ASpli: Integrative analysis of splicing landscapes through RNA-Seq assays

Mancini, Estefania [1†], Rabinovich, Andres [2,3†], Iserte, Javier [2,3], Yanovsky, Marcelo [2,3*], and Chernomoretz, Ariel [2,4*]

[1] CRG, Barcelona, Spain
[2] Fundacion Instituto Leloir, Buenos Aires, Argentina
[3] Instituto de Investigaciones Bioquímicas de Buenos Aires, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina
[4] Departamento de Fisica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Instituto de Fisica de Buenos Aires, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina
[†] Equally contributed

Abstract

Motivation: Genome-wide analysis of alternative splicing has been a very active field of research since the early days of Next Generation Sequencing technologies. Since then, ever-growing data availability and the development of increasingly sophisticated analysis methods have uncovered the complexity of the general splicing repertoire. A large number of splicing analysis methodologies exist, each of them presenting its own strengths and weaknesses. For instance methods exclusively relying on junction information do not take advantage of the large majority of reads produced in an RNA-seq assay, isoform reconstruction methods might not detect novel intron retention events, some solutions can only handle canonical splicing events, and many existing methods can only perform pairwise comparisons.

Results: In this contribution, we present ASpli, a computational suite implemented in R statistical language, that allows the identification of changes in both, annotated and novel alternative splicing events and can deal with simple, multi-factor or paired experimental designs. Our integrative computational workflow, that considers the same GLM model applied to different sets of reads and junctions, allows computation of complementary splicing signals. Analyzing simulated and real data we found that the consolidation of these signals resulted in a robust proxy of the occurrence of splicing alterations. While the analysis of junctions allowed us to uncover annotated as well as non-annotated events, read coverage signals notably increased recall capabilities at a very competitive performance when compared against other state-of-the-art splicing analysis algorithms. ASpli is freely available from the Bioconductor project site https://www.bioconductor.org/packages/ASpli

1 Introduction

The vast majority of protein coding genes in eukaryotic organisms are transcribed into precursor messenger RNA molecules carrying protein coding regions (exons) interleaved by non-coding ones (introns). The later are removed in a co-transcriptional dynamical maturation process called splicing. Alternative splicing (AS) occurs whenever distinct splicing sites are selected in this process resulting in different mature mRNA molecules (Breitbart *et al.*, 1987; Nilsen and Graveley, 2010).

Far from being an exception, it was found that AS is a rather common mechanism of gene regulation that serves to expand the functional diversity of a single gene (Brett *et al.*, 2002). Five basic modes of AS are generally recognized: the skipping of a given exon (exon skipping, ES), the boundary change of a given exon produced by the use of alternative 5' donor (Alt5') or 3' acceptor (Alt3') sites, the retention of an intron in the mature mRNA form (intron retention IR), and the alternative use of mutually exclusive exons (MEX). These canonical forms of AS are prevalent among eukaryotes, although their relative incidence might vary between them (Reddy *et al.*, 2013). Despite their ubiquity, these simple patterns that mainly involve binary choices of exons, donor and acceptor sites, do not exhaust the splicing repertoire. On the contrary, much more complex biologically relevant patterns could arise (Kahles *et al.*, 2016; Vaquero-Garcia *et al.*, 2016).

Nowadays, the analysis of AS at genomic scale is routinely done using RNA-seq assays (Ozsolak and Milos, 2010; Wang *et al.*, 2009). Roughly speaking, there are three main computational approaches to study splicing diversity from RNA-seq data. For one hand there are transcript reconstruction methods, like MISO (Katz *et al.*, 2010) or Cufflink (Trapnell *et al.*, 2012) that aim at inferring usage frequency using probabilistic models for each isoform from the read distribution mapped to a given gene. In the same spirit, Kallisto (Bray *et al.*, 2016) and Salmon(Patro *et al.*, 2017) are two recently introduced methods that leverage on light-weight pseudo-alignment heuristics to quantify transcript abundances. On the other hand, methods like rMATS (Shen *et al.*, 2014), MAJIQ (Vaquero-Garcia *et al.*, 2016) or LeafCutter (Li *et al.*, 2018) use junction information to infer both, annotated and novel splicing events. Finally, methods like voom-limma (Law *et al.*, 2014), DEXSeq (Anders *et al.*, 2012) and edgeR (McCarthy *et al.*, 2009, 2012) focus on the analysis of differential usage of subgenic features (e.g. exons) between conditions.

In practice, the study of AS faces many technical challenges, with quantitative approaches suffering methodological or scope limitations (Ding *et al.*, 2017; Liu *et al.*, 2014; Mehmood *et al.*, 2019; Zhang *et al.*, 2017). For instance, methods that exclusively rely on junction information might compromise their sensitivity performance as they do not take advantage of the large majority of reads produced in a sequencing run, isoform quantification approaches might not detect novel intron retention events, many existing methods can only handle pairwise comparisons between conditions and some focus only on the analysis of canonical splicing events.

In this paper we present ASpli, a computational suite that was specifically designed to integrate coverage and junction splicing cues in order to address these challenges. In order to weigh ASpli's performance we compared it against three different state-of-the-art methodologies: rMATS (Shen *et al.*, 2014), LeafCutter (Li *et al.*, 2018) and MAJIQ (Vaquero-Garcia *et al.*, 2016). The first one is a widely used piece of software that can integrate coverage and junction information to assess changes in splicing patterns. Additionaly, LeafCutter and MAJIQ are two recently introduced methodologies that are commonly used by the bioinformatics community. Both approaches focus on the analysis of clusters of junctions to study local splicing patterns of varying complexity. However, they differ in many technical and statistical aspects (Vaquero-Garcia *et al.*, 2016). For instance, LeafCutter was not designed to handle intron retention events and considers a Dirichlet-multinomial generalized linear model to test for differential intron excision between two groups of samples. MAJIQ, on the other hand, relies on a bayesian estimation of the posterior Percent Selected Index to identify splicing affected junctions.

The paper is organized as follows. In section 2.1 we introduced the typical ASpli workflow. In Section 2.2 we analyzed a simulated dataset in order to evaluate the specificity and sensitivity of ASpli discoveries and contrasted these results against LeafCutter, MAJIQ and rMATS outcomes. In Section 2.3 we explored the ability of ASpli to uncover consistent splicing-patterns from two independent RNA-seq assays that probed the alterations of splicing patterns of *A. thaliana* transcriptome caused by the knock out of PRMT5, a methyl transferase that, among other proteins, targets several Sm spliceosomal proteins (Hernando *et al.*, 2015a; Roworth *et al.*, 2019; Sanchez *et al.*, 2010). This analysis was also performed with the other considered methodologies in order to compare their capacity to generate reproducible results. In this section we also aimed to quantify the level of agreement of ASpli, LeafCutter, MAJIQ and rMATS discoveries with qRT-PCR based alternative splicing evidence. To that end, we took advantage of two independent studies that analyzed splicing altered events in PRMT5 mutants using qRT-PCR assays (Deng *et al.*, 2010; Sanchez *et al.*, 2010). Finally, in Section 2.4, we considered a 28 samples paired-study of human prostate cancer (Ren *et al.*, 2012). Using this dataset we analyzed how the performance, time and memory requirements scaled with the number of considered samples in a paired experimental design. Finally, we discussed our results in Section 4 and presented our conclusions in Section 5.

2 Results

2.1 ASpli description

Like DEXseq (Anders *et al.*, 2012) ASpli analyzes read coverage over *bins*, defined as maximal sub-genic features entirely included or entirely excluded from annotated mature tanscripts (see Figure 1.1, and Sup.Mat. 1.1 for details). Read junctions are considered in order to provide supporting evidence of differential bin usage. Additionally, annotation-independent signals are estimated based on the complete set of experimentally detected splice junctions (see Figure 1.2, Material and Methods 3.1, and Sup.Mat. 1.2 for further details.).

We leverage on the statistical framework implemented in edgeR (Robinson *et al.*, 2010) to asses for the statistical significance of different splicing signals in a consistent and unified way. As depicted in Figure 1.3, statistically significant evidence is collected from: bin coverage differential signals, relative changes in junction's flanking genomic regions (junction-anchorage signal), and junction usage variations displayed inside junction clusters (junction-locale signal). A detailed description of the considered differential splicing signals and event acceptance criteria can be found in Material and Methods 3.2 and 3.3 respectively.

In practice, a typical ASpli workflow involves several steps: parsing the genome annotation into bins, counting aligned reads, assessing for differential bin and junction usage, consolidating junction and bin-coverage splicing evidence and, finally, reporting integrated splicing signals (see Supplementary Figure

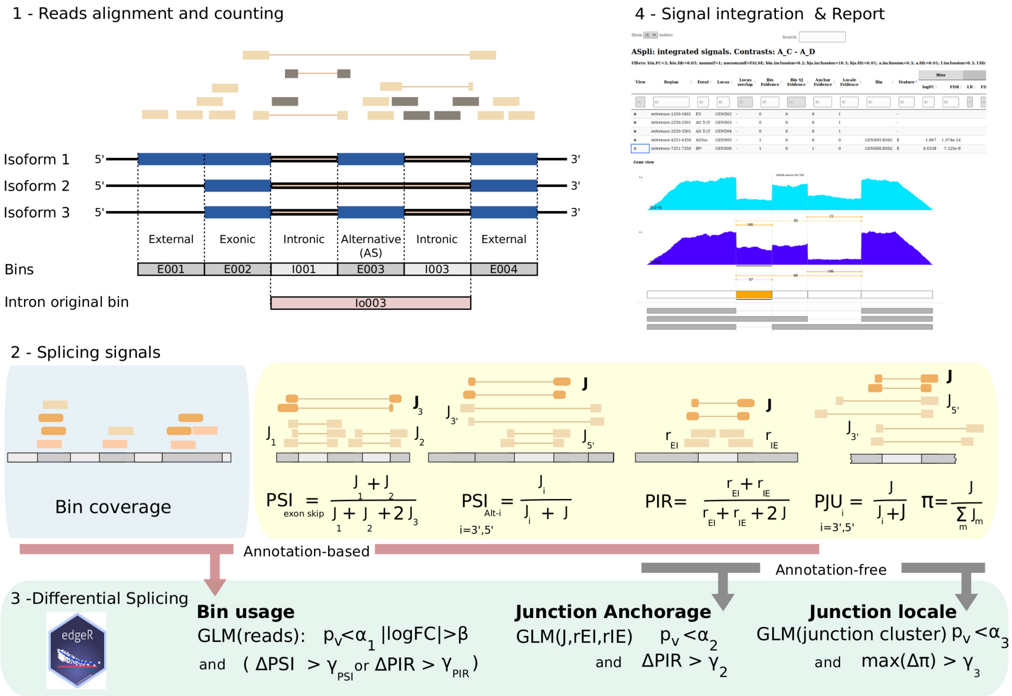


Figure 1. Panel 1: binning of a three-isoform gene. Six bins (plus one intron-original bin) were defined. According to the local bin classification model two external, two intronic, one exonic and one alternative bin were recognized. Reads aligned to the genome are used to tally bin counts. In the figure, darker boxes represent non-annotated junctions and reads. Panel 2: Summary of splicing signals. Bin coverage is depicted in the blueish sub-panel. Junction-based signals are shown in the yellow sub-panel. The Percent-Spliced-In (PSI) statistics is used to quantify the inclusion evidence of exon-skipping and for the alternative use of 5' or 3' splicing sites. On the other hand, the Percent-of-Intron-Retention (PIR) is considered for IR events. It combines counts for the junction of interest and counts tallied over exon-intron boundaries (r_{IE} and r_{EI}). The Percent-Junction-Usage (PJU) statistics (rightmost junction-panel) quantifies the relative strength of a given junction with respect to locally *competing* ones for alternative 5' or 3' splicing sites. The participation coefficient Π serves to quantify the relative strength of a given junction J with respect to the set of junctions of the corresponding junction-cluster. For annotation-compliant events ASpli takes into account bin-coverage along with $\text{PSI}_{\text{exon skip}}$, $\text{PSI}_{\text{Alt i}}$ or PIR signals depending on the event-class assigned to the analyzed bin. PIR and PJU_{i} signals estimated for experimentally detected junctions are use to produce annotation-free statistics. Panel 3: Summary of differential splicing signals. ASpli applies a GLM approach on three different sets of reads and junctions to define complementary splicing cues: differential bin-coverage, junction anchorage and junction-locale signals. 4: Example of ASpli generated output.

2.2 Synthetic dataset

Changes in splicing patterns were simulated in a treatment vs control setup (three samples per condition) for genes of the *Arabidopsis thaliana* plant genome. For the sake of simplicity, and without loss of generality, we just considered chromosome 1 of *A.th*).

We implemented a computational pipeline relying on the Flux Simulator (FS) software (Griebel *et al.*, 2012) in order to produce a controlled set of splicing events. We used FS to generate a power-law transcript abundance distribution template with $15e6$ molecules spread among the 10646 available transcript variants of 8433 genes of chromosome-one of the *A.th* genome. Then, we generated a 'treatment' transcriptional profile altering the original molecule distribution in order to simulate genome-wide differential changes in gene expression and splicing patterns. For the first case, we increased (or decreased) the assigned number of simulated sequenced molecules for a randomly selected fraction of genes to get a desired fold change level. For splicing events, we re-distributed the number of simulated sequenced molecules among different isoforms of a given gene. In our simulations, the differential usage of splicing variants affected 2451 genomic bins hosted in 915 genes (see Material and Methods 3.4 for further details).

In Table 1 we reported the number of correctly detected simulated events, number of false positives and number of events exclusively detected by each kind of signal: bin-coverage, junction-locale and junction-anchorage. A graphical summary of the detection signal landscape can be appreciated in Figure 2.

It can be seen that ASpli correctly uncovered 974 (40%) of the 2451 simulated bin events. Moreover, we found that most of the ASpli undetected simulated events (1341 out of 1477) took place in genes that presented scarce expression levels in at least one simulated condition. These barely expressed genes were filtered out before any statistical testing (see 3.3). The rationale behind this pre-filtering step before any statistical testing was three-fold. First, we assumed that for a splicing analysis to be biologically sound, a gene should exhibit a well established minimal expression level (a gene can not be spliced if it is not expressed). In second place, statistical tests applied to genes with few counts return unstable results. Finally, leaving out these genes would alleviate the effect of multiple testing correction. In the context of our analysis, this means that 1341 of the undetected simulated events involved barely expressed genes and, according to our criteria, were probably unreliable events. Noticeably, only 136 out of the 1110 events (12%) that did pass the gene-expression pre-filtering step were found to be false negative cases.

About 95% of ASpli true discoveries were identified by the analysis of significant changes at the bin-coverage level. Junction-based detection, on the other hand, could correctly identified 574 simulated events (60% of true discoveries). The null overlap between locale and anchorage detection illustrated that they probed complementary aspects of splicing events. Additionally, it can be appreciated that 41% (399) of true discoveries were only detected by bin-coverage signals, whereas junction-based analysis contributed only 5% (50) of specific detections. We included in Supplementary Material 4 a further characterization of the junction-support and fold-change signals involved in bin-coverage detection calls. In particular, we found that with the adopted 3-fold threshold ASpli achieved high recall and precision levels ($\sim 80\%$ and $\sim 95\%$ respectively) laying at rather moderate levels of false positive rates ($\sim 14\%$).

In Table 2 we reported the detection performance of ASpli along with outcomes from the other considered algorithms (see Sup Mat 2 and 3 for calculation details). Precision and recall values estimated at gene-level (in which a gene was reported as a discovery whenever at least one alternative-splicing event was detected within its genomic range) were reported between parenthesis. ASpli outcomes considering only coverage signal or just junction signals were included in the table as ASpli$_c$ and ASpli$_j$ rows respectively.

It can be seen from the table that all tested algorithm shown rather high precision values. However, ASpli benefited from larger recall scores than any other methodology. Moreover, it can be appreciated that ASpli$_j$ displayed only marginally larger recall

ASpli signal	TP	FP
bin coverage	924 (399)	42
junction locale	393 (35)	2
junction anchorage	182 (15)	6
overall	974	48

Table 1. Splicing detection performance of the three different ASpli signals. True positive and false positive calls are shown in the second and third columns respectively. The number of specific discoveries exclusively reported by each signal is reported between brackets.

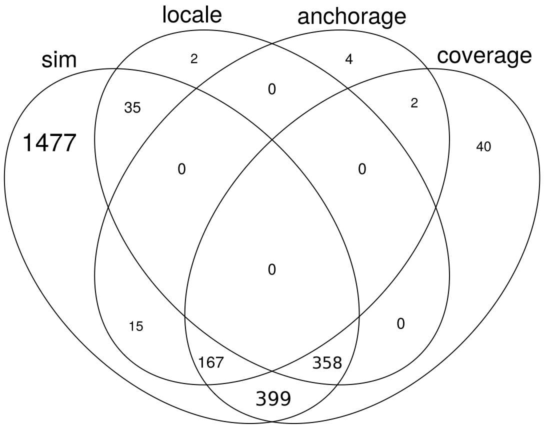


Figure 2. Distribution of detection calls produced by different ASpli signals. Simulated splicing events are included in the *sim* set. ASpli calls detected by junction-locale, junction-anchorage and bin-coverage signals are counted in the respective Venn sets

method	size	precision	recall
ASpli	1022 (631)	0.95 (0.99)	0.40 (0.68)
ASpli$_c$	966 (591)	0.96 (0.99)	0.38 (0.64)
ASpli$_j$	583 (456)	0.99 (0.99)	0.23 (0.50)
LeafCutter	204 (163)	0.93 (0.91)	0.08 (0.16)
MAJIQ	538 (381)	0.84 (0.87)	0.18 (0.36)
rMATS	405 (352)	0.87 (0.91)	0.14 (0.35)

Table 2. Number of discoveries, precision and recall levels are reported for different detection methodologies. $ASpli_c$ and $ASpli_j$ correspond to ASpli discoveries detected using just coverage or just junction signals respectively. Values between parenthesis report quantities estimated at gene-level.

levels than other methodologies implying that ASpli leveraged on coverage signals to increase this figure of merit. All of these results suggested that ASpli was capable of reliably detect the simulated splicing events achieving notably high recall values at very competitive levels of precision and specificity.

