dSreg: A bayesian model to integrate changes in splicing and RNA binding protein activity

Carlos Martí-Gómez[1], Enrique Lara-Pezzi*[1], and Fátima Sánchez-Cabo*[1]

[1]Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC), Melchor Fernández Almagro 3, Madrid, Spain

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. 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 (RBPs). 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 alternative splicing changes in simulated data, even at low read coverage. dSreg also showed improved performance when analyzing a collection of knock-down RBPs experiments from ENCODE, as opposed to traditional enrichment methods such as Over-representation Analysis (ORA) and Gene Set Enrichment Analysis (GSEA). 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. dSreg was implemented in python using stan and is freely available to the community at https://bitbucket.org/cmartiga/dsreg.

Bayesian models | Alternative splicing | RNA binding proteins | Splicing regulation

Correspondence: ***elara@cnic.es, fscabo@cnic.es***

Introduction

Eukaryotic genes are generally constituted by exons and introns (25). 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) (19, 36). There is evidence of AS for most mammalian genes (35, 38) and of widespread changes of AS patterns throughout brain and heart development (4, 5, 15, 18, 23, 42, 43, 56). Defects in mRNA processing of some specific genes often lead to disease (5, 26, 27) and have been associated with complex neurological disorders, such as autistic syndrome (23, 28, 42, 55), and cancer (12, 49). 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 work-flow is generally applied to identify the regulatory mechanisms underlying the changes in AS (see Figure 1 for a schematic representation). First, AS changes must be identified. For this, short reads from RNA sequencing are typically mapped using splice junctions (SJ) aware aligners such as STAR or Hisat2 (13, 40). Alternative mRNA processing can be studied at two different levels: 1) transcript quantification level, which can be based on a prior alignment (40, 52), or can be directly estimated from fast pseudoalignment methods (9, 39); and 2) event level quantification, as performed by popular tools such as MISO, MATS, vast-tools, DEXseq or SUPPA (1, 3, 23, 24, 46, 53). Recent toolsshowed improved performance for the estimation of AS changes by using information about exon features and perturbation experiments (21, 58). 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 (46).

The second step aims to identify regulatory features, such as RBPs motifs, associated with AS changes. Such features often include nucleotide hexamers, predicted motifs, experimentally determined or predicted binding sites (14, 17, 44, 57). Over-representation Analysis (ORA) enables the discovery of features co-ocurring 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) (48, 50). Although these tools were designed for functional analysis, they have been used to perform enrichment of known targets of regulatory elements (45, 53). 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 consid-

