Bioinformatics, 36(7), 2020, 2134–2141
doi: 10.1093/bioinformatics/btz915
Advance Access Publication Date: 13 December 2019
Original Paper

Gene expression

dSreg: a Bayesian model to integrate changes in splicing and RNA-binding protein activity

Carlos Martí-Gómez, Enrique Lara-Pezzi* and Fátima Sánchez-Cabo*

Molecular Regulation of Heart Failure (CMG and ELP); Bioinformatics Unit (FSC), Centro Nacional de Investigaciones Cardiovasculares (CNIC), Madrid 28029, Spain

*To whom correspondence should be addressed.

Associate Editor: Jan Gorodkin

Received on January 18, 2019; revised on September 9, 2019; editorial decision on November 30, 2019; accepted on December 10, 2019

Abstract

Motivation: Alternative splicing (AS) is an important mechanism in the generation of transcript diversity across mammals. AS patterns are dynamically regulated during development and in response to environmental changes. Defects or perturbations in its regulation may lead to cancer or neurological disorders, among other pathological conditions. The regulatory mechanisms controlling AS in a given biological context are typically inferred using a two-step framework: differential AS analysis followed by enrichment methods. These strategies require setting rather arbitrary thresholds and are prone to error propagation along the analysis.

Results: To overcome these limitations, we propose dSreg, a Bayesian model that integrates RNA-seq with data from regulatory features, e.g. binding sites of RNA-binding proteins. dSreg identifies the key underlying regulators controlling AS changes and quantifies their activity while simultaneously estimating the changes in exon inclusion rates. dSreg increased both the sensitivity and the specificity of the identified AS changes in simulated data, even at low read coverage. dSreg also showed improved performance when analyzing a collection of knock-down RNA-binding proteins' experiments from ENCODE, as opposed to traditional enrichment methods, such as over-representation analysis and gene set enrichment analysis. dSreg opens the possibility to integrate a large amount of readily available RNA-seq datasets at low coverage for AS analysis and allows more cost-effective RNA-seq experiments.

Availability and implementation: dSreg was implemented in python using stan and is freely available to the community at https://bitbucket.org/cmartiga/dsreg.

Contact: elara@cnic.es or fscabo@cnic.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Eukaryotic genes are generally constituted by exons and introns (Kim *et al.*, 2007). Alternative mRNAs may be generated from the same gene by inclusion or skipping of a particular exon in the mature transcript, in a process known as alternative splicing (AS) (Graveley, 2001; Nilsen and Graveley, 2010). There is evidence of AS for most mammalian genes (Barbosa-Morais *et al.*, 2012; Merkin *et al.*, 2012) and of widespread changes of AS patterns throughout brain and heart development (Auinash and Cooper, 2012; Baralle and Giudice, 2017; Fogel *et al.*, 2012; Giudice *et al.*, 2014; Irimia *et al.*, 2014; Quesnel-vallières *et al.*, 2015; Raj *et al.*, 2014; Weyn-Vanhentenryck *et al.*, 2018). Defects in mRNA processing of some specific genes often lead to disease (Baralle and Giudice, 2017; Lara-Pezzi *et al.*, 2013, 2017) and have been associated with complex neurological disorders, such as autistic syndrome (Irimia *et al.*, 2014; Lee *et al.*, 2016; Quesnel-vallières *et al.*, 2015; Wagnon *et al.*, 2012), and cancer (Climente-GonzáLez *et al.*, 2017; Stricker *et al.*, 2017). Therefore, understanding the regulatory mechanisms underlying physiologic and pathological changes in AS patterns is crucial, not only to understand RNA biology, but also to identify potential therapeutic targets with a more general effect in complex diseases.

A two-step workflow is generally applied to identify the regulatory mechanisms underlying the changes in AS (see Fig. 1 for a schematic representation). First, AS changes must be identified. For this, short reads from RNA sequencing are typically mapped using splice junctions aware aligners, such as STAR or Hisat2 (Dobin *et al.*, 2013; Pertea *et al.*, 2016). Alternative mRNA processing can be studied at two different levels: (i) transcript quantification level, which can be based on a prior alignment (Pertea *et al.*, 2016; Trapnell *et al.*, 2013), or can be directly estimated from fast pseudoalignment methods (Bray *et al.*, 2016; Patro *et al.*, 2017); and (ii) event level quantification, as performed by popular tools such as MISO, MATS, vast-tools, DEXseq or SUPPA (Alamancos *et al.*, 2014; Anders *et al.*, 2012; Irimia *et al.*, 2014; Katz *et al.*, 2010; Shen *et al.*, 2012; Trincado *et al.*, 2018). Recent tools showed improved