2.3 Reproducibility Analysis

PRMT5 is a methyl transferase that, among other proteins, targets several spliceosome proteins. Its deletion has been proved to provoke major splicing alterations (Hernando *et al.*, 2015a; Sanchez *et al.*, 2010). We analyzed two independent RNA-seq assays that were conducted at different times probing the same biology. Experiments *A*

(GSE149429) and B (GSE149430) were originally carried out to analyze splicing alterations in the PRMT5 knock-out mutant in *Arabidopsis thaliana*. Both assays probed the PRMT5-KO and wild-type transcriptomes in Columbia ecotype plants as part of larger and different studies (see Material and Methods 3.5).

2.3.1 Reproducibility assessment

We analyzed RNA-seq assays A and B with ASpli and the other considered algorithms. For ASpli, we used the detection-call criteria specified in Section 2.2. Default parameters were considered to run the other tested methodologies (command lines used to execute them were included as Supplementary Material 2). For LeafCutter and rMATS we considered events presenting fdr corrected pvalues smaller than 0.05 and changes in junction inclusion indices larger than 0.1. For MAJIQ we sought for events presenting a posterior probability larger than 0.95 of having a change in inclusion index larger than a 0.2 level.

Overall, $6350, 951, 412$ and 158 genomic regions affected by altered splicing patterns were reported by ASpli, LeafCutter, MAJIQ and rMATS algorithms respectively in at least one experiment. In Table 3 we summarized reproducibility statistics for each examined methodology (a more in-depth comparison of discoveries was included as supplementary material in Section 6). Splicing analysis algorithms, including ASpli, usually perform a pre-filtering step before the statistical analysis. So we found it useful to explicitly state how many subgenic regions got actually tested in each algorithm to better contextualise the number of discoveries and the between-experiments agreement level according each algorithm. Column *universe* of Table 3 reports the actual number of sub-genic regions that, upon passing different pre-filtering steps, were actually examined for statistically significant changes in splicing patterns. Notice that the extent of this background list was noticeably larger for ASpli as our methodology tested not only junction-related signals but also alterations in the usage of genomic bins.

Columns A and B outline the number of regions reported as differentially spliced in each experiment and column $A \cap B$ the discovery intersection size, i.e. number of sub-genic regions reported as differentially spliced in both data-sets (we verified that the found differences were in the same direction in A and B). In parenthesis, we included the *overlap coefficient value*, defined as $A \cap B / \min(A, B)$. Expected overlaps, fold enrichment (i.e. ratio between observed and expected overlaps) and p-values were estimated using the SuperExactTest R-package (Wang *et al.*, 2015) and reported in EO, FE and *pval* columns respectively.

Method	universe	A	B	$A \cap B$	EO	FE	pval
ASpli	140191	4687	3904	2241 (0.57)	130.5	17.2	0.0e+00
leafCutter	8113	603	675	327 (0.54)	50.2	6.5	3.6e-219
MAJIQ	16441	277	284	149 (0.54)	4.8	31.1	9.5e-203
rMATS	6413	310	401	310 (1.00)	19.4	16.0	0.0e+00

Table 3. Reproducibility statistics. The numbers of statistically analyzed sub-genic regions (after prefiltering steps) for each algorithms are shown in the *universe* column. The number of splicing events reported for each experiment and the number of concordant discoveries are displayed at columns A, B, and $A \cap B$ respectively. The expected overlap, fold enrichment level and significance pvalue are displayed in columns EO, FE and pval respectively

It can be seen from Table 3 that the agreement between experiments was highly significant for all the examined methods. In all cases, more than 50% of events detected in one experimental instance was also reported in the other. At the same time it can be appreciated that ASpli provided the largest (and highly significant) overlap-set. Noticeably, the total number of concordant splicing-affected genomic regions detected by ASpli presented up to a 15-fold increase with respect to the size of concordant sets reported by others methodologies.

Overall our analysis showed that results obtained at different and independent experimental instances were reproducible, in the sense that statistically significant agreement was found for every methodology. These results were robust against using different overlap quantification criteria (see Sup.Mat 6).

2.3.2 PRMT5 RT-PCR detected events

The PRMT5 methyltransferase has been the target of many studies as deficiencies in this protein causes genomewide splicing alterations (Deng *et al.*, 2010; Hernando *et al.*, 2015a; Roworth *et al.*, 2019). In this section we focused on two specific works that provided independent RT-PCR validated lists of splicing alterations linked to PRMT5 in *Arabidopsis thaliana* (Deng *et al.*, 2010; Sanchez *et al.*, 2010).

For one hand, Deng and collaborators studied PRMT5 mutant *Arabidospis thaliana* plants and presented a list of 12 RT-PCR validated intron retention events (see Fig 2 in (Deng *et al.*, 2010)). On the other, using the same biological model, Sanchez and collaborators indentified changes in alternative splicing using a high-resolution qRT–PCR panel that included several known alternative splicing events (Simpson *et al.*, 2008). They found that PRMT5 mutants had significant alterations in 44 events which included exon skipping, alternative donor and acceptor splice sites, as well as intron retention events (Supplementary Table 4 in (Sanchez *et al.*, 2010)).

We aimed to contrast these findings with the results reported by the different methodologies on datasets A and B. In Table 4 we summarized, for each study, the number of concordant findings uncovered by different algorithms on datasets A and B. We also reported ASpli results for a consolidated dataset *AB*, obtained taking advantage of ASpli capabilities to handle complex experimental designs (see Supplementary Material 7). Quantities between brackets represent the number of ASpli discoveries reported by coverage and junction-based signals respectively.

It can be seen that ASpli recovered the largest number of events and that the majority of ASpli validated discoveries originated in differential coverage signal calls. Had we only considered junction related detection-signals, ASpli would have achieved similar levels of agreement than the other junction-based algorithms (for instance we got a similar performance than LeafCutter on Sanchez dataset for the consolidated case).

In Table ST.2, included as supplementary material, we further characterized the agreement between the 23 splicing events that ASpli uncovered for the consolidated AB case, and Sanchez qRT-PCR validated events. It can be seen that in 15 out of the 23 cases (65%), the very genomic region probed by the PCR analysis was recognized by ASpli. For the other 8 cases, ASpli detected actually occurrying changes in isoform usage but from splicing signals taking place at genomic locations not probed by the PCR primers.

2.4 ASpli scalability analysis

In this section we leveraged on a mid-size RNA-seq study presented by Ren and collabrators to characterize aberrant splicing patterns occurying in prostate cancer patients (Ren *et al.*, 2012). We aimed to analyze this sample-paired assay to see how ASpli performance (statistical power, precision, time and memory requirements) scaled with the number of samples. In particular, we followed the approach suggested in (Mehmood *et al.*, 2019) to characterize ASpli in terms of statistical power and expected false discovey rate for a varying number of samples.

Method	Dataset	Deng 2010	Sanchez 2010
RT-PCR	-	12	44
ASpli	AB	8 [8,4]	23 [21,13]
	A	10 [8,5]	24 [19,7]
	B	9 [8,2]	20 [18,6]
LeafCutter	A	3	16
	B	3	17
MAJIQ	A	5	8
	B	2	8
rMATS	A	1	12
	B	1	3

Table 4. Number of RT-PCR splicing validated events reported in Deng and Sanchez studies are reported in the first row. For each study, the number of concordant findings uncovered by different algorithms on dataset A and B are reported in subsequent rows (ASpli results for the consolidated dataset *AB*, were also included). Quantities between brackets represent the number of ASpli discoveries reported by coverage and junction-based signals respectively.

2.4.1 Statistical power

Ren and collaborators presented a comprehensive study of splicing alterations detected using RNA-seq transcriptome profiles of 14 primary prostate cancers and their paired normal counterparts from the Chinese population (Ren *et al.*, 2012). On average, the 28 fastq files presented 34.6 ± 1.7 million reads per sample and 31.4 ± 1.6 millions of them were actually mapped to the ENSEMBL HG38.98 version of the human genome (see Material and Methods 3.5). The genome's GTF and BAMs files were then used as inputs to drive an alternative splicing paired-sample analysis with ASpli. We considered the $y \sim patient + tissue$ model to identify genomic regions differentially spliced in tumor samples compared to normal tissue controls. The *patient* term served to pair tumor and normal tissue samples coming from the same individual. The two-level *tissue* factor reported average differences between tumor and normal cases over the observed population of patients.

In order to study the dependency of the statistical power on the number of samples, we sampled without replacement (10 times) subsets of 3, 5, 7 and 10 individuals. For each case, we reported, in the first column of Table 5, the median (and standard error, in brackets) of the number of genomic regions found to be alternatively spliced between tumor and normal samples.

In order to estimate false discovery rates we considered mock comparisons between normal samples (we sampled 10 times normal tissue samples of 3vs3, 5vs5 and 7vs7 individuals). We then estimated FDR as the ratio between the number of mock discoveries and the median number of discoveries found in true comparisons of the same number of samples. In the second column of Table 5 median and standard errors (in brackets) were reported.

It can be seen from Table 5 that the median number of detected splicing events increased with the number of examined samples, up to a maximum of 1465 events obtained when the 28 paired samples were considered. The large variability observed between bootstrap realizations was consistent with the large variability already observed across prostate cancer transcriptomes (see (Ren *et al.*, 2012) and Supplementary

Material 9). FDR estimated values showed a huge decrease with increasing number of samples, and for the 5x5 case seemed to have already leveled off. Similar trends were observed when splicing alterations were reported at the level of hosting genes.

2.4.2 Time and memory requirements

In Table 5 we reported median values and standard errors for the elapsed time and peak memory usage required for calculations (performed on single thread on an Intel Xeon Silver 4116 2.1GHz Lenovo ThinkSystem SR650)

Execution time scaled linearly with the number of paired samples at a rate of 25.5 minutes per pair of samples (about 90% of execution time was used for BAMs reading and feature counting). The memory peak column shows that RAM requirement linearly scaled with the number of samples at a rate of about 880MB per sample pair. A simple extrapolation suggests that about 65GBshould be enough to handle 100 samples of the same sequencing depth ($\sim 3.510^6$ reads per sample).

Samples	Splicing events		Affected genes		Benchmark	
	number	FDR	number	FDR	time [min]	memory peak[GB]
3x3	67 (155)	0.2 (0.4)	44 (113)	0.25 (0.03)	67 (1)	20.25 (0.38)
5x5	486 (387)	0.02 (0.002)	371 (218)	0.02 (0.002)	111 (2)	22.15 (0.20)
7x7	759 (220)	0.005 (0.02)	481 (131)	0.004 (0.0007)	156 (4)	24.13 (0.03)
10x10	850 (418)	-	664 (191)	-	231 (5)	26.57 (0.04)
14x14	1465	-	1030	-	348	30.07

Table 5. Summary of the 10-fold bootstrapped analysis of ASpli performance on the prostate cancer data set. For each number of paired samples (first column) the median number of genomic-regions displaying a statistically significant 'tissue' effect were included in the second column. Median values of false discovery rate estimations obtained from the analysis of normal-tissue samples were shown in the third column. Median number of genes hosting splicing alterations are included in the fourth column. Median values of false discovery rate estimations obtained from the analysis of normal-tissue samples for genes hosting splicing alterations were shown in the fifth column. Median time and memory used in the analysis were reported in the sixth and seventh columns respectively. Standard error estimations were reported between brackets.

3 Material and Methods

3.1 Splicing signals

For each bin, ASpli uses the Percent Spliced-In (PSI) and the Percent of Intron Retention (PIR) metrics to quantify the relative weight of inclusion evidence (Schafer *et al.*, 2015). The definition of this quantities are pictured in the three leftmost junction-cases displayed in Figure 1.2.

For annotation-free splicing detection, every experimentally detected junction that passed a pre-filtering step is considered for the analysis. Local splicing patterns are then analyzed through three different metrics (see two rightmost junction cases of Figure 1.2). 1) the PIR metric is used to characterize hypothetical intron retention events. 2) the Percent of Junction Usage (PJU) statistics quantifies relative abundance of a given junction with respect to locally competing ones sharing, alternatively, a 5' or 3' splicing site, and 3) the participation coefficient, π, quantifies the relative abundance of the analyzed junction with respect to junctions belonging to the same junction-cluster (in the same spirit than MAJIQ and LeafCutter, ASpli defines junction-clusters as sets of junctions that share at least one end with another junction of the same cluster).

3.2 Differential analysis scheme

ASpli leverages on the statistical framework developed by Smyth and collaborators, implemented in the edgeR R-package (McCarthy *et al.*, 2012; Robinson *et al.*, 2010), to assess for statistically significant changes in gene-expression, bin coverage and junction splicing signals. Under this approach, count data is modeled using a negative binomial model, and an empirical Bayes procedure is considered to moderate the degree of overdispersion across units.

Differential expression signals Differential expression signals are estimated via generalized linear models (GLM). In this way, contrasts can be tested in experiments with multiple experimental factors. Using this statistical setting, for each gene, ASpli quantifies differential gene expression signals reporting log-fold changes, p-values, and FDR adjusted q-values.

Differential splicing signals In order to study splicing patterns, gene expression changes should be deconvolved from overall count data. We considered the same GLM model, applied to different sets of reads and junctions, in order to compute complementary splicing signals. On a very general setting, what we are looking for is to test whether a given unit (e.g. a bin) of a certain group of elements (e.g. the whole set of bins of the corresponding gene) displays differential changes respect to the collective or average behavior. ASpli uses this general idea to assess for statistically significant changes in splicing patterns probed with different genomic features (see Figure 1.3)

- bin-coverage signal: differential usage of a bin is assessed comparing single bin log-fold-changes against the expression log-fold-change of the corresponding gene.
- junction locale signal: In order to characterize changes for a given junction along experimental conditions, ASpli weighs log-fold-change of the junction of interest relative to the mean log-fold-change of junctions belonging to the same junction-cluster.
- junction anchorage signal: For every experimentally detected junction, ASpli analyzes differential intron retention changes by considering log-fold-changes of a given experimental junction relative to coverage changes of left and right junction flanking regions.

ASpli makes use of the functionality implemented in the diffSpliceDGE function of the edgeR package to perform all of this comparisons within a unified statistical framework. Given a set of elements (i.e. bins or junctions) of a certain group (i.e. genes, anchorage group or junction-cluster), a negative binomial generalized log-linear model is fit at the element level, considering an offset term that accounts for library normalization and collective changes. Differential usage is assessed by testing coefficients of the GLM. At the single element-level, the relative log-fold-change is reported along with the associated p-value and FDR adjusted q-values. In addition a group-level test is considered to check for differences in the usage of any element of the group between experimental conditions (see *diffSpliceDGE* documentation included in edgeR package for details (Robinson *et al.*, 2010)).

3.3 Filtering and detection criteria

Statistical analysis of differential splicing is performed only on expressed genes (i.e. read counts spanning the gene genomic range should be larger than a minimal number of reads, 5 by default across all the samples of the contrasted conditions). Furthermore, analyzed bins and junctions should present a minimal number of counts (5 by default) in every replicate of at least one contrasted condition. Additionally, marginally present junctions are filter-out looking at the maximal value of their *participation* coefficient, defined as the relative abundance of a given junction within its group for a given experimental condition.

Besides statistical figures of merit, ASpli considers PSI, PIR, PJU and Π statistics to ease the identification of biologically relevant events. In this way, a bin is called differentially-used by ASpli if it displays statistically significant coverage changes (fdr < 0.05, by default) and, additionally, one of the two supplementary conditions hold: either the bin fold-change level is greater than a given threshold (3 fold changes, by default) or changes in inclusion levels of bin-supporting junctions (ΔPIR or ΔPSI according to the bin class, see Table ST.1) surpasses a predefined threshold (0.2 by default).

Anchorage splicing signals, on the other hand, are reported whenever statistically significant changes are found at the cluster level (cluster.fdr < 0.01 by default) for the considered $\{r_{EI}, r_{IE}, J\}$ read set and, at the same time, $|\Delta\text{PIR}_J|$ is larger than a given threshold (0.3 by default)(see Figure 1.3).

Finally, junction locale differential splicing signals are reported whenever statistically significant changes are found at the cluster level (cluster.fdr < 0.01 by default) for the analyzed junction cluster $\{J_1, ..., J_S, ..., J_n\}$ (see SF.1-E) and, at the same time, there is at least one junction J_S within the cluster presenting statistically significant changes at the single unit level (junction.fdr < 0.05, by default) with $|\pi_{J_S}|$ larger than a given threshold (0.3 by default). In the case that statistically significative changes were detected at the unit-level for more than one junction of a given cluster, the one displaying the largest participation change was considered and reported as the cluster's representative junction.

3.4 Splicing simulation

We used the Flux Simulator (FS) software (Griebel *et al.*, 2012) in order to produce a controlled set of splicing events. Once 'control' and 'treatment' samples were generated as explained in Section 2.2, we obtained simulated biological replicates from the two *seed* transcriptomes, considering a Gamma distribution for molecule abundances. We chose to work with a $CV = 0.1$ level of variability in gene abundance between replicates. Therefore, we considered *shape* ($k = 100$), and *scale* ($\theta = 0.01\mu$) parameter values, where μ was the gene expression level in the corresponding *seed* transcriptome used for replicate generation.

