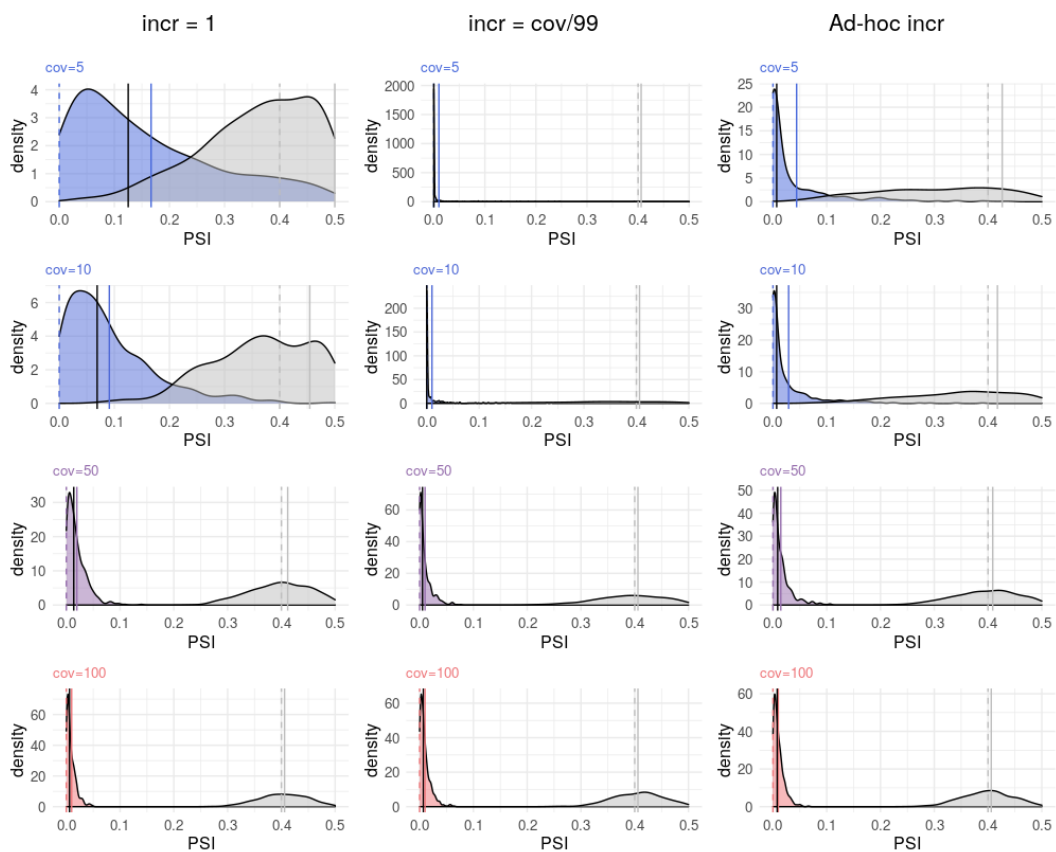
As an alternative, one could consider a coverage-dependent dynamic increment such that a  $\text{PSI} = 0$  will be corrected to be 0.01 (an arbitrarily set value for the minimum biological PSI that is not zero):

$$\frac{\text{incr}}{\text{incr} + \text{cov}} = 0.01 \Leftrightarrow \text{incr} = \frac{\text{cov}}{99}$$

The increment would then be 1 for coverage  $\geq 99$ . However, PSI estimates supported by very low coverage get an artificially high precision (Fig. SM2 below), while PSI dispersion should decrease as coverage increases. We therefore tried instead a non-parametric approach based on the median of emitted values.

### Ad-hoc empirically determined incr

For a measured PSI of 0 or 1, we established the increment to inc or exc to be lower than one, as one read is the minimum unit of evidence for a sequence. Moreover, the increment should be such that the median of the beta distribution is arbitrarily limited to  $\leq 0.01$  or  $\geq 0.99$ , respectively. For each coverage/increment pair, with coverage values being all integers between 1 and 100 reads and increments incr equally spaced by 0.01 between 0 and 1, 100 beta distributions of 500 values were generated, with parameters  $\alpha = \text{incr}$  and  $\beta = \text{coverage}$ , to find, for each coverage, the maximum increment factor such that no more than 5 simulated distributions had a median higher than 0.01.