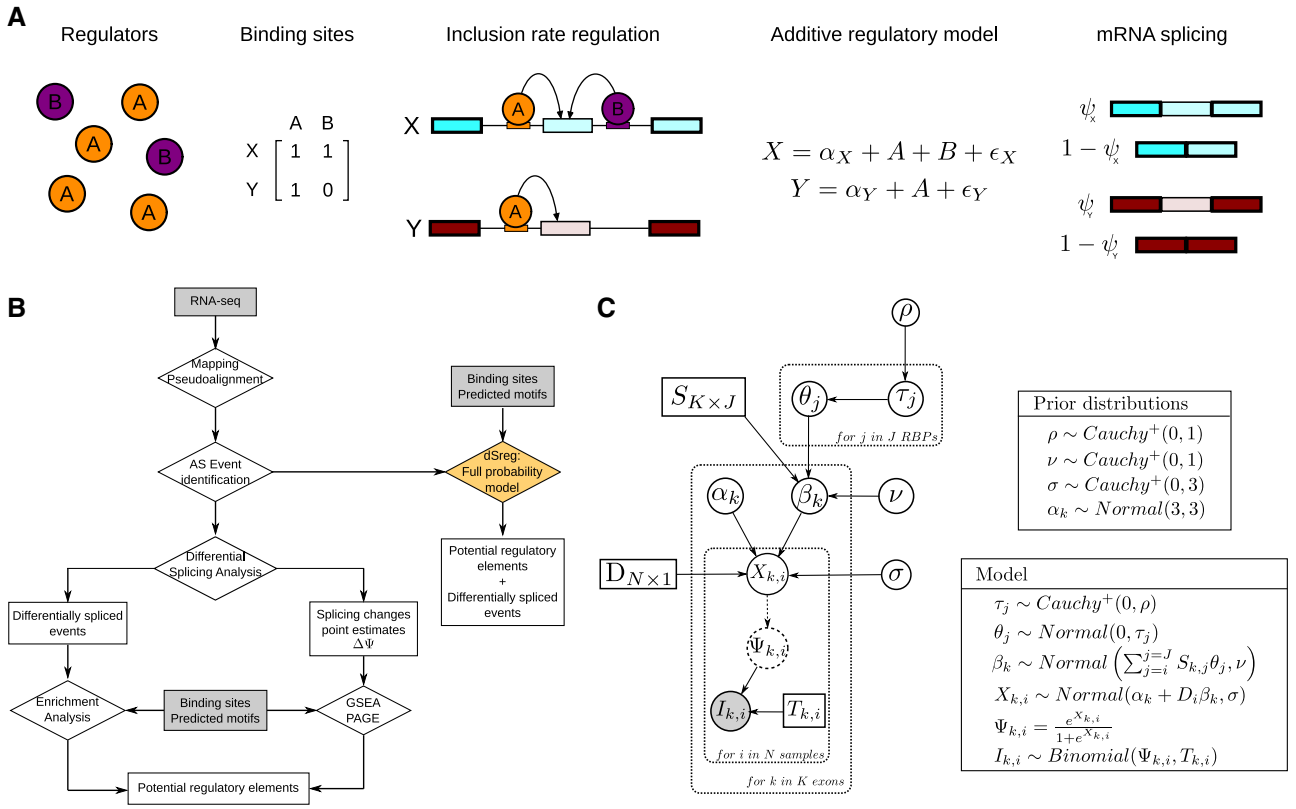

Fig. 1. General and proposed work-flows for AS regulation analysis. (**A**) Schematic representation of idealized model for the regulation of splicing rates by RBPs through direct binding to their binding motifs in the pre-mRNA. (**B**) Diagram representing the different steps required for a classical analysis of regulation of AS using RNA-seq data and the proposed model in dSreg. (**C**) Directed acyclic graph representing the full probabilistic model integrating both differential AS analysis with binding sites presence and changes in the activity of RBPs

performance for the estimation of AS changes by using information about exon features and perturbation experiments (Huang and Sanguinetti, 2017; Zhang *et al.*, 2019). Since regulation is expected to take place locally, the event level analysis is the preferred approach to study the regulatory mechanisms underlying changes in AS profiles. Once AS events have been identified and quantified, different statistical approaches can be applied to determine AS changes between biological conditions, being a generalized linear model (GLM) with binomial likelihood the most natural parametric approach (Shen *et al.*, 2012).

The second step aims to identify regulatory features, such as RNA-binding proteins (RBPs) motifs, associated with AS changes. Such features often include nucleotide hexamers, predicted motifs, experimentally determined or predicted binding sites (Dominguez *et al.*, 2018; Giudice *et al.*, 2016; Ray *et al.*, 2013; Yang *et al.*, 2015). Over-representation analysis (ORA) enables the discovery of features co-occurring with significant AS changes more often than expected by chance. Therefore, a sufficiently large set of significantly changed events is required to reach enough power to detect enrichment of regulatory features. ORA requires the categorization of splicing changes into different groups e.g. included or skipped, ignoring quantitative information about AS changes. To make use of quantitative information in the enrichment procedure, several approaches have been developed, including the widely known gene set enrichment analysis (GSEA) (Simillion *et al.*, 2017; Subramanian *et al.*, 2005). Although these tools were designed for functional analysis, they have been used to perform enrichment of known targets of regulatory elements (Sebestyén *et al.*, 2016; Trincado *et al.*, 2018). However, the inherently noisier nature of the estimation of differences in AS compared to those of differential gene expression may limit the applicability of GSEA-like methods. Moreover, an additional limitation affecting both ORA and GSEA approaches lies on the high number of different features or binding sites and on the potential co-linearities among them that should be considered, resulting otherwise in a high false positive rate derived from confounding effects.

In this work, we used simulated data to study, for the first time, the performance and the limitations of the classical enrichment approaches (ORA and GSEA) for the detection of regulatory elements driving AS changes. To tackle some of these limitations, we developed dSreg, a probabilistic model integrating differential splicing and regulation analyses. dSreg models latent changes in inclusion rates as a linear combination of the regulatory effects of the RBPs binding to relevant regions of the pre-mRNA for every identified AS event. Moreover, we used a hierarchical shrinkage prior distribution to model the changes in the activity of RBPs to formalize the assumption that only a few RBPs would show changes in their activity. dSreg was applied to simulated and real data, including data from systematic RBPs knock-down experiments, to assess its performance against ORA and GSEA. Finally, we applied dsReg to a real RNA-seq dataset obtained from a cardiomyocyte differentiation experiment, for which a limited number of AS regulators might be assumed.

2 Materials and methods

2.1 dSreg: a mechanistic probability model for differential splicing

dSreg models the AS changes between two different conditions, *a* and *b*, as a function of changes in the activity of a few of the existing RBPs acting through their known binding sites. Given K AS events detected across N samples, we observe $I_{k,i}$ reads supporting exon inclusion out of a total of $T_{k,i}$ reads mapping to the kth exon skipping event in sample i, which depends on the unknown probability of

inclusion $\Psi_{k,i}$. The conditional probability of observing $I_{k,i}$ reads given $T_{k,i}$ and $\Psi_{k,i}$ is given by the binomial distribution,

$$p(I_{k,i} \mid T_{k,i}, \Psi_{k,i}) = Binomial(I_{k,i} \mid T_{k,i}, \Psi_{k,i}). \quad (1)$$

$\Psi_{k,i}$ is therefore different for each sample i, but depends on the condition or group to which it belongs. Since probabilities are bound between 0 and 1, to model this dependency, we take the logit transformation $X_{k,i}$,

$$X_{k,i} = log\left(\frac{\Psi_{k,i}}{1-\Psi_{k,i}}\right). \quad (2)$$

We assume that $X_{k,i}$ is drawn from a normal distribution with a common standard deviation $\boldsymbol{\sigma}_k$ and different means per condition: $\boldsymbol{\alpha}_k$ for condition a; and $\alpha_k + \beta_k$ for condition b, such that $\boldsymbol{\beta}_k$ represents the difference between the two conditions. For simplicity, we assume here that the standard deviation is the same across all K AS events ($\sigma_k = \sigma$).

$$p(X_{k,i} \mid D_i, \alpha_k, \beta_k, \sigma) = Normal(X_{k,i} \mid \alpha_k + D_i\beta_{k,i}, \sigma) \quad (3)$$

where D_i is a constant that takes the value 1 when the sample belongs to condition b, and 0 when it belongs to condition a:

$$D_i = \begin{cases} 1 & \text{if sample } i \text{ in group } a \\ 0 & \text{if sample } i \text{ in group } a \end{cases}.$$

So far, this model is a simple logistic regression for each event with the only assumption that the sample variance is common across events and conditions. However, the changes in the probability of inclusion of exon k between two conditions, indirectly modeled by $\boldsymbol{\beta}_k$, should depend on the change in the activity $\boldsymbol{\theta}_j$ of a particular regulatory RBP j and on whether it can bind to exon k. The binding information is encoded in a matrix $\mathbf{S}_{K\times J}$, with value 1 whenever the RBP j binds to the exon k and 0 otherwise. Position dependent effects can be easily included by considering RBP j binding to different relative locations as different and independent RBPs. At the same time, the matrix $\mathbf{S}$ could also contain continuous values such as the probabilities of binding, affinities or scores given by Position Weighted Matrices (Ray *et al.*, 2013) or any other predictive tool (Alipanahi *et al.*, 2015; Maticzka *et al.*, 2014).