Simulated changes in isoform concentrations for a gene produce specific patterns of bin and junction differential usage that depend on the exonic architecture of the different gene variants. For instance, a splicing alteration that involves switching between Isoform 1 and Isoform 3 of the gene depicted in Figure 1.1 is expected to produce differential usage signals only for the first and third exonic bins (E001 and E003). In our case we simulated changes in variants usage for 915 genes that should altered, in principle, the coverage signal of 2451 bins. It is worth mentioning that as alternative splicing was modeled exclusively through differential variant usage, no intron retention events were simulated in the synthetic data set (and thus, anchorage-signals were not actually produced in this dataset).

3.5 Data and Code availability

PRMT5: Datasets A and B are both available from the GEO repository super-series GSE149431. The goal of these studies was to compare the transcriptional profile (RNA-seq) of wild type and PRMT5 Arabidopsis mutants plants grown under continuous light at 22 degrees centigrades. On average, 19.3 ± 5.3 million 100 long and 28.3 ± 7.7 million 150 long paired-end reads were generated per sample library for datasets A and B respectively. For both cases more than 96% of reads were uniquely mapped to TAIR10 Arabidopsis genome using STAR (command-line invocation was included in Sup Mat 2).

Prostate cancer dataset: Fifty-six paired-end fastq files from the E-MTAB-567 experiment were downloaded from the ArrayExpress server. Reads were aligned against ENSEMBL HG38.98 reference genome using the STAR aligner with default parameters and a junction overhang parameter equal to $(\text{readlength} - 1)$.

Code availability: ASpli package is freely available from Bioconductor site `https://www.bioconductor.org/packages/ASpli`. Several examples, use-cases and a detailed description of produced outputs can be found in ASpli package's vignette. Scripts and additional material to reproduce the analysis presented in this paper can be found at the github repo: `https://github.com/chernolab/ASpli_SM`. Issues and suggestions can be uploaded at `https://github.com/chernolab/ASpli`

4 Discussion

RNA high-throughput sequencing methods provide powerful means to study alternative splicing under multiple conditions in a genome-wide manner. However, the detection and understanding of general splicing patterns still present considerable technical challenges. Here we presented ASpli, a computational suite to comprehensively test bin coverage and junction usage differential splicing signals.

The analysis methodology implemented in ASpli came out as a result of several software maturation cycles of our in-house splicing analysis procedures. Over the last years, the presented core functionality has been extensively used in different projects to study: the role of AS in circadian rhythms and light response (Mancini *et al.*, 2016; Perez-Santángelo *et al.*, 2014; Romanowski *et al.*, 2019; Rugnone *et al.*, 2013; Xin *et al.*, 2017) and AS in spliceosome mutants (Hernando *et al.*, 2015b; Schlaen *et al.*, 2015) in A. thaliana model organism. In addition, ASpli in-house versions have been used to study AS and rhythmic behavior in D.melanogaster (Beckwith *et al.*, 2017) and to characterize AS in dengue's viral infection in humans (De Maio *et al.*, 2016).

In order to quantify ASpli's performance we compared it against three different state-of-the-art methodologies: LeafCutter (Li *et al.*, 2018), MAJIQ (Vaquero-Garcia *et al.*, 2018) and rMATS (Shen *et al.*, 2014). As a general rule we considered default parameters to run these analysis pipelines for our intention was not to present here an extensive benchmark between bioinformatics approaches, nor to propose the definitive analysis methodology. Rather we wanted to establish whether ASpli produced reasonable and competitive results.

Different scenarios were considered to chart ASpli performance. We first analyzed a synthetic dataset and quantified the ability of each considered methodology to detect splicing changes in terms of precision and sensitivity figures of merit. Using this controlled dataset we found that all the analysed methods presented rather high precision levels. However ASpli systematically displayed larger recall values ($\sim 40\%$), mainly because the use of coverage signals. This is an important result as highlights the benefits of not loosing effective sequencing depth by relaying not only on junction information but on the complete set of reads of RNA-seq sequencing runs.

We then aimed to outline ASpli's performance over more realistic setups. As no internal gold-standards are usually available for real world datasets we focused on the analysis of two independent RNA-seq assays that probed the same biological conditions. This allowed us to quantify the consistence and coherence of outcomes produced by each methodology in terms of reproducibility of discoveries. Our results suggested that detection agreement between studies was highly significative for every methodology. However ASpli was far superior in terms of total number of concordant discoveries reported.

It is worth noting that a necessary condition implicit in this analysis was that biological variability largely exceeded possible technical biases between studies. Using ASpli, we were able to consider a generalized linear model to define a consolidated dataset integrating data from both studies and verified that this was actually the case (see Supplementary Material 7). In addition, the possibility to implement a two-factor model greatly improved the statistical power to uncover consistent discoveries. We could identify 4314 events displaying a statistically significative genotype effect and no evidence of experiment-genotype interactions. This represented almost a two-fold increase in the number of reproducible discoveries when compared against the naive integrative approach that merely considered the 2241 splicing events simultaneously detected in both studies.

As a general result we found that for both, the synthetic scenario and the PRMT5 analysis, ASpli reported the largest number of discoveries. We think that this is an encouraging feature because large recall performance at controlled precision levels means more statistically supported hypothesis that can be further analyzed and biologically validated.

Finally, an important aspect of the presented approach is that ASpli's core functionality was implemented as user-friendly functions that produce self-contained output results for each step of the analysis. This is an important feature from the user's perspective. It provides the user valuable intermediate information eventually facilitating the integration of ASpli with other analysis pipelines.

5 Conclusions

In this paper we presented ASpli, a computational suite to study alternative splicing events. It is implemented as a flexible R modular package that allows users to fulfill gene-expression and splicing analysis following a set of simple steps. Noticeably, ASpli can deal with multi-factor or paired experimental designs using a unified statistical framework to assess for differential coverage of sub-genic features and junctions. By combining statistical information from exons, introns, and splice junctions ASpli can provide an integrative view of splicing landscapes that might include canonical and non-canonical splicing patterns occurring in annotated as well as in novel splicing variants. Examples and use-cases can be found in the ASpli package's vignette.

Acknowledgements

We thank Ruben Schlaen, Julieta Mateos, Andres Romanowski and Esteban Beckwith for helpful discussions. In memory of Hugo Mancini who suddenly passed away during the revision of this article, and whose unconditional love and support for her daughter EM were essential for her to become a scientist

Funding

This work has been supported by grants from Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT). AC also acknowledges support from University of Buenos Aires (grant 20020170100356BA). AC and MY are members of Carrera de Investigador of Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET).

6 Bibliography

References

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from rna-seq data. *Genome Research*, **22**(10), 2008–2017.

Beckwith, E. J., Hernando, C. E., Polcowñuk, S., Bertolin, A. P., Mancini, E., Ceriani, M. F., and Yanovsky, M. J. (2017). Rhythmic behavior is controlled by the srm160 splicing factor in drosophila melanogaster. *Genetics*, **207**(2), 593–607.

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic rna-seq quantification. *Nature Biotechnology*, **34**(5), 525–527.

Breitbart, R. E., Andreadis, A., and Nadal-Ginard, B. (1987). Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes. *Annu. Rev. Biochem.*, **56**.

Brett, D., Pospisil, H., Valcárcel, J., Reich, J., and Bork, P. (2002). Alternative splicing and genome complexity. *Nature Genetics*, **30**(1), 29–30.

De Maio, F. A., Risso, G., Iglesias, N. G., Shah, P., Pozzi, B., Gebhard, L. G., Mammi, P., Mancini, E., Yanovsky, M. J., Andino, R., *et al.* (2016). The dengue virus ns5 protein intrudes in the cellular spliceosome and modulates splicing. *PLoS pathogens*, **12**(8).

Deng, X., Gu, L., Liu, C., Lu, T., Lu, F., Lu, Z., Cui, P., Pei, Y., Wang, B., Hu, S., and Cao, X. (2010). Arginine methylation mediated by the arabidopsis homolog of prmt5 is essential for proper pre-mrna splicing. *Proceedings of the National Academy of Sciences*, **107**(44), 19114–19119.

Ding, L., Rath, E., and Bai, Y. (2017). Comparison of alternative splicing junction detection tools using rna-seq data. *Curr Genomics*, **18**(3), 268–277. 28659722[pmid].

Griebel, T., Zacher, B., Ribeca, P., Raineri, E., Lacroix, V., Guigo, R., and Sammeth, M. (2012). Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res.*, **40**(20), 10073–10083.

Hernando, C. E., Sanchez, S. E., Mancini, E., and Yanovsky, M. J. (2015a). Genome wide comparative analysis of the effects of prmt5 and prmt4/carm1 arginine methyltransferases on the arabidopsis thaliana transcriptome. *BMC Genomics*, **16**(1), 192.

Hernando, C. E., Sanchez, S. E., Mancini, E., and Yanovsky, M. J. (2015b). Genome wide comparative analysis of the effects of prmt5 and prmt4/carm1 arginine methyltransferases on the arabidopsis thaliana transcriptome. *BMC genomics*, **16**(1), 192.

Kahles, A., Ong, C. S., Zhong, Y., and Ratsch, G. (2016). Spladder: identification, quantification and testing of alternative splicing events from rna-seq data. *Bioinformatics*, **32**(12).

Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B. (2010). Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nature Methods*, **7**, 1009 EP –. Article.

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biology*, **15**(2), R29.

Li, Y. I., Knowles, D. A., Humphrey, J., Barbeira, A. N., Dickinson, S. P., Im, H. K., and Pritchard, J. K. (2018). Annotation-free quantification of rna splicing using leafcutter. *Nature Genetics*, **50**(1), 151–158.

Liu, R., Loraine, A. E., and Dickerson, J. A. (2014). Comparisons of computational methods for differential alternative splicing detection using rna-seq in plant systems. *BMC Bioinformatics*, **15**(1), 364.

Mancini, E., Sanchez, S. E., Romanowski, A., Schlaen, R. G., Sanchez-Lamas, M., Cerdan, P. D., and Yanovsky, M. J. (2016). Acute effects of light on alternative splicing in light-grown plants. *Photochemistry and photobiology*, **92**(1), 126–133.

McCarthy, D. J., Smyth, G. K., and Robinson, M. D. (2009). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139–140.

McCarthy, D. J., Y., C., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Research*, **40**(10).

Mehmood, A., Laiho, A., Venäläinen, M. S., McGlinchey, A. J., Wang, N., and Elo, L. L. (2019). Systematic evaluation of differential splicing tools for RNA-seq studies. *Briefings in Bioinformatics*. bbz126.

Nilsen, W. T. and Graveley, B. R. (2010). Expansion of the eukaryiotic proteome by alternative splicing. *Nature*, **463**(1).

Ozsolak, F. and Milos, P. M. (2010). Rna sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, **12**, 87 EP –. Review Article.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, **14**(4), 417–419.

Perez-Santángelo, S., Mancini, E., Francey, L. J., Schlaen, R. G., Chernomoretz, A., Hogenesch, J. B., and Yanovsky, M. J. (2014). Role for lsm genes in the regulation of circadian rhythms. *Proceedings of the National Academy of Sciences*, **111**(42), 15166–15171.

Reddy, A. S., Marquez, Y., Kalyna, M., and Barta, A. (2013). Complexity of the alternative splicing landscape in plants. *The Plant Cell*, **25**(10), 3657–3683.

Ren, S., Peng, Z., Mao, J.-H., Yu, Y., Yin, C., Gao, X., Cui, Z., Zhang, J., Yi, K., Xu, W., Chen, C., Wang, F., Guo, X., Lu, J., Yang, J., Wei, M., Tian, Z., Guan, Y., Tang, L., Xu, C., Wang, L., Gao, X., Tian, W., Wang, J., Yang, H., Wang, J., and Sun, Y. (2012). Rna-seq analysis of prostate cancer in the chinese population identifies recurrent gene fusions, cancer-associated long noncoding rnas and aberrant alternative splicings. *Cell Research*, **22**(5), 806–821.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1).

Romanowski, A., Schlaen, R. G., Perez-Santangelo, S., Mancini, E., and Yanovsky, M. J. (2019). Global transcriptome analysis reveals circadian control of splicing events in arabidopsis thaliana. *bioRxiv*, page 845560.

Roworth, A. P., Carr, S. M., Liu, G., Barczak, W., Miller, R. L., Munro, S., Kanapin, A., Samsonova, A., and La Thangue, N. B. (2019). Arginine methylation expands the regulatory mechanisms and extends the genomic landscape under e2f control. *Science advances*, **5**(6), eaaw4640–eaaw4640. 31249870[pmid].

Rugnone, M. L., Soverna, A. F., Sanchez, S. E., Schlaen, R. G., Hernando, C. E., Seymour, D. K., Mancini, E., Chernomoretz, A., Weigel, D., Más, P., *et al.* (2013). Lnk genes integrate light and clock signaling networks at the core of the arabidopsis oscillator. *Proceedings of the National Academy of Sciences*, **110**(29), 12120–12125.

Sanchez, S. E., Petrillo, E., Beckwith, E. J., Zhang, X., Rugnone, M. L., Hernando, C. E., Cuevas, J. C., Godoy Herz, M. A., Depetris-Chauvin, A., Simpson, C. G., Brown, J. W. S., Cerdán, P. D., Borevitz, J. O., Mas, P., Ceriani, M. F., Kornblihtt, A. R., and Yanovsky, M. J. (2010). A methyl transferase links the circadian clock to the regulation of alternative splicing. *Nature*, **468**(7320), 112–116.

Schafer, S., Miao, K., Benson, C. C., Heinig, M., Cook, S. A., and Hubner, N. (2015). Alternative splicing signatures in rna-seq data: Percent spliced in (psi). *Current Protocols in Human Genetics*, **87**(1), 11.16.1–11.16.14.

Schlaen, R. G., Mancini, E., Sanchez, S. E., Perez-Santángelo, S., Rugnone, M. L., Simpson, C. G., Brown, J. W., Zhang, X., Chernomoretz, A., and Yanovsky, M. J. (2015). The spliceosome assembly factor gemin2 attenuates the effects of temperature on alternative splicing and circadian rhythms. *Proceedings of the National Academy of Sciences*, **112**(30), 9382–9387.

Shen, S., Park, J. W., Lu, Z.-x., 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. *Proceedings of the National Academy of Sciences*, **111**(51), E5593–E5601.

Simpson, C. G., Fuller, J., Maronova, M., Kalyna, M., Davidson, D., McNicol, J., Barta, A., and Brown, J. W. S. (2008). Monitoring changes in alternative precursor messenger rna splicing in multiple gene transcripts. *The Plant Journal*, **53**(6), 1035–1048.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012). Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature Protocols*, **7**, 562 EP –.

Vaquero-Garcia, J., Barrera, A., Gazzara, M. R., Gonzalez-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.

Vaquero-Garcia, J., Norton, S., and Barash, Y. (2018). Leafcutter vs. majiq and comparing software in the fast moving field of genomics. *bioRxiv*.

Wang, M., Zhao, Y., and Zhang, B. (2015). Efficient test and visualization of multi-set intersections. *Scientific Reports*, **5**(1), 16923.

Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**, 57 EP –. Perspective.

Xin, R., Zhu, L., Salomé, P. A., Mancini, E., Marshall, C. M., Harmon, F. G., Yanovsky, M. J., Weigel, D., and Huq, E. (2017). Spf45-related splicing factor for phytochrome signaling promotes photomorphogenesis by regulating pre-mrna splicing in arabidopsis. *Proceedings of the National Academy of Sciences*, **114**(33), E7018–E7027.

Zhang, C., Zhang, B., Lin, L.-L., and Zhao, S. (2017). Evaluation and comparison of computational tools for rna-seq isoform quantification. *BMC Genomics*, **18**(1), 583.

Supplementary Material

Aspli Integrative analysis of splicing landscapes through RNA-Seq assays

Mancini, Estefania[1,†], Rabinovich, Andres[2,3,†], Iserte, Javier[2,3], Yanovsky, Marcelo[2,3*], and Chernomoretz, Ariel[2,4*]

[1]CRG, Barcelona, Spain
[2]Fundacion Instituto Leloir, Buenos Aires, Argentina
[3]Instituto de Investigaciones Bioquímicas de Buenos Aires, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina
[4]Departamento de Fisica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Instituto de Fisica de Buenos Aires, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina
[†]Equally contributed

1 Feature counting in ASpli

1.1 Genomic feature extraction: binGenome()

Sub-genic features are analyzed using user-provided annotation files. Exon and intron coordinates are extracted from annotation for multi-exonic genes. When more than one isoform exists, some exons and introns from different isoforms will generally overlap. In the same spirit of (Anders et al., 2012), exons and introns are then subdivided into non-overlapping sub-genic features dubbed bins, defined by the boundaries of different exons across transcript variants. In this way, these so defined *bins* are maximal sub-genic features entirely included or entirely excluded from any mature transcript.

Bins are flagged as: exonic (E), intronic (I) or alternative-splicing (AS) bins, depending on the exonic/intronic character of the bin across variants . In addition, original intronic (Io) bins are defined for every intronic region of annotated isoforms (see panel A of Figure SF.1).

As a general rule, the extreme portions of a transcript probed by RNAseq assays show a highly non-uniform coverage that might obscure differential usage analysis. ASpli flags bins that overlap with the beginning or ending of any transcript as *external*. An external bin of a transcript may overlap with a non-external one of another transcript. Whenever this happens the bin is still labelled as external. Additionally, in order to avoid confounding effects in the analysis of splicing events, ASpli identifies and flags loci where more than one gene is present in the genome.