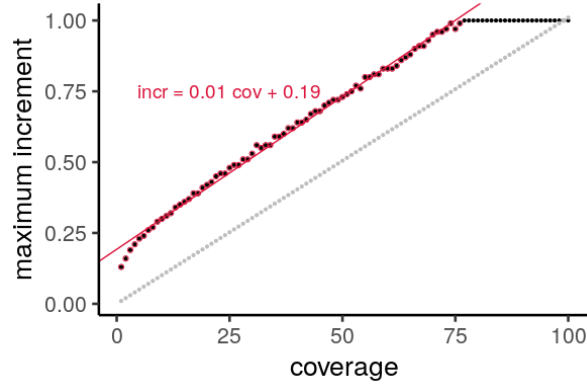


**Figure SM2: Comparison of the impacts on PSI between increment approaches**

In colour: beta distributions of PSI = 0 for each coverage in read counts (cov), after summing *incr* reads to the number of inclusion reads to avoid inc = 0; grey: beta distributions of PSI = 0.4 for each coverage, after summing *incr* to the number of inclusion reads; vertical dashed line: reference PSI (coloured for PSI = 0, grey for PSI = 0.4); vertical solid line: mean PSI after summing *incr* reads (coloured for PSI = 0, grey for PSI = 0.4); solid black line: median PSI after summing *incr* reads for PSI = 0.

Left: impact of summing 1 to zero-valued inc read counts. Middle: impact of summing cov/99 reads to zero-valued inc read counts. Right: impact of summing an ad-hoc empirically determined incr (see description in this document's main text) to zero-valued inc read counts.

We empirically observed that this is approximately equivalent to considering a linear relationship between *incr* and *coverage*:  $\text{incr} = 0.01 \cdot \text{cov} + 0.19$  (Fig. SM3). Moreover, we observe the expected decrease in PSI dispersion as coverage increases (Fig. SM2).



**Figure SM3: Empirical determination of the maximum increment as a function of coverage**

Scatterplot showing the maximum increment found per coverage (black) as well as the dynamic increment corresponding to the  $\text{incr} = \text{cov}/99$  approach (grey). Linear regression equation and line for the empirically obtained values with a maximum increment below 1 (red).

After this procedure, for each event, sets of  $N_{\text{ind}}$  *rbeta*-emitted values per sample are grouped per condition (ESC and Neurons in Fig. 2) with the median per group as their summary average PSI (Fig. 2C-D). The effect size of differential inclusion of the AS sequence between conditions is represented by  $\Delta\text{PSI}$ :

$$\text{PSI}_{\text{betAS}}(\text{condition}) = \text{median}(\text{rbeta } 1, \text{rbeta } 2, \dots, \text{rbeta } N)$$

(where *rbeta i* represents the set of *rbeta*-emitted values for sample *i*)

$$\Delta\text{PSI} = \text{PSI}_{\text{betAS}}(\text{ESC}) - \text{PSI}_{\text{betAS}}(\text{Neurons})$$

Analogously to the vast-tools *diff* module (4), the betAS pipeline takes the proportion of differences between the two shuffled sets of emitted values per condition that are greater than a minimum value (*minDiff*) as the significance metric for each estimated  $\Delta\text{PSI}$  (Supplemental Fig. S2). Using  $\text{minDiff} = 0$  has the same interpretation as asking what proportion of *rbeta*-emitted values for ESC are higher than those emitted for Neurons, thus reflecting the probability of differential splicing,  $P_{x=0}$  of  $\text{PSI}_{\text{betAS}}(\text{ESC})$  being greater than  $\text{PSI}_{\text{betAS}}(\text{Neurons})$ :

$$P_{x=0} = \frac{\#(\text{differences that are } > 0)}{\# \text{ all differences}}$$

with **differences** = emitted values<sub>ESC</sub> - emitted values<sub>Neurons</sub>