$$S_{k,j} = \begin{cases} 1 & \text{if the combination of RBP}-\text{region } j \text{ is present in event } k \\ 0 & \text{otherwise} \end{cases}.$$

Now we can model $\boldsymbol{\beta}_k$, the change in the logit-transformed inclusion rate of exon k, as a normal distribution centered at a linear combination of regulatory effects θ and S_k (the binding profile of exon k) with certain standard deviation $\boldsymbol{\nu}$. Adding variance $\boldsymbol{\nu}$ to the distribution of $\boldsymbol{\beta}_k$ allows the existence of some changes in AS not necessarily explained by the regulatory features included in the model,

$$p(\beta_k \mid \theta, S_k, \nu) = Normal\left(\beta_k \mid \sum_{j=0}^{j=J} S_{k,j}\theta_j, \nu\right). \quad (4)$$

In this type of exploratory analysis, large numbers of regulatory proteins are usually tested. However, we expect that AS changes are driven by only a few RBPs. We formalize this prior belief setting a horseshoe prior for the change in the activity of regulator j $\boldsymbol{\theta}_j$ (Carvalho *et al.*, 2009). The horseshoe prior, a member of the family of hierarchical shrinkage priors, specifies a normal prior for $\boldsymbol{\theta}_j$ with mean 0 and a standard deviation $\boldsymbol{\tau}_j$, where $\boldsymbol{\tau}_j$ is not a fixed value, but drawn from a common half Cauchy distribution with mean 0 and $\boldsymbol{\rho}$ scale parameter. $\boldsymbol{\tau}_j$ represents a local shrinkage parameter, as it only affects protein j, whereas $\boldsymbol{\rho}$ can be understood as a global shrinkage parameter. We further set a half Cauchy prior in $\boldsymbol{\rho}$ with mean 0 and standard deviation 1 as recommended (Carvalho *et al.*, 2009). Note that this prior can be adapted according to the expected number of non-zero parameters (Piironen and Vehtari, 2017):

$$p(\theta_j \mid \tau_j) = Normal(\theta_j \mid 0, \tau_j) \quad (5)$$

$$p(\tau_j \mid \rho) = Cauchy^{+}(\tau_j \mid 0, \rho) \quad (6)$$

$$p(\rho) = Cauchy^{+}(\rho \mid 0, 1). \quad (7)$$

Finally, we need to specify prior distributions for the remaining parameters $\boldsymbol{\alpha}_k$ and $\boldsymbol{\sigma}$. Since we expect most of the exons to be included most of the times ($\Psi \sim 1$) and $\boldsymbol{\alpha}_k$ is the logit transformation of the inclusion rate in condition a, we set a normal prior centered at 3 (which reflects an expected $\Psi = 0.95$), with standard deviation 3 for each exon k to enable some deviation from this expectation. Moreover, as we expect little variation among samples, we set a half Cauchy prior distribution with 0 mean and standard deviation 1 on $\boldsymbol{\sigma}$,

$$p(\alpha_k) = Normal(\alpha \mid 3, 3) \quad (8)$$

$$p(\sigma) = Cauchy^{+}(\sigma \mid 0, 1). \quad (9)$$

The joint posterior probability of the parameters $\boldsymbol{\Theta}$ given the data ($\mathbf{I}$) is proportional to the joint probability distribution of the data and $\boldsymbol{\Theta}$, since the marginal probability of obtaining the data $p(\mathbf{I})$ is constant for any $\boldsymbol{\Theta}$,

$$p(\Theta|\mathbf{I}) = \frac{p(\Theta,\mathbf{I})}{p(\mathbf{I})} \propto p(\Theta,\mathbf{I}). \quad (10)$$

Using the conditional probabilities and prior distributions that we have defined for each variable, we can calculate this joint probability distribution applying the chain rule,

$$\begin{aligned} &p(\Theta,\mathbf{I}) = \\ &= p(\mathbf{I},\mathbf{T},X,\alpha,\beta,\nu,\theta,\tau,\rho,D,S) = \\ &= p(\Theta,\mathbf{I}) = p(\sigma)p(\nu)p(\rho)\prod_j^J[p(\theta_j|\tau_j)p(\tau_k|\rho)]\prod_k^K[p(\beta_k|\mathbf{S},\theta,\nu)p(\alpha_k)P(I_k)] \end{aligned} \quad (11)$$

where,

$$P(I_k) = \prod_i^N \left(p(I_{k,i}|T_{k,i},X_{k,i})p(X_{k,i}|\alpha_k,\beta_k,\sigma,D_i)\right) \quad (12)$$

Once the full posterior distribution is completely specified, it can be explored using Markov Chain Monte Carlo algorithms. We implemented this model in stan (Carpenter *et al.*, 2017), using a non-centered parameterization whenever possible to alleviate sampling difficulties from hierarchical models (Betancourt and Girolami, 2013). The full model is represented as a directed acyclic graph to show dependencies among parameters in Figure 1B.

3 Results

3.1 Adding information about regulatory elements improves the detection of AS changes even at low sequencing depth

Using simulated data we first compared the performance of a standard GLM to detect changes in AS at different sequencing depths ($\boldsymbol{\lambda}$). Correlation between the estimated $\hat{\beta}_k$ and the real $\boldsymbol{\beta}_k$ used for the simulations was generally low and did not increase with sequencing depth. When focusing on detection of AS changes, we found that, as expected, at low sequencing depths ($log(\lambda) \leq 3$), the sensitivity at a 5% false discovery rate (FDR) was smaller than 10% when using a simple GLM. As $\boldsymbol{\lambda}$ increased, so did the sensitivity of the GLM (Fig. 2B). Interestingly, the $F1$ score, which integrates both sensitivity and specificity, saturated with depth, suggesting that after some point, there was not much gain by increasing sequencing depth (Fig. 2C). To avoid the need to select an arbitrary threshold to assess the performance of the different methods, we additionally calculated the receiver operating characteristic (ROC) curves for each simulated dataset and the area under them (AUROC, Fig. 2D and E). These results showed that, at low sequencing depths ($log(\lambda) < 3$),