Local splicing classification model Each AS bin is further classified considering a three-bin *minimum local gene model*, that assigns splicing-event categories to a given bin based on the intronic/exonic character of the analyzed bin and its first neighbors (Figure SF.1, panel B).

For genes presenting two isoforms, this model is able to unambiguously assign a well defined splicing event category to the analyzed bin (see first row of Figure SF.1, panel B). An exon-skipping event (ES) should involve a sequence of intronic-exonic-intronic bins in one isoform and an intronic bin spanning the same region in the other one. Intron retention (IR) events correspond to a combination of exonic-intronic-exonic bins in one isoform and an exonic bin spanning the same region in the other isoform. Alternative five prime splicing site (Alt5SS) should present a combination of exonic-exonic-intronic bins in one isoform and exonic-intronic-intronic bins in the same region of the other isoform. Finally, alternative three prime splicing site (Alt3SS) events correspond to intronic-intronic-exonic bins in one isoform and intronic-exonic-exonic bins in the same region of the other isoform.

When more than two isoforms are present, we still found it useful to use the three-bin local model to segment follow up analysis. For these cases (see rows 2-4 of Figure SF.1, panel B) ASpli identify splicing events that involve: intronic subgenic regions surrounded by exons in at least one isoform (bin labelled as IR*), exonic subgenic regions surrounded by two introns in at least one isoform (bin labelled as ES*), exonic regions surrounded by intronic and exonic neighbor bins (bin labelled as Alt5SS* or Alt3SS*).
When it is not possible to get a clear splicing-type assignation (last row of Figure SF.1, panel B), bins are labeled as undefined AS (UAS).

As a last step of the genomic feature extraction process, annotated junctions from all the transcripts are also identified. Junction coordinates are defined as the last position of the five prime exon (donor position) and the first position of the three prime exon (acceptor position).

1.2 Annotation based feature counting: gbCounts()

Reads are overlaid on features derived from annotation, and count tables are produced at different genomic levels: genes, bins, and intron flanking regions used to identify and quantify intron retention events. Reads corresponding to annotated junctions are also tallied, along with genomic relevant information such as identity of spanned bins, and the existence of possible *exintronic* events.

1.3 De-novo junction counting: jCounts()

ASpli takes advantage of experimentally detected splice junctions to perform two different type of analysis. For one hand, junction data is considered in order to provide junction support to AS events detected through bin coverage analysis. For the other, it is used to quantify novel splicing events.

Junction support of bin coverage statistics: ASpli makes use of junction data as supporting evidence of alternative usage of bins. For a general differential splicing event affecting a given bin, it is always possible to define exclusion and inclusion junctions. The first class of junctions (noted as J_3) pass over the bin of interest, whereas the second ones (note as J_1 and/or J_2) quantify and support the inclusion of start and/or end bin boundaries in the mature transcript. Panel C of Figure SF.1 illustrates this point for the different types of splicing events that could affect a given bin. ASpli considers for this analysis junctions that are completely included within a unique gene and have more than a minimum number of reads supporting them (by default this number is five).

PSI (percent spliced in) Schafer et al. (2015) and PIR (percent of intron retention) metrics are two well known statistics that can be used to quantify the relative weight of inclusion evidence for different kind of splicing events (see Panel C of Figure SF.1). For each bin, ASpli quantifies the inclusion strength in every experimental condition using the appropriate inclusion index (see Table ST.1). Only junctions that pass an abundance filter criterium (a minimum number of counts should be attained in all samples of at least one condition) are considered for the estimations.

For each bin, a PIR or a PSI metric is calculated, according to the splicing event category assigned to that bin (see last column of table ST.1). If no splice event was assigned, meaning that the bin is not alternative, an exon will be considered to be involved in a putative exon skipping splicing event, and an intron will be considered to be involved in a putative intron retention splicing event.

Novel and non-canonical splicing patterns: ASpli relies on the direct analysis of experimentally observed splicing junctions in order to study novel (i.e. non-annotated) splicing patterns. For every experimental junction, ASpli characterizes local splicing patterns considering two hypothetical scenarios. For one hand, assuming that every detected junction might be associated to a possible intron that could be potentially retained, a PIR_{junc} value is computed (panel D of Figure SF.1).

On the other hand, every junction also defines potential 5' and 3' splicing sites. It can be the case that one (in an alternative 5' or 3' scenario), or both ends (in case of exon skipping) were shared by other junctions. In this context, it is informative to characterize the relative abundance of the analyzed junction (dubbed J_3) with respect to the locally *competing* ones. ASpli estimates *percentage junction-usage* indices, PJU_{J_1} and PJU_{J_2}, in order to evaluate and quantify this quantities (see Panel D of figure SF.1 and Table ST.1). In order to illustrate this point, we show in Panel E of figure SF.1 an hypothetical splicing scenario for a given junction of interest, J_3. It can be appreciated that PJU_{J_1} quantifies the participation of this junction in the context of a splicing pattern involving the two orange competing junctions, whereas PJU_{J_2} reports on the usage of J_3 in connection with the green competing junction.

2 Command-line running arguments

Command lines used to invoked algorithms and further calculation details:

- STAR aligner
  For PRMT5 datasets

```
$ STAR --runThreadN 30 --genomeDir TAIR10_GENOME_DIR --
    ↪ twopassMode Basic --outSAMtype BAM
    ↪ SortedByCoordinate --outFilterMultimapNmax 2 --
    ↪ outFilterType BySJout --outSJfilterReads Unique --
    ↪ sjdbOverhang PARAM --alignSJoverhangMin 6 --
    ↪ alignSJDBoverhangMin 3 --alignIntronMin 20 --
    ↪ alignIntronMax 5000 --readFilesIn ../01_FASTQ/
    ↪ Col_3_1.fq ../01_FASTQ/Col_3_2.fq --
    ↪ outFileNamePrefix Col_3/Col_3
```

  We used a sjdbOverhang parameter value equal to 99 and 149 for PRMT5 datasets A and B respectively.

  For the prostate dataset we aligned using default STAR parameters.

```
$ STAR --runThreadN 30 --genomeDir ENSEMBL_HG38_PATH --
    ↪ readFilesCommand zcat --twopassMode Basic --
    ↪ outSAMtype BAM SortedByCoordinate --sjdbOverhang 89
    ↪ --readFilesIn 1.fq 2.fq
```

  We used a sjdbOverhang parameter value equal to 99 and 149 for PRMT5 datasets A and B respectively.

- LeafCutter (synthetic dataset)
  BAM files were first processed using the provided *bam2junc.sh* script. The generated *juncfiles.txt* was then used to build junction clusters via the provided python script

```
$ python PATH leafcutter_cluster.py -j juncfiles.txt -m 30
    ↪ -l 500000
```

Finally, we used the provided *leafcutter_ds* R-script to run the statistical analysis (min_samples_per_intron=3).

- rMATS Command line used to analyze PRMT5 assays:

```
rMATS.4.0.2/rMATS-turbo-Linux-UCS4/rmats.py --b1 bam_prmt5.
    ↪ txt --b2 bam_col.txt --gtf /data1/genomeData/ath/
    ↪ Ensembl_illumina_iGenomes/TAIR10/Annotation/Genes/
    ↪ genes.gtf --od rl150 -t paired --nthread 20 --
    ↪ readLength 150 --tstat 10
```

- MAJIQ Script used to analyze PRMT5 assays:

```
#builder
majiq build /data1/genomeData/ath/Ensembl/TAIR10_20190827/
    ↪ genes.gff3 -c majiq.config -j 25 -o 01_build

#psi calculation
majiq psi 01_build/Col_1.star.majiq 01_build/Col_3.star.
    ↪ majiq 01_build/Col_4.star.majiq -j 28 -o 02_psi -n
    ↪ col0
majiq psi 01_build/prmt5_Col_12.star.majiq 01_build/
    ↪ prmt5_Col_11.star.majiq 01_build/prmt5_Col_9.star.
    ↪ majiq -j 28 -o 02_psi -n prmt5

#delta_psi
majiq deltapsi -grp1 01_build/Col_1.star.majiq 01_build/
    ↪ Col_3.star.majiq 01_build/Col_4.star.majiq -grp2 01
    ↪ _build/prmt5_Col_11.star.majiq 01_build/prmt5_Col_12.
    ↪ star.majiq 01_build/prmt5_Col_9.star.majiq -j 8 -o 04
    ↪ _dpsi -n col0 prmt5
```

3 Splicing affected regions detected by different algorithms

Each algorithm reports splicing altered genomic features in different ways. In order to standardize the identification of regions of interest we proceeded as follows:

- LeafCutter: We first identified clusters presenting adjusted pvalues< 0.05 as reported in 'leafcutter_ds_cluster_significance.txt' file. For each of these statistically significant clusters we considered the associated genomic-regions reported in 'leafcutter_ds_effect_size.txt' file with $|\Delta\Psi| > 0.1$.

- MAJIQ: We considered the genomic-region covering junction clusters presenting at least one junction with $P(|\Delta\Psi| > 0.2) > 0.95$.

- rMATS: We considered the values reported in 'JCEC.txt' files. This means that we considered a model that evaluated splicing with reads that spanned splicing junctions and reads on targets bins (i.e. alternatively spliced exons). We kept junctions presenting adjusted FDR< 0.05 and inclusion signal larger than a 0.1 level. Genomic regions were then defined according the following rules:
  - A3SS' (A3SS.MATS.JCEC.txt file): We considered the genomic region between 'shortEE' and 'longExonEnd' coordinates for negative strand and by 'longExonStart_0base' and 'shortES' for positive strand cases.
  - A5SS' (A5SS.MATS.JCEC.txt file): We considered the genomic region between 'shortEE' and 'longExonEnd' coordinates for positive strand and by 'longExonStart_0base' and 'shortES' for negative strand cases.
  - MXE (MXE.MATS.JCEC.txt file): We considered two regions per event defined by: '1stExonStart_0base', '1stExonEnd' and '2ndExonStart_0base', '2ndExonEnd'.
  - SE (SE.MATS.JCEC.txt file): We considered the regions between 'exonStart_0base' and 'exonEnd'.
  - RI (RI.MATS.JCEC.txt file): We considered the regions between 'riExonStart_0base' and 'riExonEnd'.

4 Bin-coverage detection calls in the synthetic dataset

We decided to further characterized some aspects of bin-coverage detection calls for the synthetic dataset, as this signal provided the major number of discoveries. It can be seen in panel-(A) of Figure SF.2 that fold-change and junction-support signals used in the bin-coverage analysis reported relevant and non-redundant information. Whereas the first one accounted for 37% of true positive instances exclusively detected by this signal, the second one accounted for the specific identification of 12% of the total number of true events. The impact of the selected fold-change threshold value, FC*, on specificity, precision and recall can be appreciated with the aid of the Receiver-Operator and Precision-Recall curves shown in panels (B) and (C) of Figure SF.2. It can be recognized from these figures that with the adopted 3-fold threshold ASpli achieved high recall and precision levels ($\sim$ 80% and $\sim$ 95% respectively) laying at rather moderate levels of false positive rates ($\sim$ 14%).

5 Analysis of false positive calls in the simulated dataset

In this section we analyzed the origin of ASpli detected false positive events.

Simulated changes in isoforms concentration for a given gene produced specific patterns of bin and junction differential usage that depended on the exonic architecture of the gene-isoforms. For each gene for which a change in isoforms concentration was simulated, it was possible to anticipate which bins would present differential usage (we call them *active bins*). As in our simulations we admitted a 20% level of random variability in isoform concentration profiles, some background differential bin usage could also take place for bins belonging to genes for which alternative splicing was not explicitly simulated (we called them *inactive-bins*).

In order to quantify this effect we introduced the *a-priori* Splicing Activation Signal (SAS) value. For each gene, SAS was estimated as the maximum absolute change in isoform concentration actually simulated between conditions. We found that background isoform concentration variability produced non-zero SAS levels for inactive events, but this noise-originated signal remained well bellow the signal reported for active ones. The left-most first and second boxplots in Figure SF.3 depict the distribution of this quantity for the 915 genes for which a splicing event was simulated (*active*), and for the remaining 7518 genes (*inactive*) respectively . On the other hand, the four right-most boxplots show the SAS distribution for false positive calls obtained with different methods. Non explicitely splicing simulated changes were reported for $9, 4, 48$ and 23 genes according to ASpli, LeafCutter, MAJIQ and rMATS algorithms respectively.

6 Comparison of discoveries

A comprehensive comparison of discoveries appeared at first-sight problematic as each algorithm is focused on different genomic features in order to chart splicing landscapes.

For instance, rMATS analyzes genomic regions flanked by upstream and downstream exons to examine canonical splicing events. MAJIQ and LeafCutter, on the other hand, exclusively rely on clusters of split reads that share start or ending junction-ends. Finally ASpli considers both, junction clusters and bin features, i.e. genomic regions defined from disjoint ranges of annotated junctions.

In this context, a first coarse grained comparison could be established at gene-level, comparing the identity of genes housing splicing-altered patterns according to the different analyzed methods. Panel (A) of Figure SF.4 displays a color-coded overlap matrix of affected genes in experiments A and B according to the four examined methodologies. Each cell reports the intersection size and, in brackets, the corresponding overlap coefficient. At gene level, rMATS achieved the largest agreement factor (83% of genes identified in experiment B, were also reported in experiment A). However, it also produced the lowest number of discoveries (119). ASpli, on the other hand, presented a comparable level of agreement (71%), highlighting a significatively larger number of concordant genes (2109). Typically, more than 50% of genes identified by any methodology was also reported by ASpli (first and second rows of Figure SF.4). Moreover, the number of concordant discoveries between experiments considering a given methodology was comparable to the agreement level achieved between each experiment-metodology combination and the correpsonding ASpli result. Noticeably, more than 90% of MAJIQ's genes were also spotted by ASpli.

A more in-depth comparison could be established analyzing the overlap of identified genomic regions. In panels (b) and (c) of Figure SF.4 we informed the extent of the overlaps between genomic regions found to be affected by differential splicing patterns according to each algorithm (see Material and Methods **??**) to map events reported by each method to a common set of genomic coordinates). While any kind of overlap was registered for panel (b), only complete inclusion of genomic regions identified by one method inside the ones identified by a second one was considered for panel (c). Statistically significant overlaps were marked with asterisks. Note that overlap coefficients (in brackets) exceeding unity were detected in between-experiments comparisons for LeafCutter and rMATS as a result of the presence of one-to-many region mappings.

For the loose overlap criterium we found statistically significant concordance between discoveries for almost every cell (Fig SF.4-b). Only specific comparisons involving MAJIQ and rMATs failed the statistical significance test. At the same time, overlap coefficient values were similar to the ones estimated at the gene-level analysis. Noticeably, we recognised a sensible reduction in this quantity for the MAJIQ vs ASpli comparison. This finding highlighted that gene-level agreement should in general be considered with caution. A more detailed examination at the sub-genic level might be necessary to assess for discovery consistencies between algorithms. Results for the most stringent overlap criterion are shown in Figure SF.4(c). As expected, a major decrease on overlap coefficient values was observed . However, statistically significant agreement between results was still found as a general rule. Only comparisons involving MAJIQ's discoveries failed the statistical assessments.

7 Data consolidation

We took advantage of ASpli capabilities to deal with complex experimental designs to consolidate datasets A and B in a statistically sound way. We considered the following generalized linear model:

$$y \sim experiment + genotype + experiment : genotype \qquad \text{(SE.1)}$$

'experiment' was a fixed effect to cope with specific technical biases, and the 'genotype' factor was meant to capture the PRMT5 vs wild-type effect. The third term was an interaction term, and was used to enforced the exclusion of non-coherent signals between experiments. ASpli detected 4360 genomic regions displaying strong evidence of a genotype effect (fdr < 0.05). In addition, 99% of these PRMT5-related events (4314 out of 4360) passed a filtering step to enforce they presented no-detectable evidence of experiment-genotype interactions (experiment:genotype associated fdr > 0.5). These 4314 events defined the consolidated AB data set.

We found that 99% (2209 out of 2241) of the concordant discoveries independently detected in both assays were also included in the consolidated dataset (we included a Venn diagram of the discoveries reported for experiments A, B, and the consolidated data-set AB in Sup.Fig. SF.6). Noticeably, the consideration of the AB data-set allowed to almost double the number of detected genomic regions displaying robust evidence of differential splicing patterns.

8 PRMT5 PCR events

We characterized the agreement between the 23 splicing events that ASpli uncovered for the consolidated AB case, and the 44 Sanchez qRT-PCR validated events in Table ST.2. For each assayed event we included the kind of the original event and the reported qRT-PCR splicing signal value in the second and third columns respectively (Sanchez and collaborators calculated the fraction of the shortest isoform in PRMT5 mutants and wildtype plants detected by qRT-PCR, and used the relativized difference between them as a quantitative proxy of splicing changes (Table 4 of Sanchez et al. (2010))). In the fourth column we informed whether the PCR-interrogated genomic region overlapped with the one signaled by ASpli. Finally, the type of splicing event detected by ASpli was included in the last column of the table.

9 Prostate cancer dataset:Transcriptomic variability

In order to visualize the transcriptomic variability across patients at gene expression levels we considered the 30% most variable genes across the 28 expression profiles that presented more than 10 counts per million reads in at least 3 samples. With this informative set of 1386 genes we built a multidimensional scaling plot of distances between gene expression profiles estimated with the edgeR package Robinson et al. (2010). Results are shown in Fig SF.5 (reported results were very robust against changes in the number of top-variable genes chosen to characterize the transcriptome of each sample). In this kind of plot, samples lay on a two-dimensional scatterplot so that distances on the plot approximate the typical log2 fold changes between the samples (function plotMDS of edgeR Robinson et al. (2010)).