betAS allows users to tune the significance of estimated  $\Delta\text{PSI}$ s, by setting the minimum difference *minDiff* in order to calculate  $P_{x=\text{minDiff}}$ :

$$P_{x=\text{minDiff}} = \frac{\#(\text{differences that are } > \text{minDiff})}{\# \text{ all differences}}$$

with **differences** = emitted values<sub>A</sub> - emitted values<sub>B</sub>

This calculation is performed spanning a set of  $\Delta$ PSI values between -1 and 1 with increments of 0.002 (Supplemental Fig. S2E-G). In a comparison between two groups with different number of samples, the set of *rbeta*-emitted values from the largest group is sampled to match the dimension of the vector representing the smaller group.

Importantly, since the comparison is sided (in this case, emitted values<sub>ESC</sub> - emitted values<sub>Neurons</sub> > 0), both low and high probabilities reflect the most extreme differences. For instance, with minDiff = 0 a  $P_{x=0} = 0.8$  would reflect that there is a 0.8 probability of higher inclusion values in ESC than in Neurons (while the probability of Neurons having greater inclusion values than ESC is 0.2) but a probability of  $P_{x=0} = 0.2$  would reflect the opposite case, with the probability of inclusion levels in ESC being greater than in Neurons equal to 0.2 and that of levels in Neurons being greater than in ESC equal to 0.8. Accordingly, to simplify the interpretation of this metric, betAS transforms the values of  $P_{x=0}$  from the interval [0,1] to the interval [0.5,1]:

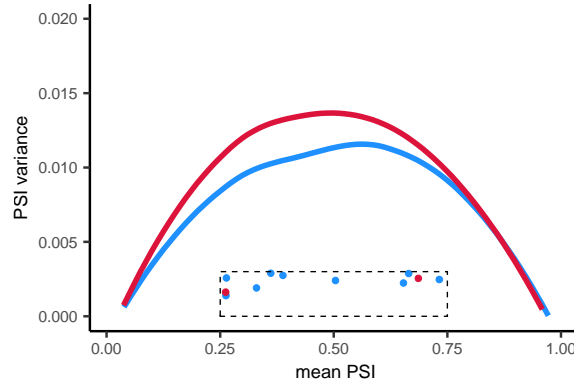
$$P_{\text{diff}} = |P_{x=0} - 0.5| + 0.5$$

This defines an intelligible significance metric that, coupled with the  $\Delta$ PSI, allows an informed selection of interesting (according to the user's criteria) differentially spliced sequences between two conditions (Fig. 2C-D).

### **Pipeline to simulate empirically-inspired RNA-seq junction read counts**

With the purpose of simulating RNA-seq junction read counts, inspired in real PSIs and coverages, on which to test the accuracy of betAS in quantifying AS differences, we started by identifying a set of GTEx (5) studied cerebellum- and muscle-specific AS events reflecting the lowest biological variability, i.e., the least observed biological noise.

By studying the relationship between the PSI variance and mean for the subset of 1143 tissue-unimodal events (Supplemental Fig. S4B-D) while considering the mean and variance relationship for a beta distribution with shape parameters  $\alpha$  and  $\beta$  (equations 1 and 2), a smaller set of representative events showing the lowest observed variance with PSI mean values ranging between 0.25 and 0.75 was considered to determine the combination of  $\alpha$  and  $\beta$  that is equivalent to the lowest biological variance.



**Figure SM4: PSI variance and mean relationship for a subset of tissue-specific unimodal events.**

PSI variance plotted against PSI mean per event across muscle (red) or cerebellum (blue) GTEx samples. Curves show the local polynomial regression fitting (“loess”) for all the 1143 events considered in muscle and cerebellum. The subset of events selected to feed the empirical estimate of the lowest observed variance are highlighted. Dashed rectangle indicates the region in which events were considered as representative of the lowest biological variance.

$$\text{PSI mean} = \frac{\alpha}{\alpha + \beta} \text{ (equation 1) and } \text{PSI variance} = \frac{\alpha \cdot \beta}{(\alpha + \beta)^2 \cdot (\alpha + \beta + 1)} \text{ (equation 2)}$$