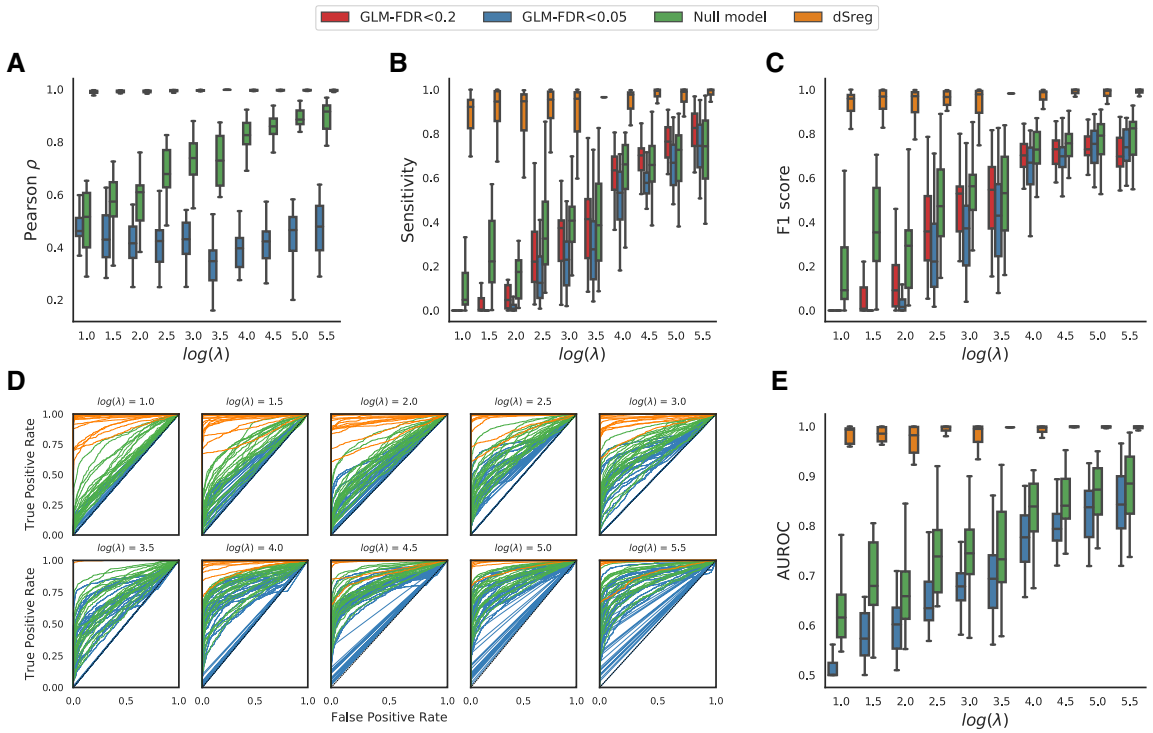

Fig. 2. Comparison of the performance for the identification of different event inclusion rates of a standard method using a single GLM per exon considering two FDR thresholds (0.05 and 0.2), a Bayesian model that pools variance across all exons (*Null model*) and dSreg. Performance was analyzed in simulations with increasing sequencing depths λ (the mean of the Poisson distribution used to simulate the total number of reads mapping to an exon skipping event). (**A**) Pearson correlation between real and estimated β_j. (**B**) Sensitivity. (**C**) *F*1 score. (**D, E**) ROC curves (D) and the area under them (E)

the performance was rather poor, with AUROC values of 0.7 at most.

In order to check whether potential improvements of dSreg were due to the inclusion of binding sites and changes in RBPs activity in the model or just to variance pooling, we ran dSreg and a reduced model that only pools variance from all exons without taking into account of the binding sites and changes in regulatory activities (*Null model*). We defined significantly changed events as those with a posterior probability higher than 95% of having a $\beta_k > 0$. The *Null model* already outperformed the GLM at the single exon level and improved quantitative estimation of $\boldsymbol{\beta_k}$ with depth (Fig. 2A). However, dSreg showed a much greater improvement in correlation and sensitivity, even at very low sequencing depths ($log(\lambda) < 3$), when there was practically no information from individual events (Fig. 2). This increased sensitivity did not come with a decrease in specificity as could be expected, since it showed also very high *F*1 scores and AUROC, suggesting that differences in performance are intrinsic to the method and not threshold dependent (Fig. 2C–E). Results with the *Null model* suggest that pooling variance across events does only marginally improve the inference of splicing changes, at least with the low variance used in these simulations. dSreg, in contrast, additionally used the information about the underlying regulatory mechanisms to correct differences that may easily arise by chance in datasets with limited sample size, given that simulations were done with only three samples per condition.

3.2 dSreg improves the detection of the RBPs driving AS changes

Once AS changes have been identified, we focused on the detection of the regulatory elements potentially controlling these events. Using our simulated datasets, we compared dSreg with the traditional ORA and GSEA approaches. As FDR<0.2 filtering showed higher *F*1 score in the identification of splicing changes (Fig. 2C), we used this threshold to select significantly changed events to perform the downstream enrichment analyses. The dependency of ORA on the detection of significant changes led to low *F*1 scores for GLM results at any tested FDR threshold, especially at low sequencing depths (Fig. 3A). We also used an in-house version of GSEA to take advantage of quantitative information in the identification of regulatory elements. Briefly, events were ranked according to their maximum likelihood estimation of the coefficient of the GLM, which represents the log of the odds ratio of inclusion between the two conditions. Then, we looked for non-random distributions of binding sites along the ranked list (Subramanian *et al.*, 2005) (see Section 2 for details). We found a substantial improvement over ORA, with higher *F*1 scores, especially at low sequencing depths, but did not seem to benefit from higher sequencing depths (Fig. 3A). dSreg outperformed both ORA and GSEA at every evaluation metric, and was barely affected by low sequencing depths (Fig. 3). Therefore, integration of the two sources of information improves results both in terms of inference of differential inclusion rates and the identification of the mechanisms driving those changes.

3.3 Increasing the number of potential regulatory elements does not decrease dSreg performance

We have so far used simulated data to explore the effect of sequencing depth on both the detection of splicing changes and on the identification of the key RBPs driving these changes. We next assessed the impact of the number of regulators, which may increase the number of false positives, particularly in presence of co-linearities among binding profiles of different RBPs. To study this potential limitation, we simulated datasets with only five active RBPs as in the previous simulations, but increasing the number of total RBPs included in the analysis up to 250. We found that the *F*1 score tended to decrease as the number of potential regulators increased with either ORA or GSEA, despite multiple test correction to control FDR. Once more, dSreg outperformed both methods and remained unaffected by the inclusion of other inactive regulatory elements (Fig. 3B).