Emtpy and filled symbol correspond to tumor and normal tissue samples respectively. Pair of points of a given patient are equally colored and joined by a dashed edge. It can be seen that tumor and normal samples were well separated across the leading reduced dimension. The second largest projected dimension, however, let us appreciate internal structure and some variability between patients. There was a group of 5 patients (top left empty points) that displayed a rather homogeneous pattern of changes between tumor affected and normal tissues. On the contrary, the 9 bottom-left tumor samples seemed to segregate into a different cluster of transcriptomes. Moreover, the corresponding patients presented different kinds of alterations between tumor and control samples.

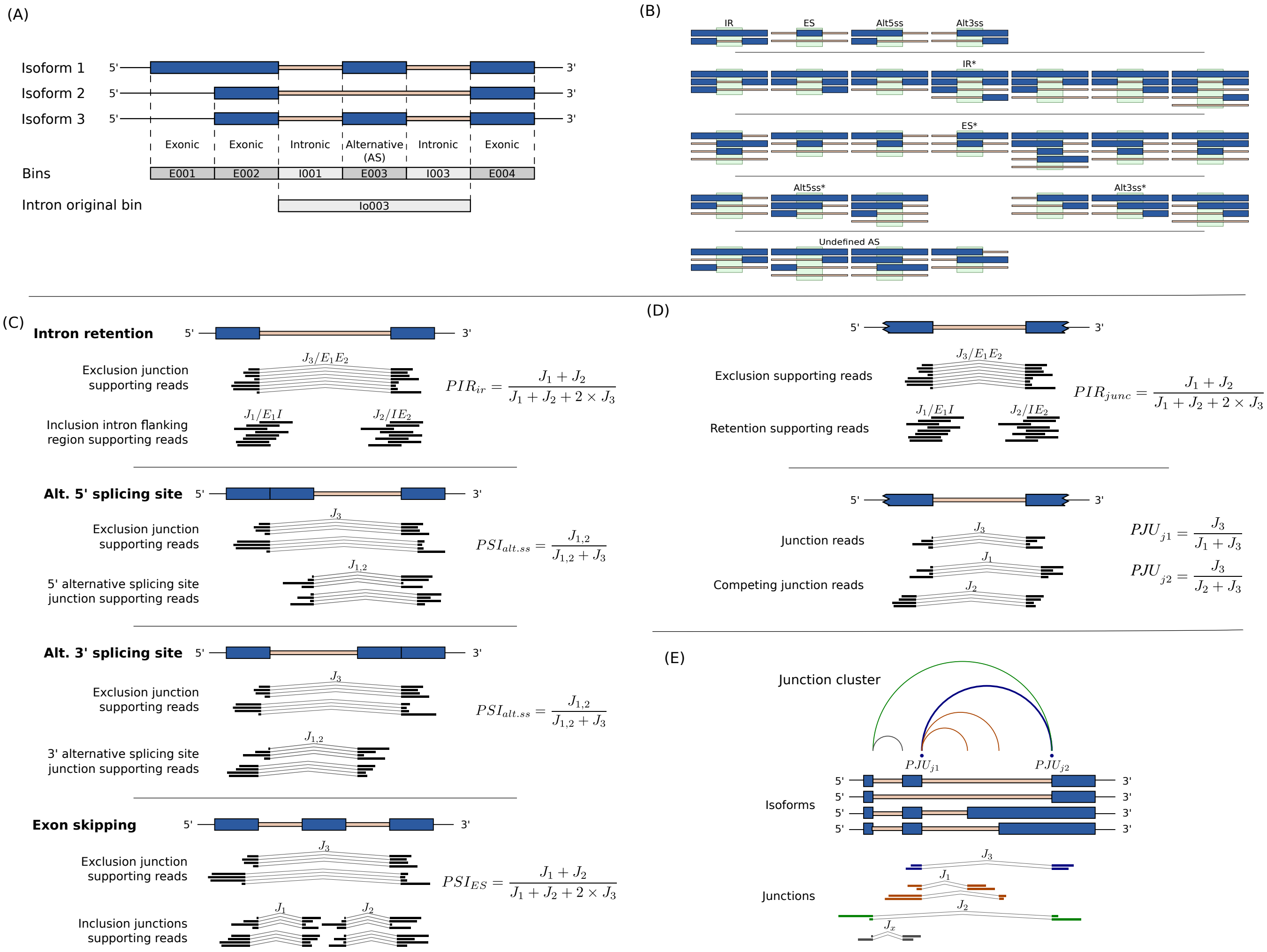

(A)
Isoform 1
Isoform 2
Isoform 3
5'
3'
Exonic
Exonic
Intronic
Alternative (AS)
Intronic
Exonic
Bins
E001
E002
I001
E003
I003
E004
Intron original bin
Io003
(B)
IR
ES
Alt5ss
Alt3ss
IR*
ES*
Alt5ss*
Alt3ss*
Undefined AS
(C)
Intron retention
Exclusion junction supporting reads
J_3/E_1E_2
$PIR_{ir} = \frac{J_1 + J_2}{J_1 + J_2 + 2 \times J_3}$
Inclusion intron flanking region supporting reads
J_1/E_1I
J_2/IE_2
Alt. 5' splicing site
Exclusion junction supporting reads
J_3
$PSI_{alt.ss} = \frac{J_{1,2}}{J_{1,2} + J_3}$
5' alternative splicing site junction supporting reads
$J_{1,2}$
Alt. 3' splicing site
Exclusion junction supporting reads
J_3
$PSI_{alt.ss} = \frac{J_{1,2}}{J_{1,2} + J_3}$
3' alternative splicing site junction supporting reads
$J_{1,2}$
Exon skipping
Exclusion junction supporting reads
J_3
$PSI_{ES} = \frac{J_1 + J_2}{J_1 + J_2 + 2 \times J_3}$
Inclusion junctions supporting reads
J_1
J_2
(D)
Exclusion supporting reads
J_3/E_1E_2
$PIR_{junc} = \frac{J_1 + J_2}{J_1 + J_2 + 2 \times J_3}$
Retention supporting reads
J_1/E_1I
J_2/IE_2
Junction reads
J_3
$PJU_{j1} = \frac{J_3}{J_1 + J_3}$
Competing junction reads
J_1
J_2
$PJU_{j2} = \frac{J_3}{J_2 + J_3}$
(E)
Junction cluster
PJU_{j1}
PJU_{j2}
Isoforms
Junctions
J_3
J_1
J_2
J_x

Figure (SF.1) Panel (A) shows how bin-features are defined and classified as: external, exonic, intronic or intron_original bins using genome annotation. The local splicing classification scheme is illustrated in panel (B). The definition of PSI and PIR metrics for bin features are pictured in panel (C). Definition of junction PIR and PJU statistics are shown in panel (D). Panel (E) shows a possible junction cluster and highlights the definition of type J_1, J_2 and J_3 junctions for the analysis of PJU statistics for the blue junction.

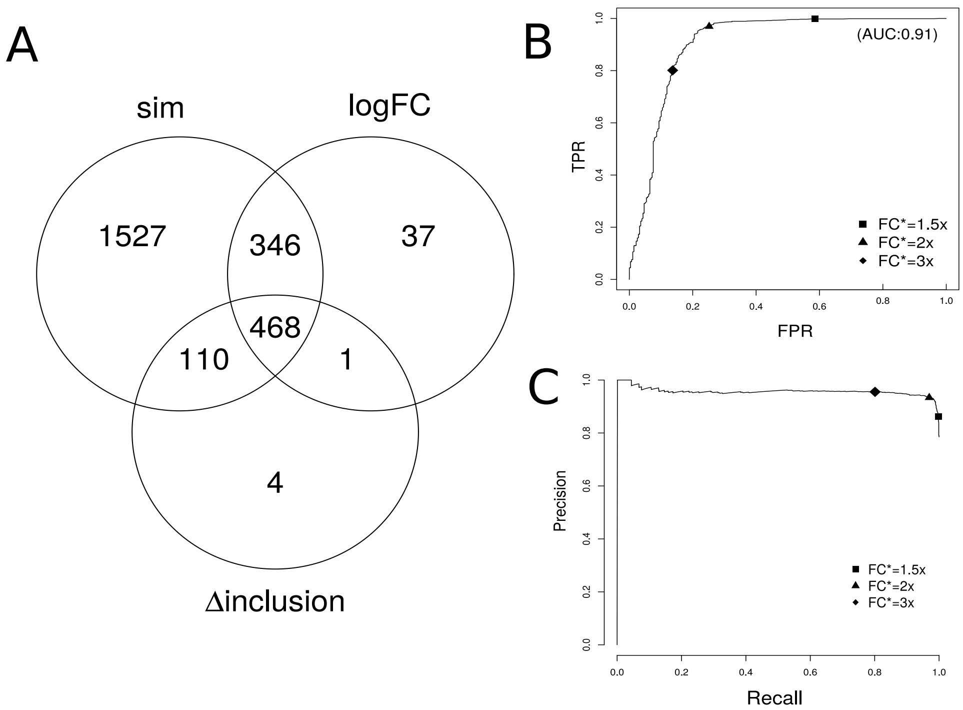


Figure (SF.2) (A) Graphical summary of bin-coverage detection calls. The *sim* set correspond to simulated events. *logFC* and *D-inclusion* sets are associated to statistically significant discoveries presenting large enough fold change and large bin-supporting junction inclusion signals respectively. ROC and Precision-Recall curves, parameterized by the considered fold-change threshold level, are shown for statistically significant bins in panels (B) and (C) respectively.

feature	assesment	index		bin class
bin	inclusion	PIR_{ir}	$\frac{J_1+J_2}{J_1+J_2+2*J_3}$	UAS, I, I*, I_0
		PSI_{es}		UAS, E, E*
		PSI_{alt5ss}	$\frac{J_{1,2}}{J_{1,2}+J_3}$	Alt5ss, Alt5ss*
		PSI_{alt3ss}		Alt3ss, Alt3ss*
junction		PIR_{junc}	$\frac{J_1+J_2}{J_1+J_2+2*J_3}$	-
	usage	PJU_{J_1}	$\frac{J_3}{J_1+J_3}$	-
		PJU_{J_2}	$\frac{J_3}{J_2+J_3}$	-

Table (ST.1) Junction usage and inclusion strength figure of merits for different bin classes and for experimentally detected junctions. The definition of J_1, J_2 and J_3 junction counts is depicted in panels C and D of Figure SF.1 for annotated and experimentally detected junctions respectively.

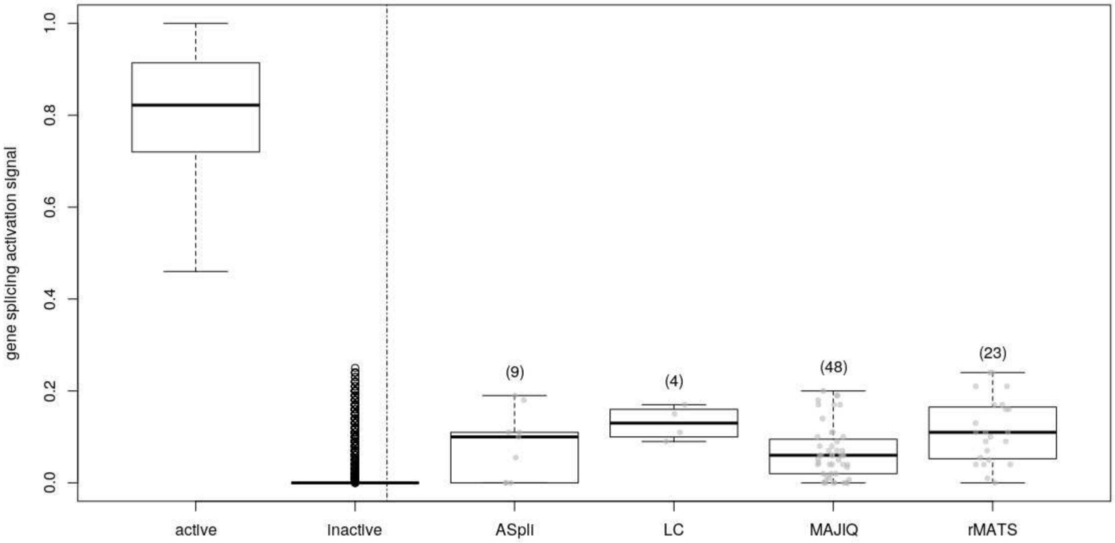


Figure (SF.3) Distribution of Splicing Activation Signals for bins in simulated (*active*) and non-simulated (*inactive*) splicing events are displayed in the first two boxplots. The four right-most boxplot show the SAS distributions for false positive events reported for each algorithm

	Gene ID	Event	qRT-PCR signal	Region overlap	Detected event
1	AT1G53650	5'ss	18.93	yes	IR (next to 5)
2	AT1G54360	5'ss	39.21	yes	Alt5ss
3	AT1G76510	5'ss	-28.00	yes	Alt5ss
4	AT2G04790	5'ss	11.14	yes	IR
5	AT2G15530	5'ss	21.12	yes	Alt 5'/3'
6	AT2G33480	5'ss	-27.16	yes	IR
7	AT2G38880	5'ss	-10.44	no	IR
8	AT2G46790	5'ss	35.20	yes	Alt5ss (plus additional IR)
9	AT3G01150	ES	-13.70	no	IR
10	AT3G12250	5'ss	-16.29	no	ES*
11	AT3G16800	IR/3'ss	-31.59	yes	IR,Novel Alt 5'/3'
12	AT3G19840	5'ss	-26.20	no	IR
13	AT3G20270	ES	8.51	no	IR (next to ES)
14	AT3G23280	ES	18.21	yes	ES
15	AT3G25840	ES	6.51	yes	ES
16	AT4G02430	3'ss	27.45	no	IR
17	AT4G24740	ES	16.07	no	IR (next to ES)
18	AT4G31720	3'ss	12.37	no	IR
19	AT4G32730	5'ss	30.93	yes	Novel Alt 5'/3'
20	AT4G38510	5'ss	15.53	yes	Alt5ss*,CSP
21	AT5G05550	ES	32.72	yes	ES (plus additional IR)
22	AT5G25610	IR	-71.69	yes	IR
23	AT5G57630	5'ss	31.07	yes	Novel Alt 5'/3' (plus adjacent IR)

Table (ST.2) Agreement between ASpli and qRT-PCR results

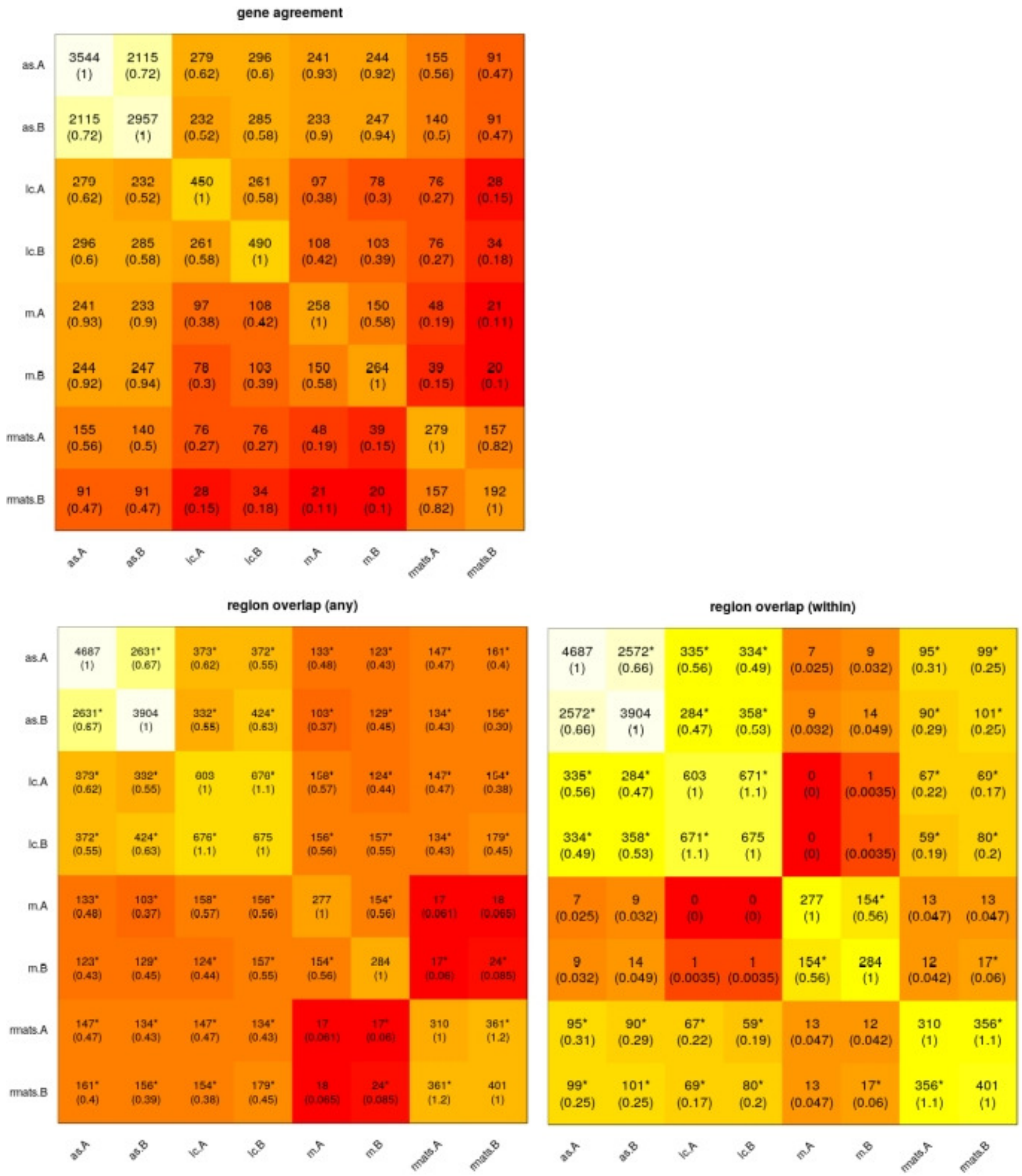


Figure (SF.4) Number of concordant discoveries produced by different methods. Gene-level, region-overlap and strong-region-overlap results are displayed in the top-left, bottom-left and bottom-right panels respectively. Overlap coefficients are included between brackets. See text for details.