From equations 1 and 2, it can be concluded that, for a fixed PSI value, its variance is inversely proportional to  $\alpha + \beta$ :

$$\text{PSI variance} = \frac{\text{PSI mean} \cdot (1 - \text{PSI mean})}{\alpha + \beta + 1}$$

We therefore solved the above equations in order to find  $\alpha$  and  $\beta$  for each of the set of low-variance representative events:

$$\alpha = \frac{\text{PSI mean}^2}{\text{PSI variance}} \cdot (1 - \text{PSI mean} - \text{PSI variance})$$

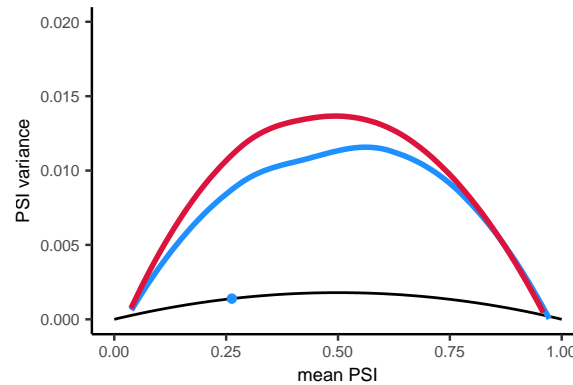
$$\beta = \frac{1 - \text{PSI mean}}{\text{PSI mean}} \cdot \alpha$$

Then, for these points, the maximum  $\alpha + \beta$ , i.e., the minimum variance was determined.

Mean	Variance	$\alpha$	$\beta$	$\alpha + \beta$
0.504	0.00240411	51.878	51.105	102.983
0.264	0.00257553	19.614	54.770	74.385
0.388	0.00274568	33.184	52.311	85.495
0.665	0.00287193	50.919	25.665	76.584
0.653	0.00223484	65.559	34.782	100.341
0.361	0.00289198	28.483	50.323	78.806
0.732	0.00247098	57.368	20.977	78.345
0.330	0.00190498	37.955	77.092	115.047
<b>0.263</b>	<b>0.00138866</b>	<b>36.395</b>	<b>102.109</b>	<b>138.504</b>

Mean	Variance	$\alpha$	$\beta$	$\alpha + \beta$
0.262	0.00162246	30.993	87.229	118.222
0.686	0.00254829	57.299	26.217	83.517

Their PSI variance to PSI mean relationship matched that of a beta distribution with shape parameters such that  $\alpha + \beta = 138.5042$ .



**Figure SM5: PSI variance and mean relationship for the low-variance events.** PSI variance plotted against PSI mean per event across muscle (red) or cerebellum (blue) GTEx samples. Curves show the local polynomial regression fitting (“loess”) for all the 1 143 events considered in muscle and cerebellum. The relationship between PSI variance and mean that is associated with the lowest variance, with  $\alpha + \beta = 138.5042$ , is marked by a black line, with the AS event associated with this maximum highlighted by the larger blue dot.

Then, for each event, beta distributions with the lowest empirical variance, i.e., such that  $\alpha + \beta = 138.5042$ , and with mean values equal to the reference PSI for that tissue/event pair ( $\text{PSI}_{\text{REF}}$ ) were used to simulate  $N$  (intended number of simulated biological replicates) PSI values.

In R:

```
simulatedPSIs <- rbeta(n = N, shape1 = a, shape2 = b),
```

with  $a + b = 138.5042$  and  $\frac{a}{a+b} = \text{PSI}_{\text{REF}}$

The Poisson distribution is used to model the probability of a given number of events occurring in a given interval and serves as a standard model for count data under the assumptions that: 1) transcripts are generated in a sample at a given transcription rate; 2) transcript fragments are detected with a given probability; 3) the number of observed reads follows a Poisson distribution with a rate (parameter) that is proportional to the product between that transcription rate and the probability of fragment detection. Along the lines of the PSI value simulation described above,  $N$  coverage values were thus simulated from a Poisson distribution with parameter  $\lambda$  equal to the reference coverage of that tissue/event pair ( $\text{cov}_{\text{REF}}$ ):

```
simulatedCoverages <- rpois(n = N, lambda = covREF)
```