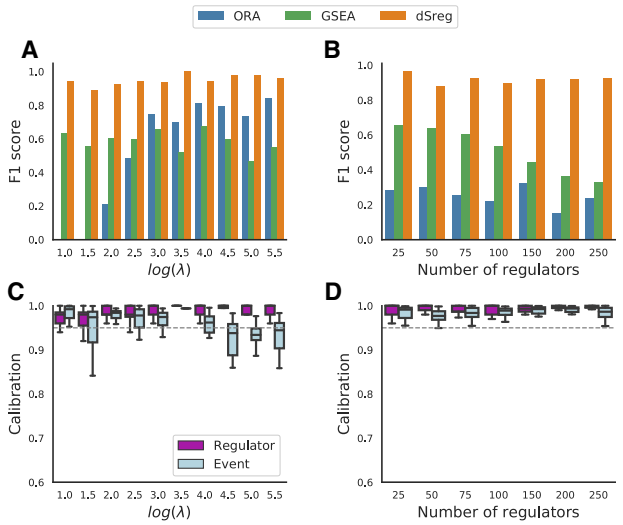

Fig. 3. Performance of methods for the detection of regulatory elements: ORA with variable FDR thresholds (0.05 and 0.2), non-parametric GSEA and dSreg. Performance was analyzed in simulations with increasing sequencing depths λ, which is the mean of the Poisson distribution used to simulate the total number of reads mapping to an exon skipping event. (**A**, **B**) Mean *F*1 scores obtained with different depths λ (A) and total number of regulators (B) for the different enrichment approaches. (**C**, **D**) Calibration, measured as the proportion of times the real value lies within the 95%CI of differentially spliced exons and regulatory elements for increasing sequencing depth (C) or increasing number of total regulatory elements (D)

3.4 Model calibration remains robust while decreasing the proportions of active RBP

We further analyzed the performance of dSreg in terms of calibration. A model is well calibrated when inferred probabilities actually represent the real frequency of a given phenomenon i.e. a model is calibrated when the uncertainty of the parameter estimate matches the evidence contained in the data. Calibration was calculated as the proportion of events and regulators whose real change in logit-transformed inclusion rates ($\boldsymbol{\beta}_k$) or activity ($\boldsymbol{\theta}_j$) is within the estimated 95%CI. Whereas changes in inclusion rates were well calibrated, the uncertainty of the changes in the activity of RBPs seemed to be slightly overestimated, given that 95%CI included the real values more often than 95% of the times, independently on the sequencing depth $\boldsymbol{\lambda}$ (Fig. 3C). We then tested how different numbers of total regulatory elements affected model calibration with the previous simulations using only five active out of an increasing number of candidate RBPs. We found that the total number of candidate regulators had no effect on calibration (Fig. 3D). These results suggest that dSreg is conservative when estimating the uncertainty of the regulatory activities $\boldsymbol{\theta}_j$ based on the data, since the real value is within the 95%CI more often than expected across all tested conditions (Fig. 3C and D).

3.5 dSreg outperforms other methods using real data

To assess whether the better performance of dSreg could be confirmed with independent real data, we used an RNA-seq dataset (around 120 M reads per sample) for which a subset of AS events were quantified using RASL-seq and can be used as gold standard (Zhang *et al.*, 2019). We used CLiP-seq data of a number of RBPs binding to upstream and downstream flanks of exon skipping events as regulatory features for dSreg (Dominguez *et al.*, 2018). Since dSreg performed particularly better than other methods at low sequencing depths, we subsampled the sequencing reads by a factor of 2 up to 512 to analyze the extent of this advantage. We analyzed the data also with MISO, BRIE and DARTS. Both BRIE and DARTS use prior information to improve detection of splicing changes (Huang and Sanguinetti, 2017; Katz *et al.*, 2010; Zhang *et al.*, 2019). dSreg and the *Null model* showed the best performance, compared to all other methods, except in extremely low coverages (dilution factor >100), in which DARTS overcame dSreg (Fig. 4A and B). In contrast to the results obtained from the simulated data, dSreg and *Null model* performed similarly, which suggests that the regulatory features that were added do not contribute much to the estimation of AS changes. However, it also shows that it remains robust to the inclusion of non-relevant regulatory features. Neither BRIE nor DARTS outperformed the *Null model*. We observed the same patterns when comparing the results to the full coverage RNA-seq dataset (Supplementary Fig. S1).

The main advantage and motivation of dSreg is the inference of the regulators driving AS changes, a feature that is not provided by any of the existing tools for AS analysis. To assess whether dSreg outperforms ORA and GSEA also with real data, we used the collection of RBP knock-down experiments from ENCODE (Nostrand *et al.*, 2018). Although it is difficult to know the actual regulatory mechanisms in each case, one may reasonably assume that at least some of the AS changes would be mediated by the down-regulation of the target RBP. dSreg detected the highest percentage of knock-down RBPs as regulatory elements compared to the random expectation in each case (Fig. 4C). If the expression of other regulatory element is affected by the perturbation, we would expect them also to contribute to explain AS changes. Regulators detected by dSreg tended to be more often differentially expressed in the same experiment than expected by chance compared to other methods (Fig. 4D). Finally, we observed that, when sorting the regulators by their evidence, the RBP that was knocked-down tended to appear higher in the ranking produced by dSreg than in those yielded by ORA and GSEA (Fig. 4E and F, respectively). Altogether, these results suggest that dSreg also outperforms previous methods in the identification of regulatory elements using real data.

3.6 AS regulation in cardiomyocyte differentiation by core-spliceosomal factors

We then tested our model on a dataset of mouse cardiomyocyte differentiation from cardiac precursors (GSE59383) with three samples per condition as in our simulated scenario. Binding sites for a number of RBPs were obtained from CLiP-seq experiments and only those located in the upstream and downstream intronic flanking 250 bp were used (see Section 2 for details). We run the three approaches explored in this work and found that ORA resulted in a high number of significantly enriched candidates, most of which are likely to represent false positives as in our simulation analysis (Fig. 5A). GSEA, on the other hand, showed no significant enrichment at FDR < 0.05, and only a few at nominal $P < 0.05$, which suggest that these *P*-values can easily arise by chance. Indeed, there is little concordance with results from ORA (Fig. 5A and B). dSreg did show an overall agreement with ORA results, but, as expected, dSreg provided a reduced number of RBPs whose combined action best explain the observed AS changes (Fig. 5 and Supplementary Table S1). Interestingly, a great deal of the identified regulatory RBPs are considered to be members of the core spliceosome (BUD13, EFTUD2, PRPF8, SF3A3, SF3B4), suggesting that changes in the activity of these particular components might be key for the AS changes underlying cardiomyocyte differentiation. In this regard, the core-spliceosomal machinery has been shown to have extensive regulatory potential (Papasaikas *et al.*, 2015) and mutations in one of these genes (EFTUD2) have been associated with congenital heart defects, among other phenotypes (Lines *et al.*, 2012).