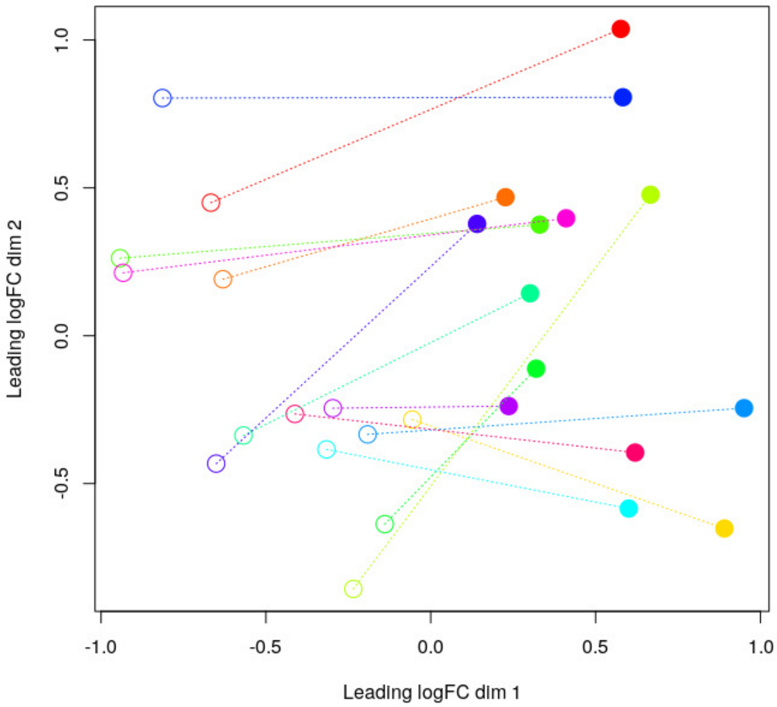


Figure (SF.5) Multidimensional scaling plot of distances between gene expression profiles. Normal and tumor samples are depicted using empty and filled circles respectively. Paired-samples were uniformily colored and connected with a dotted line.

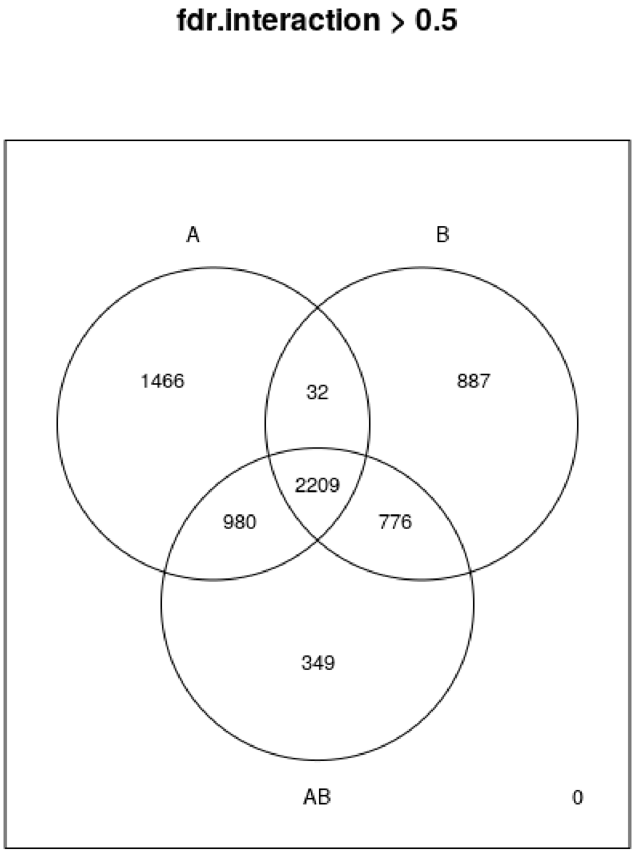


Figure (SF.6) Venn diagram of alternative splicing events detected in experiments A, B, and the consolidated data set AB (i.e. events displaying strong evidence of a genotype effect (fdr< 0.05) and no-detectable evidence of experiment-genotype interaction (experiment:genotype associated fdr> 0.5)).

10 Bibliography

References

S. Anders, A. Reyes, W. Huber, Detecting differential usage of exons from rna-seq data, Genome Research 22 (2012) 2008–2017.

S. Schafer, K. Miao, C. C. Benson, M. Heinig, S. A. Cook, N. Hubner, Alternative splicing signatures in rna-seq data: Percent spliced in (psi), Current Protocols in Human Genetics 87 (2015) 11.16.1–11.16.14.

S. E. Sanchez, E. Petrillo, E. J. Beckwith, X. Zhang, M. L. Rugnone, C. E. Hernando, J. C. Cuevas, M. A. Godoy Herz, A. Depetris-Chauvin, C. G. Simpson, J. W. S. Brown, P. D. Cerdán, J. O. Borevitz, P. Mas, M. F. Ceriani, A. R. Kornblihtt, M. J. Yanovsky, A methyl transferase links the circadian clock to the regulation of alternative splicing, Nature 468 (2010) 112–116.

M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics 26 (2010).

Supplementary Material

Aspli Integrative analysis of splicing landscapes through RNA-Seq assays

Mancini, Estefania[1,†], Rabinovich, Andres[2,3,†], Iserte, Javier[2,3], Yanovsky, Marcelo[2,3*], and Chernomoretz, Ariel[2,4*]

[1]CRG, Barcelona, Spain
[2]Fundacion Instituto Leloir, Buenos Aires, Argentina
[3]Instituto de Investigaciones Bioquímicas de Buenos Aires, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina
[4]Departamento de Fisica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Instituto de Fisica de Buenos Aires, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina
[†]Equally contributed

1 Feature counting in ASpli

1.1 Genomic feature extraction: binGenome()

Sub-genic features are analyzed using user-provided annotation files. Exon and intron coordinates are extracted from annotation for multi-exonic genes. When more than one isoform exists, some exons and introns from different isoforms will generally overlap. In the same spirit of (Anders et al., 2012), exons and introns are then subdivided into non-overlapping sub-genic features dubbed bins, defined by the boundaries of different exons across transcript variants. In this way, these so defined *bins* are maximal sub-genic features entirely included or entirely excluded from any mature transcript.

Bins are flagged as: exonic (E), intronic (I) or alternative-splicing (AS) bins, depending on the exonic/intronic character of the bin across variants . In addition, original intronic (Io) bins are defined for every intronic region of annotated isoforms (see panel A of Figure SF.1).

As a general rule, the extreme portions of a transcript probed by RNAseq assays show a highly non-uniform coverage that might obscure differential usage analysis. ASpli flags bins that overlap with the beginning or ending of any transcript as *external*. An external bin of a transcript may overlap with a non-external one of another transcript. Whenever this happens the bin is still labelled as external. Additionally, in order to avoid confounding effects in the analysis of splicing events, ASpli identifies and flags loci where more than one gene is present in the genome.

Local splicing classification model Each AS bin is further classified considering a three-bin *minimum local gene model*, that assigns splicing-event categories to a given bin based on the intronic/exonic character of the analyzed bin and its first neighbors (Figure SF.1, panel B).

For genes presenting two isoforms, this model is able to unambiguously assign a well defined splicing event category to the analyzed bin (see first row of Figure SF.1, panel B). An exon-skipping event (ES) should involve a sequence of intronic-exonic-intronic bins in one isoform and an intronic bin spanning the same region in the other one. Intron retention (IR) events correspond to a combination of exonic-intronic-exonic bins in one isoform and an exonic bin spanning the same region in the other isoform. Alternative five prime splicing site (Alt5SS) should present a combination of exonic-exonic-intronic bins in one isoform and exonic-intronic-intronic bins in the same region of the other isoform. Finally, alternative three prime splicing site (Alt3SS) events correspond to intronic-intronic-exonic bins in one isoform and intronic-exonic-exonic bins in the same region of the other isoform.

When more than two isoforms are present, we still found it useful to use the three-bin local model to segment follow up analysis. For these cases (see rows 2-4 of Figure SF.1, panel B) ASpli identify splicing events that involve: intronic subgenic regions surrounded by exons in at least one isoform (bin labelled as IR*), exonic subgenic regions surrounded by two introns in at least one isoform (bin labelled as ES*), exonic regions surrounded by intronic and exonic neighbor bins (bin labelled as Alt5SS* or Alt3SS*).
When it is not possible to get a clear splicing-type assignation (last row of Figure SF.1, panel B), bins are labeled as undefined AS (UAS).

As a last step of the genomic feature extraction process, annotated junctions from all the transcripts are also identified. Junction coordinates are defined as the last position of the five prime exon (donor position) and the first position of the three prime exon (acceptor position).

1.2 Annotation based feature counting: gbCounts()

Reads are overlaid on features derived from annotation, and count tables are produced at different genomic levels: genes, bins, and intron flanking regions used to identify and quantify intron retention events. Reads corresponding to annotated junctions are also tallied, along with genomic relevant information such as identity of spanned bins, and the existence of possible *exintronic* events.

1.3 De-novo junction counting: jCounts()

ASpli takes advantage of experimentally detected splice junctions to perform two different type of analysis. For one hand, junction data is considered in order to provide junction support to AS events detected through bin coverage analysis. For the other, it is used to quantify novel splicing events.

Junction support of bin coverage statistics: ASpli makes use of junction data as supporting evidence of alternative usage of bins. For a general differential splicing event affecting a given bin, it is always possible to define exclusion and inclusion junctions. The first class of junctions (noted as J_3) pass over the bin of interest, whereas the second ones (note as J_1 and/or J_2) quantify and support the inclusion of start and/or end bin boundaries in the mature transcript. Panel C of Figure SF.1 illustrates this point for the different types of splicing events that could affect a given bin. ASpli considers for this analysis junctions that are completely included within a unique gene and have more than a minimum number of reads supporting them (by default this number is five).

PSI (percent spliced in) Schafer et al. (2015) and PIR (percent of intron retention) metrics are two well known statistics that can be used to quantify the relative weight of inclusion evidence for different kind of splicing events (see Panel C of Figure SF.1). For each bin, ASpli quantifies the inclusion strength in every experimental condition using the appropriate inclusion index (see Table ST.1). Only junctions that pass an abundance filter criterium (a minimum number of counts should be attained in all samples of at least one condition) are considered for the estimations.

For each bin, a PIR or a PSI metric is calculated, according to the splicing event category assigned to that bin (see last column of table ST.1). If no splice event was assigned, meaning that the bin is not alternative, an exon will be considered to be involved in a putative exon skipping splicing event, and an intron will be considered to be involved in a putative intron retention splicing event.

Novel and non-canonical splicing patterns: ASpli relies on the direct analysis of experimentally observed splicing junctions in order to study novel (i.e. non-annotated) splicing patterns. For every experimental junction, ASpli characterizes local splicing patterns considering two hypothetical scenarios. For one hand, assuming that every detected junction might be associated to a possible intron that could be potentially retained, a PIR_{junc} value is computed (panel D of Figure SF.1).

On the other hand, every junction also defines potential 5' and 3' splicing sites. It can be the case that one (in an alternative 5' or 3' scenario), or both ends (in case of exon skipping) were shared by other junctions. In this context, it is informative to characterize the relative abundance of the analyzed junction (dubbed J_3) with respect to the locally *competing* ones. ASpli estimates *percentage junction-usage* indices, PJU_{J_1} and PJU_{J_2}, in order to evaluate and quantify this quantities (see Panel D of figure SF.1 and Table ST.1). In order to illustrate this point, we show in Panel E of figure SF.1 an hypothetical splicing scenario for a given junction of interest, J_3. It can be appreciated that PJU_{J_1} quantifies the participation of this junction in the context of a splicing pattern involving the two orange competing junctions, whereas PJU_{J_2} reports on the usage of J_3 in connection with the green competing junction.

2 Command-line running arguments

Command lines used to invoked algorithms and further calculation details:

- STAR aligner
  For PRMT5 datasets

  ```
  $ STAR --runThreadN 30 --genomeDir TAIR10_GENOME_DIR --
      ↪ twopassMode Basic --outSAMtype BAM
      ↪ SortedByCoordinate --outFilterMultimapNmax 2 --
      ↪ outFilterType BySJout --outSJfilterReads Unique --
      ↪ sjdbOverhang PARAM --alignSJoverhangMin 6 --
      ↪ alignSJDBoverhangMin 3 --alignIntronMin 20 --
      ↪ alignIntronMax 5000 --readFilesIn ../01_FASTQ/
      ↪ Col_3_1.fq ../01_FASTQ/Col_3_2.fq --
      ↪ outFileNamePrefix Col_3/Col_3
  ```

  We used a sjdbOverhang parameter value equal to 99 and 149 for PRMT5 datasets A and B respectively.

  For the prostate dataset we aligned using default STAR parameters.

  ```
  $ STAR --runThreadN 30 --genomeDir ENSEMBL_HG38_PATH --
      ↪ readFilesCommand zcat --twopassMode Basic --
      ↪ outSAMtype BAM SortedByCoordinate --sjdbOverhang 89
      ↪ --readFilesIn 1.fq 2.fq
  ```

  We used a sjdbOverhang parameter value equal to 99 and 149 for PRMT5 datasets A and B respectively.

- LeafCutter (synthetic dataset)
  BAM files were first processed using the provided *bam2junc.sh* script. The generated *juncfiles.txt* was then used to build junction clusters via the provided python script

  ```
  $ python PATH leafcutter_cluster.py -j juncfiles.txt -m 30
      ↪ -l 500000
  ```

Finally, we used the provided *leafcutter_ds* R-script to run the statistical analysis (min_samples_per_intron=3).

- rMATS Command line used to analyze PRMT5 assays:

```
rMATS.4.0.2/rMATS-turbo-Linux-UCS4/rmats.py --b1 bam_prmt5.
    ↪ txt --b2 bam_col.txt --gtf /data1/genomeData/ath/
    ↪ Ensembl_illumina_iGenomes/TAIR10/Annotation/Genes/
    ↪ genes.gtf --od rl150 -t paired --nthread 20 --
    ↪ readLength 150 --tstat 10
```

- MAJIQ Script used to analyze PRMT5 assays:

```
#builder
majiq build /data1/genomeData/ath/Ensembl/TAIR10_20190827/
    ↪ genes.gff3 -c majiq.config -j 25 -o 01_build

#psi calculation
majiq psi 01_build/Col_1.star.majiq 01_build/Col_3.star.
    ↪ majiq 01_build/Col_4.star.majiq -j 28 -o 02_psi -n
    ↪ col0
majiq psi 01_build/prmt5_Col_12.star.majiq 01_build/
    ↪ prmt5_Col_11.star.majiq 01_build/prmt5_Col_9.star.
    ↪ majiq -j 28 -o 02_psi -n prmt5

#delta_psi
majiq deltapsi -grp1 01_build/Col_1.star.majiq 01_build/
    ↪ Col_3.star.majiq 01_build/Col_4.star.majiq -grp2 01
    ↪ _build/prmt5_Col_11.star.majiq 01_build/prmt5_Col_12.
    ↪ star.majiq 01_build/prmt5_Col_9.star.majiq -j 8 -o 04
    ↪ _dpsi -n col0 prmt5
```

3 Splicing affected regions detected by different algorithms

Each algorithm reports splicing altered genomic features in different ways. In order to standardize the identification of regions of interest we proceeded as follows:

- LeafCutter: We first identified clusters presenting adjusted pvalues< 0.05 as reported in 'leafcutter_ds_cluster_significance.txt' file. For each of these statistically significant clusters we considered the associated genomic-regions reported in 'leafcutter_ds_effect_size.txt' file with $|\Delta\Psi| > 0.1$.

- MAJIQ: We considered the genomic-region covering junction clusters presenting at least one junction with $P(|\Delta\Psi| > 0.2) > 0.95$.

- rMATS: We considered the values reported in 'JCEC.txt' files. This means that we considered a model that evaluated splicing with reads that spanned splicing junctions and reads on targets bins (i.e. alternatively spliced exons). We kept junctions presenting adjusted FDR$<$ 0.05 and inclusion signal larger than a 0.1 level. Genomic regions were then defined according the following rules:
  - A3SS' (A3SS.MATS.JCEC.txt file): We considered the genomic region between 'shortEE' and 'longExonEnd' coordinates for negative strand and by 'longExonStart_0base' and 'shortES' for positive strand cases.
  - A5SS' (A5SS.MATS.JCEC.txt file): We considered the genomic region between 'shortEE' and 'longExonEnd' coordinates for positive strand and by 'longExonStart_0base' and 'shortES' for negative strand cases.
  - MXE (MXE.MATS.JCEC.txt file): We considered two regions per event defined by: '1stExonStart_0base', '1stExonEnd' and '2ndExonStart_0base', '2ndExonEnd'.
  - SE (SE.MATS.JCEC.txt file): We considered the regions between 'exonStart_0base' and 'exonEnd'.
  - RI (RI.MATS.JCEC.txt file): We considered the regions between 'riExonStart_0base' and 'riExonEnd'.