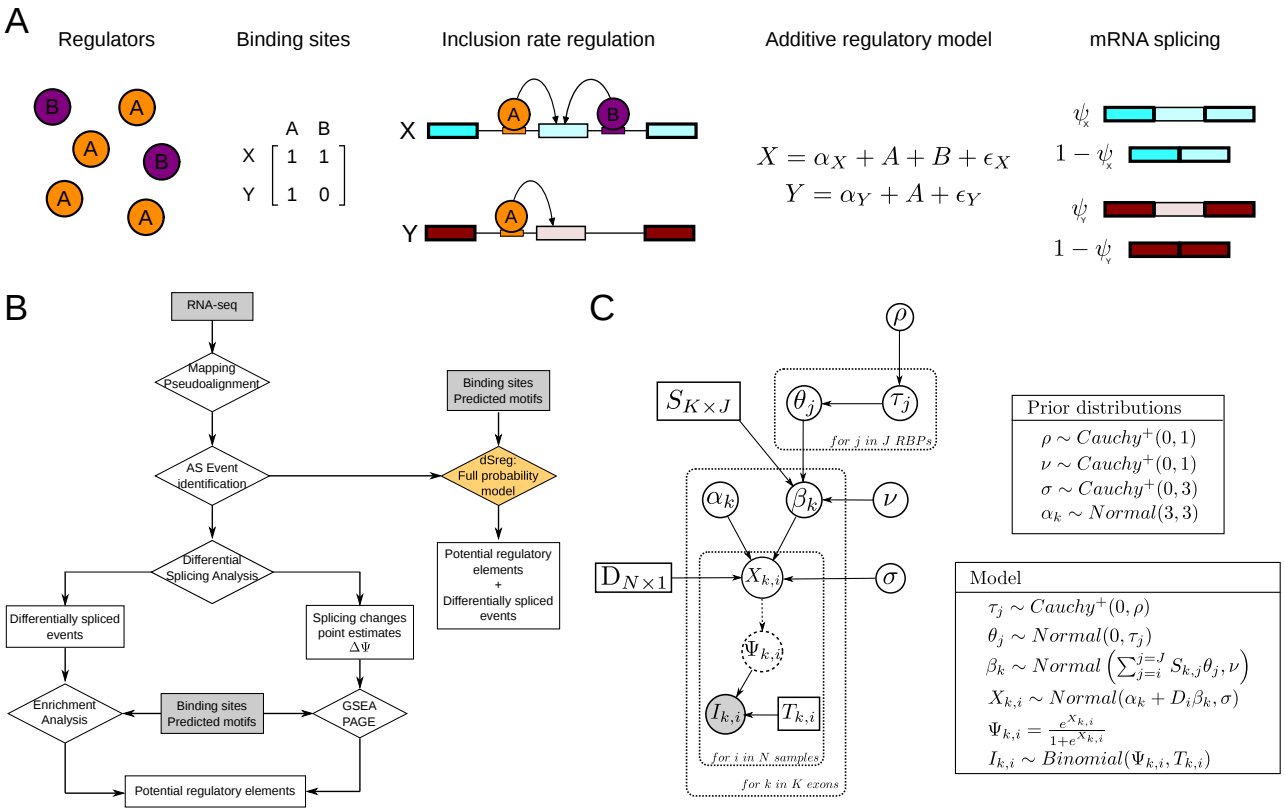

Fig. 1. General and proposed work-flows for AS regulation analysis. **A** Schematic representation of idealized model for the regulation of splicing rates by RBPs through direct binding to their binding motifs in the pre-mRNA **B**. Diagram representing the different steps required for a classical analysis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

ered, 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 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.

Methods

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.

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

$\Psi_{k,i}$ 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 $X_{k,i}$,

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

We assume that $X_{k,i}$ is drawn from a normal distribution with a common standard deviation σ_k and different means per condition: α_k for condition a; and $\alpha_k + \beta_k$ for condition b, such that β_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 ($\sigma_k = \sigma$) .

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

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 β_k, should depend on the change in the activity θ_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 (PWMs) (44) or any other predictive tool (2, 33).

$$S_{k,j} = \begin{cases} 1 & \text{if the RBP } j \text{ binds to event } k \\ 0 & \text{otherwise} \end{cases}$$

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

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

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 θ_j(11). The horseshoe prior, a member of the family of hierarchical shrinkage priors, specifies a normal prior for θ_j with mean 0 and a standard deviation τ_j, where τ_j is not a fixed value, but drawn from a common half Cauchy distribution with mean 0 and ρ standard deviation. τ_j represents a local shrinkage parameter, as it only affects protein j, whereas ρ can be understood as a global shrinkage parameter. We further set a half Cauchy prior in ρ with mean 0 and standard deviation 1 as recommended (11). Note that this prior can be adapted according to the expected number of non-zero parameters (41).

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

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

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

Finally, we need to specify prior distributions for the remaining parameters α_k and σ. Since we expect most of the exons to be included most of the times ($\Psi \sim 1$) and α_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 σ.

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

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

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

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

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.

$$p(\Theta,\mathbf{I}) = \quad \textbf{(11)}$$

$$= p(\mathbf{I},\ \mathbf{T},\ X,\alpha,\beta,\nu,\theta,\tau,\rho,D,S) =$$

$$= p(\sigma)p(\nu)p(\rho)\prod_{j}^{J}[p(\theta_j|\tau_j)p(\tau_k|\rho)]\prod_{k}^{K}[p(\beta_k|\mathbf{S},\theta,\nu)p(\alpha_k)P(I_k)]$$

where,

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

Once the full posterior distribution is completely specified, it can be explored using Markov Chain Monte Carlo (MCMC) algorithms. We implemented this model in stan (10), using a non-centered parametrization whenever possible to alleviate sampling difficulties from hierarchical models (7). The full model is represented as a Directed Acyclic Graph (DAG) to show dependencies among parameters in Fig. 1B.