4 Discussion

Here we present dSreg, a new method that integrates the analysis of differential AS and the identification of the underlying regulatory mechanisms in a single model. Our single-step model bypasses the need to call for differential splicing before enrichment and therefore improves sensitivity, especially at low sequencing depths. It also increases specificity as it uses information from the underlying changes in RBPs activity to avoid false positives derived from small sample sizes. Moreover, dSreg analyzes the regulatory activity all

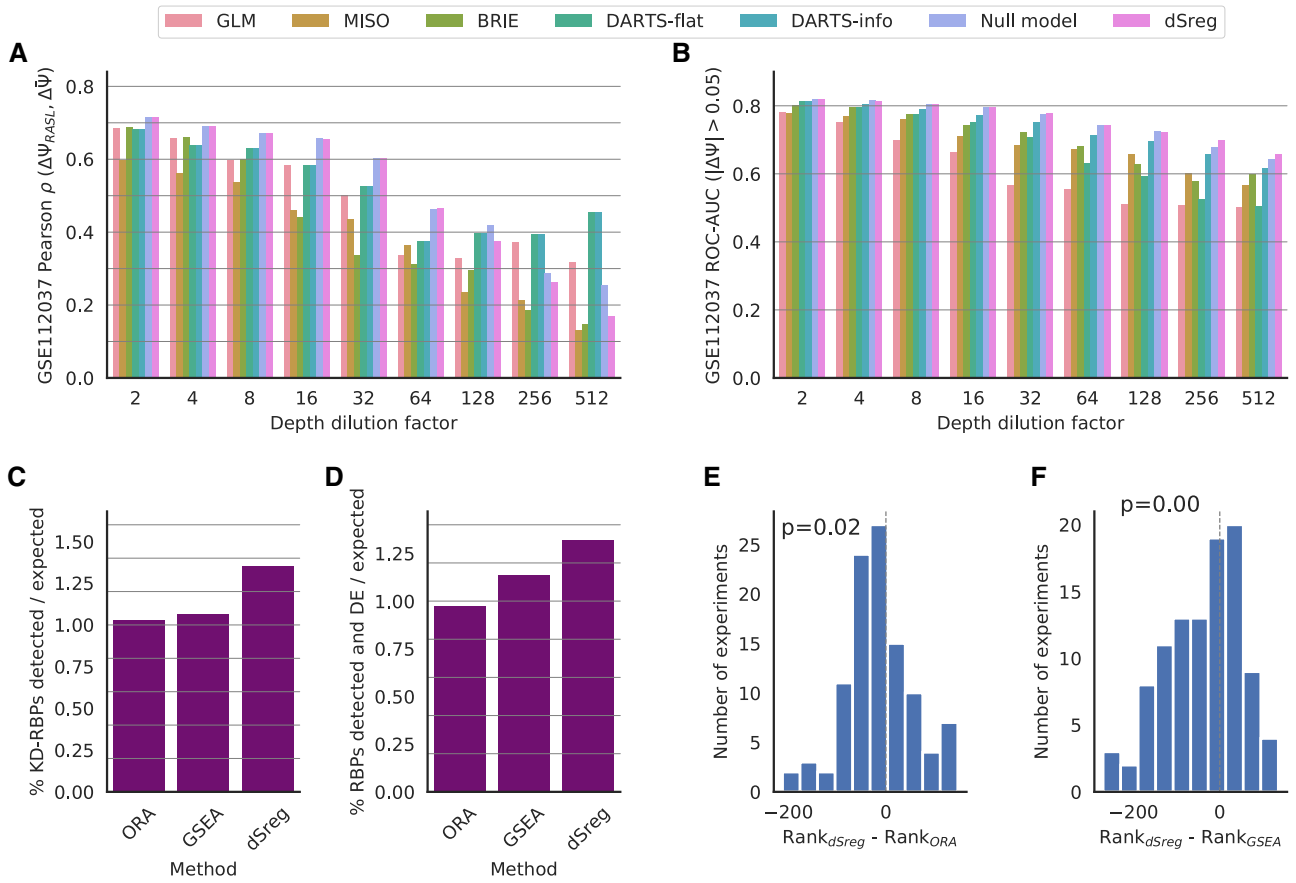

Fig. 4. Evaluation of the performance of dSreg with other methods on real data. (**A, B**) Performance of differential splicing methods using RASL-seq quantification as true values, measured by Pearson correlation of ΔΨ (A) and AUROC for exons significantly changed, defined as those with a $|\Delta\Psi| > 0.05$. Methods include a GLM. MISO, BRIE, DARTS with and without using the predictions as prior (info and flat, respectively) and dSreg and its *Null model*. (**C**) Percentage of experiments in which the knocked-down RBP was found among the regulatory elements compared to expectation. (**D**) Percentage of regulatory RBPs identified by each method that were detected to be differentially expressed compared to expectation. Expectations were calculated by 20 000 random sampling of the same number of regulators. (**E, F**) Difference in rank occupied by the knocked-down RBP in the output of dSreg with that of ORA (E) and GSEA (F)

RBPs simultaneously to correct for possible co-linearities in the binding profiles and uses a horseshoe prior to force most of the RBPs activities to remain unchanged. Joint modeling also provides higher specificity in the detection of regulatory mechanisms as it reduces the number of false positives due to co-occurrence of binding sites of different RBPs, leading to an improved overall performance compared with classical enrichment approaches for regulatory elements. Our model opens the possibility to analyze AS more accurately using RNA-seq data with low sequencing depth, both for reanalysis of previously sequenced samples or for more cost-effective new RNA-seq experiments with focus on the regulatory mechanisms. Although transcript-based methods also lower the requirements on sequencing depths (Alamancos *et al.*, 2014; Trincado *et al.*, 2018), our model works directly at the event level, reducing the dependency on the transcript annotation (Zhao and Zhang, 2015). In contrast to previous approaches, including Bayesian methods like MISO (Katz *et al.*, 2010), our model is motivated by how splicing changes are regulated between two biological conditions rather than on how inclusion and skipping reads are generated from the inclusion rate (Ψ) in a particular sample. Still, we show that not only we gain more biological insight directly from the model, but also obtain, at least, as good estimations of AS changes as provided by the best performing tools to date. dSreg still requires a previous definition of alternative mapping events and allocation of reads to inclusion or skipping isoforms, making it compatible with any of the software used in this paper.