4 Bin-coverage detection calls in the synthetic dataset

We decided to further characterized some aspects of bin-coverage detection calls for the synthetic dataset, as this signal provided the major number of discoveries. It can be seen in panel-(A) of Figure SF.2 that fold-change and junction-support signals used in the bin-coverage analysis reported relevant and non-redundant information. Whereas the first one accounted for 37% of true positive instances exclusively detected by this signal, the second one accounted for the specific identification of 12% of the total number of true events. The impact of the selected fold-change threshold value, FC*, on specificity, precision and recall can be appreciated with the aid of the Receiver-Operator and Precision-Recall curves shown in panels (B) and (C) of Figure SF.2. It can be recognized from these figures that with the adopted 3-fold threshold ASpli achieved high recall and precision levels ($\sim$ 80% and $\sim$ 95% respectively) laying at rather moderate levels of false positive rates ($\sim$ 14%).

5 Analysis of false positive calls in the simulated dataset

In this section we analyzed the origin of ASpli detected false positive events.

Simulated changes in isoforms concentration for a given gene produced specific patterns of bin and junction differential usage that depended on the exonic architecture of the gene-isoforms. For each gene for which a change in isoforms concentration was simulated, it was possible to anticipate which bins would present differential usage (we call them *active bins*). As in our simulations we admitted a 20% level of random variability in isoform concentration profiles, some background differential bin usage could also take place for bins belonging to genes for which alternative splicing was not explicitly simulated (we called them *inactive-bins*).

In order to quantify this effect we introduced the *a-priori* Splicing Activation Signal (SAS) value. For each gene, SAS was estimated as the maximum absolute change in isoform concentration actually simulated between conditions. We found that background isoform concentration variability produced non-zero SAS levels for inactive events, but this noise-originated signal remained well bellow the signal reported for active ones. The left-most first and second boxplots in Figure SF.3 depict the distribution of this quantity for the 915 genes for which a splicing event was simulated (*active*), and for the remaining 7518 genes (*inactive*) respectively . On the other hand, the four right-most boxplots show the SAS distribution for false positive calls obtained with different methods. Non explicitely splicing simulated changes were reported for $9, 4, 48$ and 23 genes according to ASpli, LeafCutter, MAJIQ and rMATS algorithms respectively.

6 Comparison of discoveries

A comprehensive comparison of discoveries appeared at first-sight problematic as each algorithm is focused on different genomic features in order to chart splicing landscapes.

For instance, rMATS analyzes genomic regions flanked by upstream and downstream exons to examine canonical splicing events. MAJIQ and LeafCutter, on the other hand, exclusively rely on clusters of split reads that share start or ending junction-ends. Finally ASpli considers both, junction clusters and bin features, i.e. genomic regions defined from disjoint ranges of annotated junctions.

In this context, a first coarse grained comparison could be established at gene-level, comparing the identity of genes housing splicing-altered patterns according to the different analyzed methods. Panel (A) of Figure SF.4 displays a color-coded overlap matrix of affected genes in experiments A and B according to the four examined methodologies. Each cell reports the intersection size and, in brackets, the corresponding overlap coefficient. At gene level, rMATS achieved the largest agreement factor (83% of genes identified in experiment B, were also reported in experiment A). However, it also produced the lowest number of discoveries (119). ASpli, on the other hand, presented a comparable level of agreement (71%), highlighting a significatively larger number of concordant genes (2109). Typically, more than 50% of genes identified by any methodology was also reported by ASpli (first and second rows of Figure SF.4). Moreover, the number of concordant discoveries between experiments considering a given methodology was comparable to the agreement level achieved between each experiment-metodology combination and the correpsonding ASpli result. Noticeably, more than 90% of MAJIQ's genes were also spotted by ASpli.

A more in-depth comparison could be established analyzing the overlap of identified genomic regions. In panels (b) and (c) of Figure SF.4 we informed the extent of the overlaps between genomic regions found to be affected by differential splicing patterns according to each algorithm (see Material and Methods **??**) to map events reported by each method to a common set of genomic coordinates). While any kind of overlap was registered for panel (b), only complete inclusion of genomic regions identified by one method inside the ones identified by a second one was considered for panel (c). Statistically significant overlaps were marked with asterisks. Note that overlap coefficients (in brackets) exceeding unity were detected in between-experiments comparisons for LeafCutter and rMATS as a result of the presence of one-to-many region mappings.

For the loose overlap criterium we found statistically significant concordance between discoveries for almost every cell (Fig SF.4-b). Only specific comparisons involving MAJIQ and rMATs failed the statistical significance test. At the same time, overlap coefficient values were similar to the ones estimated at the gene-level analysis. Noticeably, we recognised a sensible reduction in this quantity for the MAJIQ vs ASpli comparison. This finding highlighted that gene-level agreement should in general be considered with caution. A more detailed examination at the sub-genic level might be necessary to assess for discovery consistencies between algorithms. Results for the most stringent overlap criterion are shown in Figure SF.4(c). As expected, a major decrease on overlap coefficient values was observed . However, statistically significant agreement between results was still found as a general rule. Only comparisons involving MAJIQ's discoveries failed the statistical assessments.

7 Data consolidation

We took advantage of ASpli capabilities to deal with complex experimental designs to consolidate datasets A and B in a statistically sound way. We considered the following generalized linear model:

$$y \sim experiment + genotype + experiment : genotype \tag{SE.1}$$

'experiment' was a fixed effect to cope with specific technical biases, and the 'genotype' factor was meant to capture the PRMT5 vs wild-type effect. The third term was an interaction term, and was used to enforced the exclusion of non-coherent signals between experiments. ASpli detected 4360 genomic regions displaying strong evidence of a genotype effect (fdr < 0.05). In addition, 99% of these PRMT5-related events (4314 out of 4360) passed a filtering step to enforce they presented no-detectable evidence of experiment-genotype interactions (experiment:genotype associated fdr > 0.5). These 4314 events defined the consolidated AB data set.

We found that 99% (2209 out of 2241) of the concordant discoveries independently detected in both assays were also included in the consolidated dataset (we included a Venn diagram of the discoveries reported for experiments A, B, and the consolidated data-set AB in Sup.Fig. SF.6). Noticeably, the consideration of the AB data-set allowed to almost double the number of detected genomic regions displaying robust evidence of differential splicing patterns.

8 PRMT5 PCR events

We characterized the agreement between the 23 splicing events that ASpli uncovered for the consolidated AB case, and the 44 Sanchez qRT-PCR validated events in Table ST.2. For each assayed event we included the kind of the original event and the reported qRT-PCR splicing signal value in the second and third columns respectively (Sanchez and collaborators calculated the fraction of the shortest isoform in PRMT5 mutants and wildtype plants detected by qRT-PCR, and used the relativized difference between them as a quantitative proxy of splicing changes (Table 4 of Sanchez et al. (2010))). In the fourth column we informed whether the PCR-interrogated genomic region overlapped with the one signaled by ASpli. Finally, the type of splicing event detected by ASpli was included in the last column of the table.

9 Prostate cancer dataset:Transcriptomic variability

In order to visualize the transcriptomic variability across patients at gene expression levels we considered the 30% most variable genes across the 28 expression profiles that presented more than 10 counts per million reads in at least 3 samples. With this informative set of 1386 genes we built a multidimensional scaling plot of distances between gene expression profiles estimated with the edgeR package Robinson et al. (2010). Results are shown in Fig SF.5 (reported results were very robust against changes in the number of top-variable genes chosen to characterize the transcriptome of each sample). In this kind of plot, samples lay on a two-dimensional scatterplot so that distances on the plot approximate the typical log2 fold changes between the samples (function plotMDS of edgeR Robinson et al. (2010)).

Emtpy and filled symbol correspond to tumor and normal tissue samples respectively. Pair of points of a given patient are equally colored and joined by a dashed edge. It can be seen that tumor and normal samples were well separated across the leading reduced dimension. The second largest projected dimension, however, let us appreciate internal structure and some variability between patients. There was a group of 5 patients (top left empty points) that displayed a rather homogeneous pattern of changes between tumor affected and normal tissues. On the contrary, the 9 bottom-left tumor samples seemed to segregate into a different cluster of transcriptomes. Moreover, the corresponding patients presented different kinds of alterations between tumor and control samples.

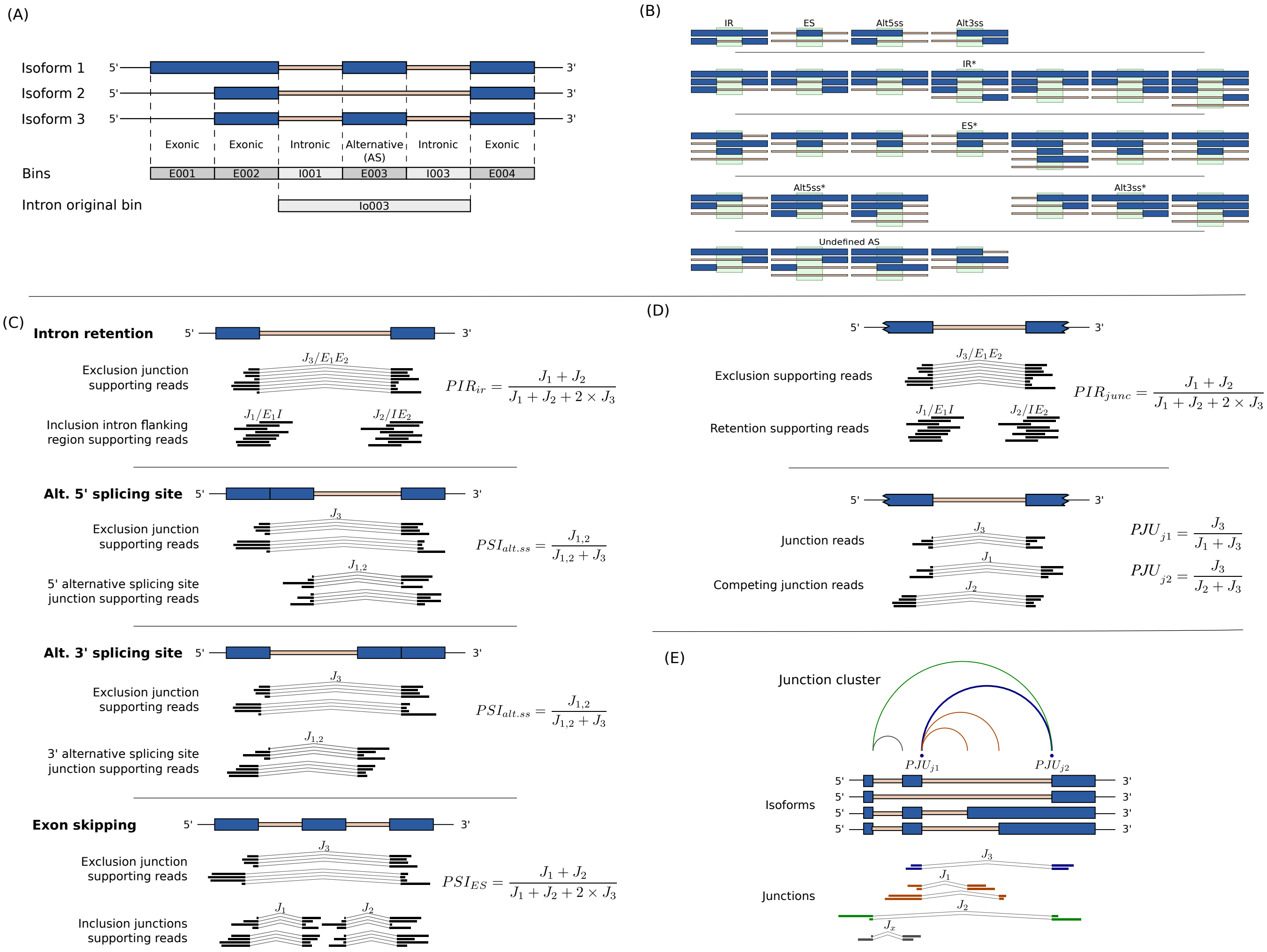

(A)
Isoform 1
Isoform 2
Isoform 3
5'
3'
Exonic
Exonic
Intronic
Alternative (AS)
Intronic
Exonic
Bins
E001
E002
I001
E003
I003
E004
Intron original bin
Io003
(B)
IR
ES
Alt5ss
Alt3ss
IR*
ES*
Alt5ss*
Alt3ss*
Undefined AS
(C)
Intron retention
Exclusion junction supporting reads
J_3/E_1E_2
$PIR_{ir} = \frac{J_1 + J_2}{J_1 + J_2 + 2 \times J_3}$
Inclusion intron flanking region supporting reads
J_1/E_1I
J_2/IE_2
Alt. 5' splicing site
Exclusion junction supporting reads
J_3
$PSI_{alt.ss} = \frac{J_{1,2}}{J_{1,2} + J_3}$
5' alternative splicing site junction supporting reads
$J_{1,2}$
Alt. 3' splicing site
Exclusion junction supporting reads
J_3
$PSI_{alt.ss} = \frac{J_{1,2}}{J_{1,2} + J_3}$
3' alternative splicing site junction supporting reads
$J_{1,2}$
Exon skipping
Exclusion junction supporting reads
J_3
$PSI_{ES} = \frac{J_1 + J_2}{J_1 + J_2 + 2 \times J_3}$
Inclusion junctions supporting reads
J_1
J_2
(D)
Exclusion supporting reads
J_3/E_1E_2
$PIR_{junc} = \frac{J_1 + J_2}{J_1 + J_2 + 2 \times J_3}$
Retention supporting reads
J_1/E_1I
J_2/IE_2
Junction reads
J_3
$PJU_{j1} = \frac{J_3}{J_1 + J_3}$
Competing junction reads
J_1
J_2
$PJU_{j2} = \frac{J_3}{J_2 + J_3}$
(E)
Junction cluster
PJU_{j1}
PJU_{j2}
Isoforms
Junctions
J_3
J_1
J_2
J_x

Figure (SF.1) Panel (A) shows how bin-features are defined and classified as: external, exonic, intronic or intron_original bins using genome annotation. The local splicing classification scheme is illustrated in panel (B). The definition of PSI and PIR metrics for bin features are pictured in panel (C). Definition of junction PIR and PJU statistics are shown in panel (D). Panel (E) shows a possible junction cluster and highlights the definition of type J_1, J_2 and J_3 junctions for the analysis of PJU statistics for the blue junction.

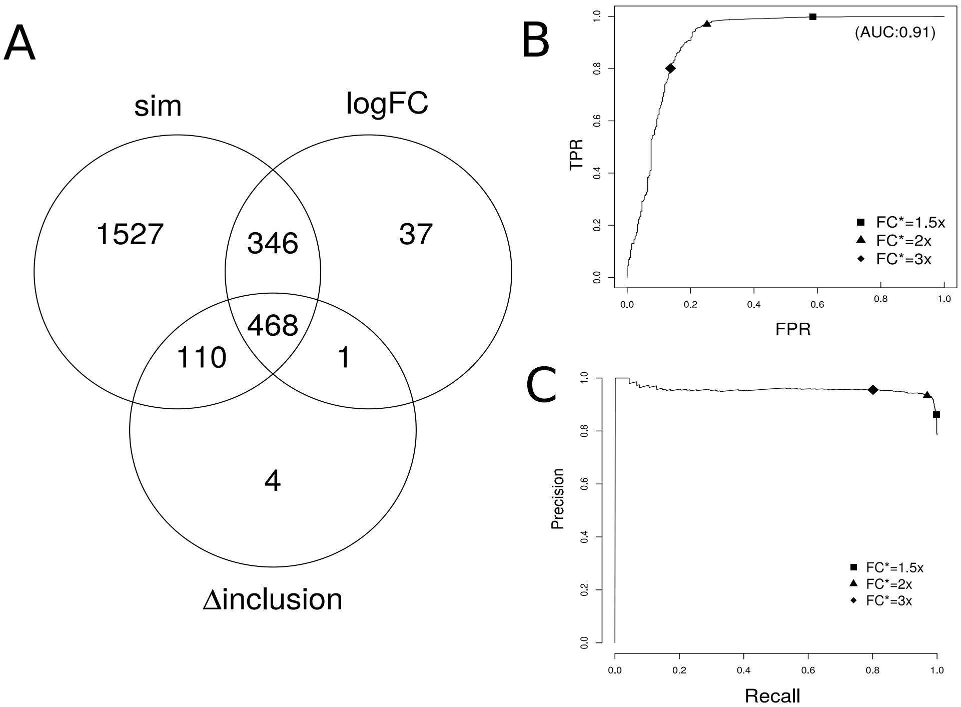


Figure (SF.2) (A) Graphical summary of bin-coverage detection calls. The *sim* set correspond to simulated events. *logFC* and *D-inclusion* sets are associated to statistically significant discoveries presenting large enough fold change and large bin-supporting junction inclusion signals respectively. ROC and Precision-Recall curves, parameterized by the considered fold-change threshold level, are shown for statistically significant bins in panels (B) and (C) respectively.

feature	assesment	index		bin class
bin	inclusion	PIR_{ir}	$\frac{J_1+J_2}{J_1+J_2+2*J_3}$	UAS, I, I*, I_0
		PSI_{es}		UAS, E, E*
		PSI_{alt5ss}	$\frac{J_{1,2}}{J_{1,2}+J_3}$	Alt5ss, Alt5ss*
		PSI_{alt3ss}		Alt3ss, Alt3ss*
junction		PIR_{junc}	$\frac{J_1+J_2}{J_1+J_2+2*J_3}$	-
	usage	PJU_{J_1}	$\frac{J_3}{J_1+J_3}$	-
		PJU_{J_2}	$\frac{J_3}{J_2+J_3}$	-

Table (ST.1) Junction usage and inclusion strength figure of merits for different bin classes and for experimentally detected junctions. The definition of J_1, J_2 and J_3 junction counts is depicted in panels C and D of Figure SF.1 for annotated and experimentally detected junctions respectively.