Data simulation.

Data can be simulated by setting fixed values on the parent nodes of the DAG (σ, $\vec{\alpha}$, $\vec{\theta}$, $\mathbf{T}$ and $\mathbf{S}$) representing the probabilistic model (Fig. 1C) and drawing samples from the corresponding distributions for each parameter. We simulated 20 datasets for each initial set of conditions, all with K=2000 events, 3 samples per condition (N=6) and J=50 potential regulatory elements with correlated binding profiles, of which

only 5 showed non-zero effects on splicing changes between the two conditions.

To simulate realistic values of inclusion rates for the condition a ($\Psi_{k,a}$) across the K=2000 exons, we assumed that 20% of the exons are alternative, with inclusion rates following a uniform distribution between 0 and 1; and 80% are consitutive, with inclusion rates drawn from a Beta(10, 1), to promote generally high inclusion rates.

$$u_k \sim Uniform(0,1) \tag{13}$$

$$\Psi_{k,a} \sim \begin{cases} Beta(10,1)\ if\ u_k > 0.2 \\ Uniform(0,1)\ if\ u_k < 0.2 \end{cases} \tag{14}$$

$$\alpha_k = logit(\Psi_{k,a}) = log\left(\frac{\Psi_{k,a}}{1-\Psi_{k,a}}\right) \tag{15}$$

We aimed to simulate matrices of correlated binding profiles to take into account that certain groups of RBPs often bind to similar regions in the exons. To do so, we first simulated a covariance matrix Σ of size J sampling from an inverse Wishart distribution,

$$\Sigma_{J\times J} \sim InvWishart\left(J+1, \frac{1}{J}\mathbb{I}_J\right) \tag{16}$$

and used it to simulate K samples from a multivariate normal distribution using a mean of -2.5. This value represents an expected 7.5% of events bound by a particular RBP.

$$\vec{M}_k \sim MvNormal(-2.5, \Sigma) \tag{17}$$

Then, we took the inverse logit to transform $\mathbf{M}$ matrix into the probability matrix $\mathbf{P}$ and use these probabilities to simulate binary binding profiles across exons ($\mathbf{S}_{K\times J}$ matrix) by sampling from a Bernoulli distribution for each element in the $\mathbf{P}_{K\times J}$ matrix.

$$P_{k,j} = InvLogit(M_{k,j}) = \frac{e^{M_{k,j}}}{1+e^{M_{k,j}}} \tag{18}$$

$$S_{k,j} \sim Bernoulli(P_{k,j}) \tag{19}$$

We randomly drew a set $A = \{A_1, A_2, A_3, A_4, A_5\}$ of 5 active regulatory proteins (with non-zero effects on changes in the inclusion rates) from the whole set of regulatory proteins $R = \{1, 2, ..., J\}$. The regulatory effect for RBP j θ_j was then drawn from a uniform distribution between -2.5 and 2.5 if j belonged to the set of active regulatory elements A and set to zero otherwise. These values of θ_j represent the mean increase in the log(odds ratio) of exons having a binding site for that protein compared with those without a binding site.

$$\theta_j \sim \begin{cases} Uniform(-2.5, 2.5)\ if\ j \in A \\ 0\ otherwise \end{cases} \tag{20}$$

Once the parent nodes of the DAG were simulated, we could easily simulate the final data by sampling parameter values along the graph according to our model. First, we drew changes in the logit-transformed inclusion rates β_k from a normal distribution with mean obtained from a linear combination of effects $\vec{\theta}$ and binding sites $\vec{S_k}$ and standard deviation $\nu = 0.1$. This way we introduced noise with small random changes in inclusion rates of exons that were not targets of any of the differentially active RBP.