Our good results on simulations are, however, restricted to those cases in which AS changes are mediated only by a subset of differentially active RBPs binding to known sites. Although inclusion of a high number of RBPs showed no effect on the changes in inclusion rates between the two tested conditions, alternative sources of errors, such as errors in the binding profiles or missing information might have a negative impact on the sensitivity of dSreg. Indeed, the improvement of dSreg on real data compared with the *Null model* is rather small, if any. The better performance of both models compared with existing methods seems therefore due to the inclusion of a parameter describing variance between samples across exons rather than to the regulatory information. Similarly, other methods including event features to improve detection of splicing changes do not outperform our *Null model* (Huang and Sanguinetti, 2017; Zhang *et al.*, 2019), except when data are very scarce. Whereas dSreg only uses AS data from the target experiment, DARTS informative prior was trained with many other datasets such that, in absence of information, is able to make relatively good predictions about the outcome of an experiment given the regulatory features. These results suggest that only a small part of splicing variation is mediated by RBPs CLiP-seq binding sites as in the model. This was not unexpected, since previous studies suggest that AS regulation is far more complex than a sum of effects of a number of RBPs and that RNA structure plays a critical role (Barash *et al.*, 2010; Leung *et al.*, 2014; Taliaferro *et al.*, 2016). In spite of these limitations, dSreg was able to detect the knocked-down and differentially expressed RBPs in loss of function models more efficiently than traditional approaches like ORA and GSEA, suggesting that the identified regulation was real. Yet, we expect that careful modeling of additional AS regulatory features will improve the results, e.g. nucleosome positioning and histone modifications (Iannone *et al.*, 2015; Luco *et al.*, 2010; Merkin *et al.*, 2015). Moreover, dSreg is

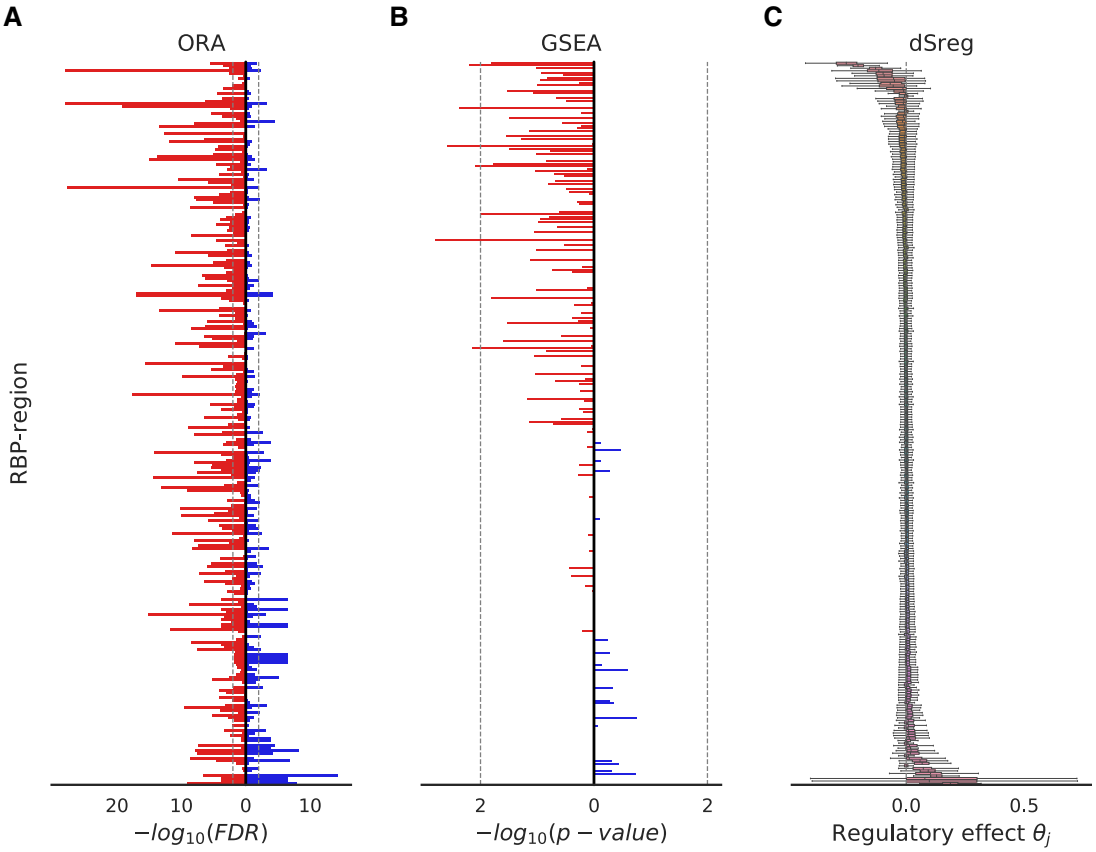

Fig. 5. Comparison of ORA, GSEA and dSreg using a real RNA-seq dataset from a cardiomyocyte differentiation experiment. RBPs on the *y*-axis are sorted for the three panels according to the posterior mean of the regulatory effect θ_j inferred by dSreg. (**A**) Candidate regulatory proteins derived from the ORA on the significantly included (blue) or skipped (red) exons represented by their significance expressed as the log transformation of the FDR. (**B**) GSEA results represented by the nominal empirical *P*-value resulting from permuting the exon labels. RBPs with positive enrichment scores are represented on the right, and those with negative scores on the left. (**C**) Posterior distributions of the regulatory effects θ_j inferred by our model. (Color version of this figure is available at *Bioinformatics* online.)

limited so far to pairwise comparisons, whereas we are often interested in analyzing enrichment over a number of conditions, such as time series and dose-response experiments. Further work will be necessary to allow for more complex and powerful experimental designs.

5 Conclusion

Our model provides an example of how joint modeling of interdependent phenomena can improve results compared with completely separated analysis relying on categorization according to rather arbitrary thresholds. Bayesian inference through Markov Chain Monte Carlo methods provides a general framework to fit very flexible models that adapt to each particular analysis and to easily extend currently existing models to integrate different sources of information. In our case, we only integrated binding sites information with AS data, but these models are flexible enough to include information about expression of AS regulators, post-transcriptional modifications or any other piece of information supporting a change in the activity of a particular regulatory protein. This model is not only limited to regulation analysis, but can also be used with functional annotations, such as the presence of functional domains, phosphorylation sites, protein–protein interaction motifs or any other property that may be associated with AS. Moreover, we have implemented the model in dSreg (https://bitbucket.org/cmartiga/pydsreg/src/master/), which enables running the model using only the matrices of inclusion and total number of reads per event and a matrix *S* with the event features e.g. the binding sites. Therefore, dSreg adds a valuable statistical tool to existing software aimed at identifying AS events, such as rMATS or vast-tools (Irimia *et al.*, 2014; Shen *et al.*, 2012), among others, for more accurate detection of AS regulatory mechanisms using RNA-seq data.

Acknowledgements

We would like to thank Victor Jimenez for critical reading and useful discussions about the manuscript and beyond. We also thank Eric Van Nostrand for his help to retrieve the processed ENCODE data.

Funding

This work was supported by grants from the European Union [CardioNeT-ITN-289600, CardioNext-608027]; the Spanish Ministry of Economy and Competitiveness [SAF2015-65722-R, SAF2012-31451]; the Spanish Ministry of Science, Innovation and Universities [RTI2018-102084-B-I00]; the Instituto de salud Carlos III (ISCIII) [CPII14/00027, RD012/0042/0066]; and the Madrid Regional Government [2010-BMD-2321 'Fibroteam']. The study also received support from the Plan Estatal de I + D+I 2013-2016 - European Regional Development Fund (ERDF) 'A way of making Europe', Spain. The CNIC is supported by the Instituto de Salud Carlos III (ISCIII), the Ministerio de Ciencia, Innovación y Universidades (MCNU) and the Pro CNIC Foundation, and is a Severo Ochoa Center of Excellence (SEV-2015-0505).