The binomial distribution with parameters  $n$  and  $p$  is appropriate to model read counts because it reflects the number of events in a series of  $n$  independent Bernoulli experiments, each with the same probability of occurrence  $p$  (see section *The beta and the binomial distributions*). Then, by combining simulated PSIs and coverages, RNA-seq splice junction read counts are simulated from a binomial distribution using the *rbinom* function:

```
bin <- rbinom(n = cov, size = 1, prob = psi)
```

```
inc <- sum(bin)
```

```
exc <- cov - inc
```

## **Visual comparison of approaches for differential splicing analyses between two conditions**

We compared the betAS approach with other tools for differential splicing analysis by making use of public RNA sequencing data (SRA project [PRJNA185305](https://www.ncbi.nlm.nih.gov/sra/PRJNA185305)) on mouse embryonic stem cells and differentiated cortical glutamatergic neurons. In order to illustrate, discuss and interpret the particularities of measuring the inclusion of alternative sequences using evidence from RNA-seq data, we considered tools that fulfil all the following criteria:

1. quantify differential AS between samples using FASTQ or BAM files;
2. provide a significance metric to assist the interpretation of splicing differences;
3. use PSIs (or equivalent) to quantify the inclusion of AS events.

With these criteria, we initially selected SUPPA2 (6), rMATS (7), Whippet (8) and MAJIQ (9) to benchmark differential AS statistics provided by betAS applied over vast-tools' AS quantifications (Supplemental Fig. S7). Overall quality of raw RNA-seq data of ESC and mature 28-DIV neurons samples was assessed on the respective FASTQ files using fastqc version 0.11.5 (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc>) with no relevant quality problem or technical bias identified.

vast-tools (10–12) version 2.5.1 was used for quantification of inclusion levels of the alternative sequences in VAST-DB annotation for the mouse mm10 genome assembly. To ensure sufficient and reliable read coverage, only events with PSIs quantified for all samples (i.e., with no NA values) and a minimum mappability-corrected read coverage score of “VLOW” across all samples were considered. Only exon skipping events were considered for simplicity and easier interpretability. Quality columns from vast-tools output table containing inclusion levels (“.Q”) were used to extract inc and exc mappability-corrected junction read counts to serve as input for betAS, for each quantified PSI per event and sample. betAS analysis based on vast-tools quantifications was performed as detailed in the previous sections.

SUPPA2 was run as detailed in their GitHub tutorial

(<https://github.com/comprna/SUPPA/wiki/SUPPA2-tutorial>). First, transcript quantification was performed using salmon version 1.3.0 indexed on mouse Gencode transcript annotation GENCODE M25 (gencode.vM25.transcripts.fa), followed by quantification in the quasi-mapping-based mode with library type automatically inferred (-l A) and --gcBias flag on. SUPPA2 in-house “multipleFieldSelection” script was used (with index row k=1 and column to extract f=4) to extract TPMs from salmon quant.sf files. Moreover, as explained in the tutorial, transcript IDs in the previous set’s output file iso\_tpm.txt were formatted to match those in the GTF file that SUPPA2 uses to calculate inclusion levels.

Next, event generation was performed using GENCODE M25 annotation and --pool-genes option for all event types: skipping exons (SE), alternative 5’/3’ splice sites (SS), mutually exclusive exons (MX), retained introns (RI) and alternative first/last exons (FL) (<https://github.com/comprna/SUPPA>) and followed by PSI quantification per event using the .ioe files generated and the formatted file containing salmon-based TPMs. To study differential AS on SUPPA2’s generated events, diffSplice was applied with .psi and isoform TPM files split between ESC and Neurons with parameters -m empirical (to calculate significance), --gene-correction (to apply multiple testing correction by gene) and --save-tpm-events. Differential splicing results provided by SUPPA2, comprising 79 983 events, were filtered in R to keep each event for which the cognate transcript had TPM quantification available and the dPSI was not a missing value.

rMATS version turbo 4.1.0 was run on FASTQ files split between ESC and Neurons groups, using GENCODE M25 mouse genome annotation to build the STAR binary index and as GTF file.