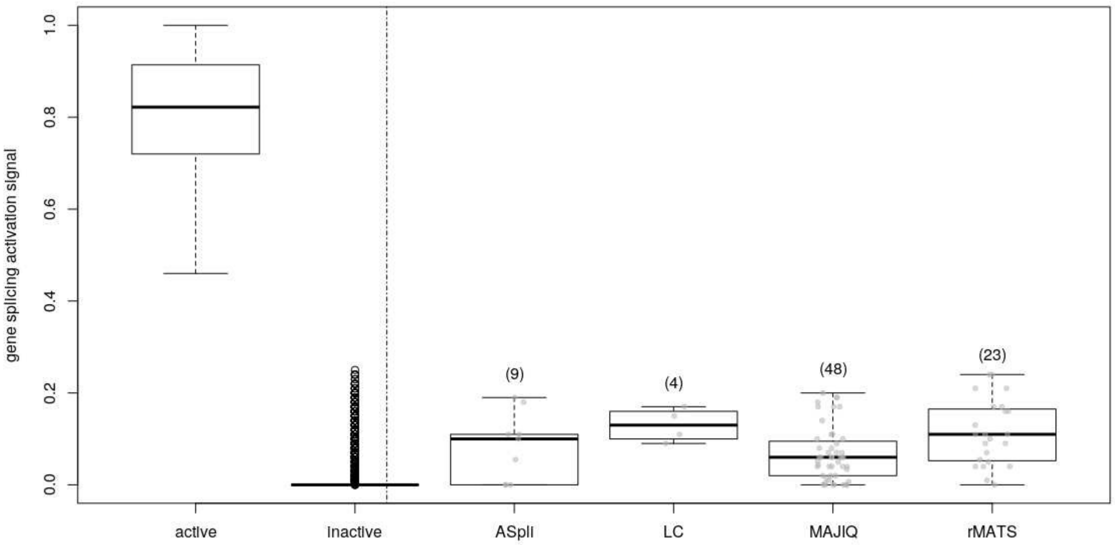
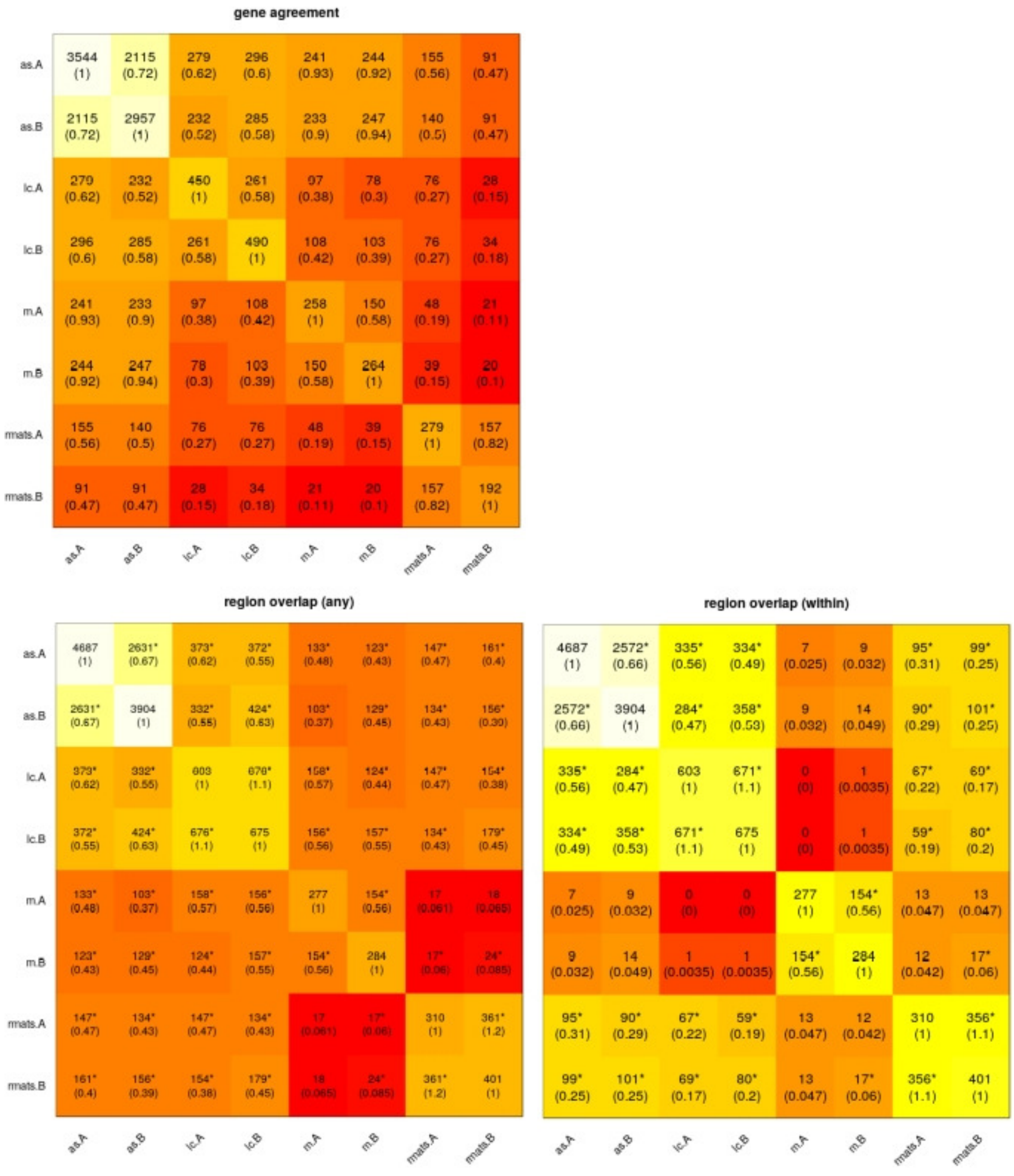


Figure (SF.3) Distribution of Splicing Activation Signals for bins in simulated (*active*) and non-simulated (*inactive*) splicing events are displayed in the first two boxplots. The four right-most boxplot show the SAS distributions for false positive events reported for each algorithm

	Gene ID	Event	qRT-PCR signal	Region overlap	Detected event
1	AT1G53650	5'ss	18.93	yes	IR (next to 5)
2	AT1G54360	5'ss	39.21	yes	Alt5ss
3	AT1G76510	5'ss	-28.00	yes	Alt5ss
4	AT2G04790	5'ss	11.14	yes	IR
5	AT2G15530	5'ss	21.12	yes	Alt 5'/3'
6	AT2G33480	5'ss	-27.16	yes	IR
7	AT2G38880	5'ss	-10.44	no	IR
8	AT2G46790	5'ss	35.20	yes	Alt5ss (plus additional IR)
9	AT3G01150	ES	-13.70	no	IR
10	AT3G12250	5'ss	-16.29	no	ES*
11	AT3G16800	IR/3'ss	-31.59	yes	IR,Novel Alt 5'/3'
12	AT3G19840	5'ss	-26.20	no	IR
13	AT3G20270	ES	8.51	no	IR (next to ES)
14	AT3G23280	ES	18.21	yes	ES
15	AT3G25840	ES	6.51	yes	ES
16	AT4G02430	3'ss	27.45	no	IR
17	AT4G24740	ES	16.07	no	IR (next to ES)
18	AT4G31720	3'ss	12.37	no	IR
19	AT4G32730	5'ss	30.93	yes	Novel Alt 5'/3'
20	AT4G38510	5'ss	15.53	yes	Alt5ss*,CSP
21	AT5G05550	ES	32.72	yes	ES (plus additional IR)
22	AT5G25610	IR	-71.69	yes	IR
23	AT5G57630	5'ss	31.07	yes	Novel Alt 5'/3' (plus adjacent IR)

Table (ST.2) Agreement between ASpli and qRT-PCR results



gene agreement

	as.A	as.B	lc.A	lc.B	m.A	m.B	rmats.A	rmats.B
as.A	3544 (1)	2115 (0.72)	279 (0.62)	296 (0.6)	241 (0.93)	244 (0.92)	155 (0.56)	91 (0.47)
as.B	2115 (0.72)	2957 (1)	232 (0.52)	285 (0.58)	233 (0.9)	247 (0.94)	140 (0.5)	91 (0.47)
lc.A	279 (0.62)	232 (0.52)	450 (1)	261 (0.58)	97 (0.38)	78 (0.3)	76 (0.27)	28 (0.15)
lc.B	296 (0.6)	285 (0.58)	261 (0.58)	490 (1)	108 (0.42)	103 (0.39)	76 (0.27)	34 (0.18)
m.A	241 (0.93)	233 (0.9)	97 (0.38)	108 (0.42)	258 (1)	150 (0.58)	48 (0.19)	21 (0.11)
m.B	244 (0.92)	247 (0.94)	78 (0.3)	103 (0.39)	150 (0.58)	264 (1)	39 (0.15)	20 (0.1)
rmats.A	155 (0.56)	140 (0.5)	76 (0.27)	76 (0.27)	48 (0.19)	39 (0.15)	279 (1)	157 (0.82)
rmats.B	91 (0.47)	91 (0.47)	28 (0.15)	34 (0.18)	21 (0.11)	20 (0.1)	157 (0.82)	192 (1)

region overlap (any)

	as.A	as.B	lc.A	lc.B	m.A	m.B	rmats.A	rmats.B
as.A	4687 (1)	2631* (0.67)	373* (0.62)	372* (0.55)	133* (0.48)	123* (0.43)	147* (0.47)	161* (0.4)
as.B	2631* (0.67)	3904 (1)	332* (0.55)	424* (0.63)	103* (0.37)	129* (0.45)	134* (0.43)	156* (0.30)
lc.A	373* (0.62)	332* (0.55)	603 (1)	676* (1.1)	158* (0.57)	124* (0.44)	147* (0.47)	154* (0.38)
lc.B	372* (0.55)	424* (0.63)	676* (1.1)	675 (1)	156* (0.56)	157* (0.55)	134* (0.43)	179* (0.45)
m.A	133* (0.48)	103* (0.37)	158* (0.57)	156* (0.56)	277 (1)	154* (0.56)	17 (0.061)	18 (0.065)
m.B	123* (0.43)	129* (0.45)	124* (0.44)	157* (0.55)	154* (0.56)	284 (1)	17* (0.06)	24* (0.085)
rmats.A	147* (0.47)	134* (0.43)	147* (0.47)	134* (0.43)	17 (0.061)	17* (0.06)	310 (1)	361* (1.2)
rmats.B	161* (0.4)	156* (0.39)	154* (0.38)	179* (0.45)	18 (0.065)	24* (0.085)	361* (1.2)	401 (1)

region overlap (within)

	as.A	as.B	lc.A	lc.B	m.A	m.B	rmats.A	rmats.B
as.A	4687 (1)	2572* (0.66)	335* (0.56)	334* (0.49)	7 (0.025)	9 (0.032)	95* (0.31)	99* (0.25)
as.B	2572* (0.66)	3904 (1)	284* (0.47)	358* (0.53)	9 (0.032)	14 (0.049)	90* (0.29)	101* (0.25)
lc.A	335* (0.56)	284* (0.47)	603 (1)	671* (1.1)	0 (0)	1 (0.0035)	67* (0.22)	69* (0.17)
lc.B	334* (0.49)	358* (0.53)	671* (1.1)	675 (1)	0 (0)	1 (0.0035)	59* (0.19)	80* (0.2)
m.A	7 (0.025)	9 (0.032)	0 (0)	0 (0)	277 (1)	154* (0.56)	13 (0.047)	13 (0.047)
m.B	9 (0.032)	14 (0.049)	1 (0.0035)	1 (0.0035)	154* (0.56)	284 (1)	12 (0.042)	17* (0.06)
rmats.A	95* (0.31)	90* (0.29)	67* (0.22)	59* (0.19)	13 (0.047)	12 (0.042)	310 (1)	356* (1.1)
rmats.B	99* (0.25)	101* (0.25)	69* (0.17)	80* (0.2)	13 (0.047)	17* (0.06)	356* (1.1)	401 (1)

Figure (SF.4) Number of concordant discoveries produced by different methods. Gene-level, region-overlap and strong-region-overlap results are displayed in the top-left, bottom-left and bottom-right panels respectively. Overlap coefficients are included between brackets. See text for details.

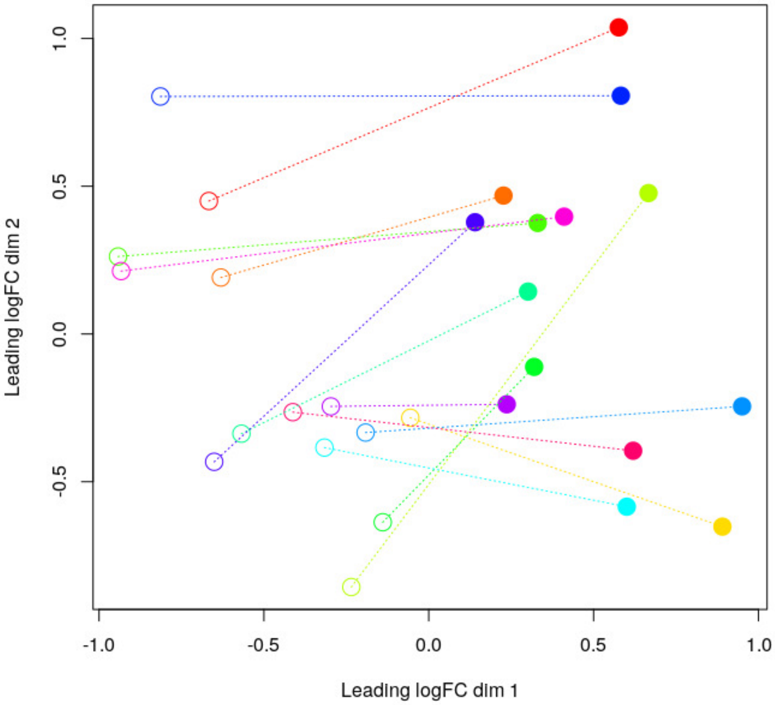


Figure (SF.5) Multidimensional scaling plot of distances between gene expression profiles. Normal and tumor samples are depicted using empty and filled circles respectively. Paired-samples were uniformily colored and connected with a dotted line.

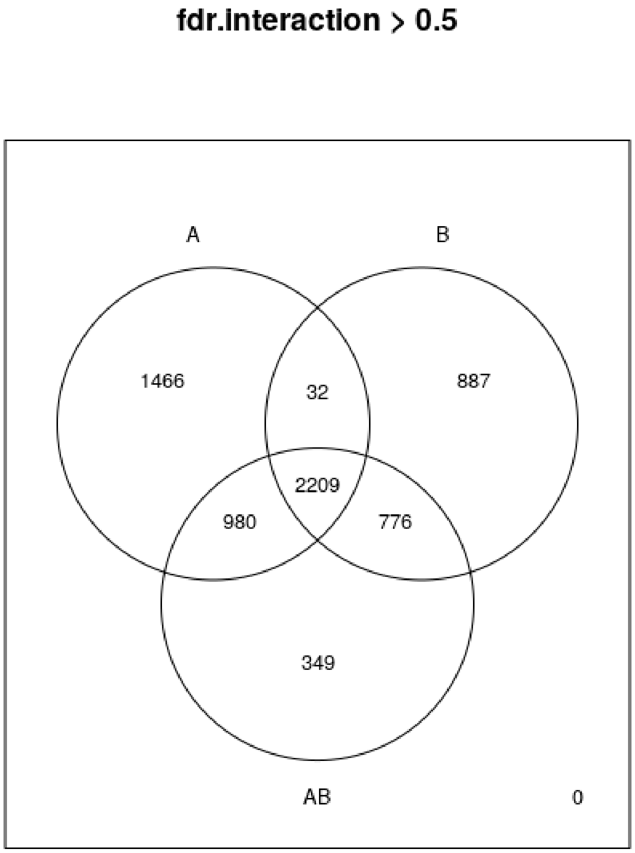


Figure (SF.6) Venn diagram of alternative splicing events detected in experiments A, B, and the consolidated data set AB (i.e. events displaying strong evidence of a genotype effect (fdr< 0.05) and no-detectable evidence of experiment-genotype interaction (experiment:genotype associated fdr> 0.5)).

10 Bibliography

References

S. Anders, A. Reyes, W. Huber, Detecting differential usage of exons from rna-seq data, Genome Research 22 (2012) 2008–2017.

S. Schafer, K. Miao, C. C. Benson, M. Heinig, S. A. Cook, N. Hubner, Alternative splicing signatures in rna-seq data: Percent spliced in (psi), Current Protocols in Human Genetics 87 (2015) 11.16.1–11.16.14.

S. E. Sanchez, E. Petrillo, E. J. Beckwith, X. Zhang, M. L. Rugnone, C. E. Hernando, J. C. Cuevas, M. A. Godoy Herz, A. Depetris-Chauvin, C. G. Simpson, J. W. S. Brown, P. D. Cerdán, J. O. Borevitz, P. Mas, M. F. Ceriani, A. R. Kornblihtt, M. J. Yanovsky, A methyl transferase links the circadian clock to the regulation of alternative splicing, Nature 468 (2010) 112–116.

M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics 26 (2010).