Conflict of Interest: none declared.

References

Alamancos,G.P. *et al.* (2015) Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA*, **21**, 1521–1531.

Alipanahi,B. *et al*. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol*., **33**, 831–838.
Anders,S. *et al*. (2012) Detecting differential usage of exons from RNA-Seq data. *Genome Res*., **22**, 2008–2017.
Auinash,K. and Cooper,T.A. (2012) Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet*., **12**, 715–729.
Baralle,F.E. and Giudice,J. (2017) Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol*., **18**, 437–451.
Barash,Y. *et al*. (2010) Deciphering the splicing code. *Nature*, **465**, 53–59.
Barbosa-Morais,N.L. *et al*. (2012) Research articles. *Science*, **338**, 1587–1594.
Betancourt,M.J. and Girolami,M. (2013) Hamiltonian Monte Carlo for hierarchical models. *Arxiv*, 1–11.
Bray,N.L. *et al*. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol*., **34**, 525–527.
Carpenter,B. *et al*. (2017) *Stan*: a probabilistic programming language. *J. Stat. Softw*., **76**, 1–32.
Carvalho,C.M. *et al*. (2009) Handling sparsity via the horseshoe. *J. Mach. Learn. Res*., **5**, 73–80.
Climente-GonzáLez,H. *et al*. (2017) The functional impact of alternative splicing in cancer. *Cell Rep*., **20**, 2215–2226.
Dobin,A. *et al*. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
Dominguez,D. *et al*. (2018) Sequence, structure and context preferences of human RNA binding proteins. *Mol. Cell*, **70**, 854–857.
Fogel,B.L. *et al*. (2012) RBFOX1 regulates both splicing and transcriptional networks in human neuronal development. *Hum. Mol. Genet*., **21**, 4171–4186.
Giudice,G. *et al*. (2016) ATtRACT-a database of RNA-binding proteins and associated motifs. *Database*, **2016**, baw035.
Giudice,J. *et al*. (2014) Alternative splicing regulates vesicular trafficking genes in cardiomyocytes during postnatal heart development. *Nat. Commun*., **5**, 3603.
Graveley,B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet*., **17**, 100–107.
Huang,Y. and Sanguinetti,G. (2017) BRIE: transcriptome-wide splicing quantification in single cells. *Genome Biol*., **18**, 123.
Iannone,C. *et al*. (2015) Relationship between nucleosome positioning and progesterone-induced alternative splicing in breast cancer cells. *RNA*, **21**, 360–374.
Irimia,M. *et al*. (2014) A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell*, **159**, 1511–1523.
Katz,Y. *et al*. (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods*, **7**, 1009–1015.
Kim,E. *et al*. (2007) Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res*., **35**, 125–131.
Lara-Pezzi,E. *et al*. (2013) The alternative heart: impact of alternative splicing in heart disease. *J. Cardiovasc. Transl. Res*., **6**, 945–955.
Lara-Pezzi,E. *et al*. (2017) Neurogenesis: regulation by alternative splicing and related posttranscriptional processes. *Neuroscientist*, **23**, 466–477.
Lee,J.A. *et al*. (2016) Cytoplasmic Rbfox1 regulates the expression of synaptic and autism-related genes. *Neuron*, **89**, 113–128.
Leung,M.K.K. *et al*. (2014) Deep learning of the tissue-regulated splicing code. *Bioinformatics*, **30**, i121–i129.
Lines,M.A. *et al*. (2012) Haploinsufficiency of a spliceosomal GTPase encoded by EFTUD2 causes mandibulofacial dysostosis with microcephaly. *Am. J. Hum. Genet*., **90**, 369–377.
Luco,R.F. *et al*. (2010) Regulation of alternative splicing by histone modifications. *Science*, **327**, 996–1000.
Maticzka,D. *et al*. (2014) GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol*., **15**, R17.
Merkin,J. *et al*. (2012) Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*, **338**, 1593–1600.
Merkin,J. *et al*. (2015) Origins and impacts of new mammalian exons. *Cell Rep*., **10**, 1992–2005.
Nilsen,T.W. and Graveley,B.R. (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature*, **463**, 457–463.
Nostrand,E.L.V. *et al*. (2018) A large-scale binding and functional map of human RNA binding proteins. *bioRxiv*, 1–111. doi: 10.1101/179648.
Papasaikas,P. *et al*. (2015) Functional splicing network reveals extensive regulatory potential of the core spliceosomal machinery. *Mol. Cell*, **57**, 7–22.
Patro,R. *et al*. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, **14**, 417–419.
Pertea,M. *et al*. (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc*., **11**, 1650–1667.
Piironen,J. and Vehtari,A. (2017) Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electron. J. Stat*., **11**, 5018–5051.
Quesnel-Vallières,M. *et al*. (2015) Essential roles for the splicing regulator nSR100/SRRM4 during nervous system development. *Genes Dev*., **29**, 746–759.
Raj,B. *et al*. (2014) A global regulatory mechanism for activating an exon network required for neurogenesis. *Mol. Cell*, **56**, 90–103.
Ray,D. *et al*. (2013) A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, **499**, 172–177.
Sebestyén,E. *et al*. (2016) Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome Res*., **26**, 732–744.
Shen,S. *et al*. (2012) MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Res*., **40**, e61.
Simillion,C. *et al*. (2017) Avoiding the pitfalls of gene set enrichment analysis with SetRank. *BMC Bioinformatics*, **18**, 151.
Stricker,T.P. *et al*. (2017) Robust stratification of breast cancer subtypes using differential patterns of transcript isoform expression. *PLoS Genet*., **13**, e1006589.
Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.
Taliaferro,J.M. *et al*. (2016) RNA sequence context effects measured in vitro predict in vivo protein binding and regulation. *Mol. Cell*, **64**, 294–306.
Trapnell,C. *et al*. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol*., **31**, 46–53.
Trincado,J.L. *et al*. (2018) SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol*., **19**, 40.
Wagnon,J.L. *et al*. (2012) CELF4 regulates translation and local abundance of a vast set of mRNAs, including genes associated with regulation of synaptic function. *PLoS Genet*., **8**, e1003067.
Weyn-Vanhentenryck,S.M. *et al*. (2018) Precise temporal regulation of alternative splicing during neural development. *Nat. Commun*., **9**, 2189.
Yang,Y.-C.T. *et al*. (2015) CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics*, **16**, 51.
Zhang,Z. *et al*. (2019) Deep-learning augmented RNA-seq analysis of transcript splicing. *Nat. Methods*, **16**, 307–310.
Zhao,S. and Zhang,B. (2015) A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC Genomics*, **16**, 97.