Differential splicing results were filtered in R to consider only exon skipping events without PSIs with missing values.

rMATS intermediate sorted BAM files generated were indexed using samtools index such that they could be used with the GENCODE M25 transcriptome annotation GFF3 file to build local splicing variation graphs using MAJIQ builder, followed by MAJIQ psi run to calculate PSI values and MAJIQ delpsi to compare those PSIs between ESC and Neuron samples.

Whippet was used to build an index for AS events on which PSIs were quantified and compared between ESC and Neurons groups. whippet-index was run using FASTA genome and GTF gene

annotation files from the GENCODE M25 annotation with the `--suppress-low-tsl` option to consider only well-supported transcript annotations (TSL2+) when building the splice graph. Inclusion levels, their confidence, and total reads for each defined node were quantified with `whippet-quant` on the FASTQ file pair for each sample under study. `whippet-delta` was then run to compare the generated `psi.gz` files, one for each sample, grouped into ESCs and Neurons. The `psi.gz` files and the differential splicing output file were filtered in R to consider only exon skipping ("CE, or core exon") events without missing PSI values.

For the between-tools comparison of differential AS splicing of the same sequences between mouse ESC and cortical glutamatergic neurons, we focused on the exon skipping events that behave as "alternative" (i.e.,  $0 < \text{PSI} < 1$ ) in all ESC and Neuron samples. Such filtering criteria returned 738 events in `vast-tools`, 3 388 in SUPPA2 and 1 909 in rMATS (Supplemental Fig. S7). MAJIQ was not considered for the comparison between tools as the PSI quantifications for their units of local splicing variation (LSV) were not directly comparable with the exon skipping events considered for the other tools. Although Whippet also considers an annotation structure, in which PSIs are associated with the nodes of contiguous splice graphs, that is not directly comparable with those of other tools, we were able to compare `whippet-delta`'s differential splicing results for a subset of 367 exon skipping events associated with "core exon" nodes whose annotated alternative sequence coordinates' match VAST-DB's (Supplemental Fig. S7).

To match inclusion quantifications of the same event across VAST-DB's, SUPPA2's, rMATS', and Whippet's annotations, we used `vast-tools`' coordinates of the annotated events in the format `chrN:exonStart-exonEnd` (available in column "COORD" of the output table) and adapted the function `parseSuppaAnnotation()` from `psichomics` (13). All processing was performed in the R environment for statistical computing.

### **betAS applied to comparisons between multiple sequential groups**

Inspired by ANOVA, we extended the `betAS` pipeline to test differential AS between more than two groups. The coverage-dependent dispersion of values emitted from a beta distribution is a surrogate for the PSI uncertainty for each sample. Such dispersion can be compared within samples from the same group and compared with that found between groups. Thus, for each event, we consider within as the set of differences between pairs of samples that are part of the same group and between as the set of differences between pairs of samples from different groups.

Comparisons between the median absolute values of between and within are used to obtain an “F-like” statistic, which compares inter- and intra-group variabilities:

$$F_{\text{stat}} = \frac{\text{variation **between** samples}}{\text{variation **within** samples}} = \frac{\text{median}(|\text{between}|)}{\text{median}(|\text{within}|)}$$

For experimental designs with groups following a sequential relationship (e.g., time course), between distances are calculated by betAS in a sequential mode. Thus, sequential between are the set of between-group distances for any pair of consecutive groups in the sequence. Considering directional differences between groups (i.e., from each group to its predecessor), the modes of this set of differences will be associated with the prevalence of increasing (if positive) or decreasing (when negative) steps of a given magnitude, from which the monotonicity of change in inclusion levels can be empirically estimated. To accommodate this, we define, for each event, its monotonicity coefficient:

$$\text{monotonicity} = \frac{\# (\text{seq. differences that are } > 0) - \# (\text{seq. differences that are } < 0)}{\# \text{ all sequential differences}}$$

$$\text{monotonicity} = \frac{\# (\text{steps that are } > 0) - \# (\text{steps that are } < 0)}{\# \text{ all steps}}$$

$$\text{with steps} = \text{emitted values}_{\text{group } i} - \text{emitted values}_{\text{group } i-1}$$

Moreover, we extended the application of the  $P_{\text{diff}}$  described in previous sections to calculate the probability that  $|\text{between}| > |\text{within}|$ :

$$P_{x=0} = \frac{\# (\text{differences } |\text{between}| - |\text{within}| \text{ that are } > 0)}{\# \text{ minimum length of between and within differences}}$$

$$P_{\text{diff}} = |P_{x=0} - 0.5| + 0.5$$

## REFERENCES

1. Donovan,T.M. and Mickey,R.M. (2019) Bayesian Statistics for Beginners: A Step-by-Step Approach Press,O.U. (ed).
2. Hubbard,K.S., Gut,I.M., Lyman,M.E. and McNutt,P.M. (2013) Longitudinal RNA sequencing of the deep transcriptome during neurogenesis of cortical glutamatergic neurons from murine ESCs. *F1000research*, **2**, 35.
3. Cardoso-Moreira,M., Halbert,J., Valloton,D., Velten,B., Chen,C., Shao,Y., Liechti,A., Ascenção,K., Rummel,C., Ovchinnikova,S., *et al.* (2019) Gene expression across mammalian organ development. *Nature*, **571**, 505–509.
4. Han,H., Braunschweig,U., Gonatopoulos-Pournatzis,T., Weatheritt,R.J., Hirsch,C.L., Ha,K.C.H., Radovani,E., Nabeel-Shah,S., Sterne-Weiler,T., Wang,J., *et al.* (2017) Multilayered Control of Alternative Splicing Regulatory Networks by Transcription Factors. *Mol Cell*, **65**, 539-553.e7.
5. Lonsdale,J., Thomas,J., Salvatore,M., Phillips,R., Lo,E., Shad,S., Hasz,R., Walters,G., Garcia,F., Young,N., *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat Genet*, **45**, 580–585.
6. Trincado,J.L., Entizne,J.C., Hysenaj,G., Singh,B., Skalic,M., Elliott,D.J. and Eyraes,E. (2018) SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol*, **19**, 40.
7. Shen,S., Park,J.W., Lu,Z., Lin,L., Henry,M.D., Wu,Y.N., Zhou,Q. and Xing,Y. (2014) rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc National Acad Sci*, **111**, E5593–E5601.
8. Sterne-Weiler,T., Weatheritt,R.J., Best,A.J., Ha,K.C.H. and Blencowe,B.J. (2018) Efficient and Accurate Quantitative Profiling of Alternative Splicing Patterns of Any Complexity on a Laptop. *Mol Cell*, **72**, 187-200.e6.
9. Vaquero-Garcia,J., Barrera,A., Gazzara,M.R., González-Vallinas,J., Lahens,N.F., Hogenesch,J.B., Lynch,K.W. and Barash,Y. (2016) A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife*, **5**, e11752.
10. Irimia,M., Weatheritt,R.J., Ellis,J.D., Parikshak,N.N., Gonatopoulos-Pournatzis,T., Babor,M., Quesnel-Vallièeres,M., Tapial,J., Raj,B., O’Hanlon,D., *et al.* (2014) A Highly Conserved Program of Neuronal Microexons Is Misregulated in Autistic Brains. *Cell*, **159**, 1511–1523.
11. Tapial,J., Ha,K.C.H., Sterne-Weiler,T., Gohr,A., Braunschweig,U., Hermoso-Pulido,A., Quesnel-Vallièeres,M., Permanyer,J., Sodaei,R., Marquez,Y., *et al.* (2017) An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. *Genome Res*, **27**, 1759–1768.
12. Braunschweig,U., Barbosa-Morais,N.L., Pan,Q., Nachman,E.N., Alipanahi,B., Gonatopoulos-Pournatzis,T., Frey,B., Irimia,M. and Blencowe,B.J. (2014) Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res*, **24**, 1774–1786.
13. Saraiva-Agostinho,N. and Barbosa-Morais,N.L. (2019) psichomics: graphical application for alternative splicing quantification and analysis. *Nucleic Acids Res*, **47**, e7–e7.