$$\beta_k \sim Normal\left(\sum_{j=1}^{j=J} S_{k,j}\theta_j,\ \nu\right) \tag{21}$$

We then combined α_k and β_k to obtain the mean $logit(\Psi)$ for condition b, and sample 3 samples from each mean using $\sigma = 0.2$ to introduce some inter-individual variability. Being D_i a variable that takes value 1 when sample i belongs to condition b and 0 otherwise,

$$X_{k,i} \sim Normal(\alpha_k + D_{k,i}\beta_k, \sigma) \tag{22}$$

The total number of reads mapping to each event $T_{k,i}$ were drawn from a Poisson distribution with $log(\lambda) = 2$ by default,

$$T_{k,i} \sim Poisson(\lambda) \tag{23}$$

They were subsequently used to sample the corresponding reads supporting inclusion $I_{k,i}$ from the binomial distribution with $p = \Psi_{k,i}$, obtained from the inverse logit transformation of $X_{k,i}$.

$$I_{k,i} \sim Binomial\left(T_{k,i}, InvLogit(X_{k,i})\right) \tag{24}$$

Using these default parameter values, we additionally simulated data for increasing sequencing depths (from $log(\lambda) = 1$ to $log(\lambda) = 5.5$) and with an increasing number of total regulatory proteins (from J=50 to J=250), maintaining a total of 5 differentially active RBPs to evaluate the effect of this variables on the methods performance.

Bayesian inference.

The probabilistic models were implemented in Stan (10) using non-centered parametrization, whenever it was possible, to improve sampling efficiency (7). The joint posterior distributions of the parameters were approximated using No-U Turn Sampler (NUTS) as implemented in Stan (20), running 4 chains along 4000 iterations, being 2000 of them for warming up. Convergence of the Markov Chain Monte Carlo (MCMC) algorithm was checked in each case by means of the split Gelman-Rubin R ($\hat{R}$) (16).

Differential splicing analysis.

In order to identify exons with significant changes in inclusion rates, a GLM with binomial likelihood was used to model the probability of inclusion of a particular exon using the sample condition D_i as only predictor. After fitting the model, we extracted the estimate and p-value for the coefficient representing the condition of interest. We then obtained adjusted p-values by means of Benjamini-Hochberg (BH) multiple test correction.

Over-representation Analysis (ORA).

We tested over-representation of binding sites for a particular RBP on the set of significantly changed exons using a Generalized Linear Model (GLM) with binomial likelihood to model the probability of being significantly changed as a function of the presence of a binding site for a particular RBP. We then extracted the p-value for the coefficient for each RBP and applied BH multiple test correction.

Gene Set Enrichment Analysis (GSEA).

We implemented an in-house algorithm for GSEA in python following (50). We sorted exons according to the estimated coefficient representing log-transformation of change in exon inclusion odds between the two conditions under study. We then used the matrix with binding sites for each exon and RBP and subtracted the mean for each column. This way, we give weight to each binding site depending on the number of binding sites present for a particular RBP. We then calculated the cumulative sum and took the maximum and minimum values as enrichment scores. We permuted 10000 times the list of exons to calculate a null distribution of enrichment scores, estimated p-values as the proportion of permutations with bigger enrichment scores and performed BH multiple test correction.

Regulatory features: CLiP-seq derived RBPs binding sites.

CLiP-seq binding sites were collected from several databases and merged in a single BED file (8, 14, 30, 57). Human binding sites and mouse binding sites in mm10 were transformed to mm9 coordinates using liftOver tool for compatibility with vast-tools. For simplicity, only binding sites mapping to the 250bp upstream or downstream the alternative exons were included in the analyses.

Bench-marking of differential splicing methods using real data.

In order to assess the performance of dSreg in real biological data, we used the GSE112037 dataset, which contained an independent quantification of exon inclusion rates using RASL-seq for the quantitative evaluation of the performance of different methods (58). We also evaluated the impact of sequencing depth on the performance of the different methods by serial down-sampling of sequencing up to 1/512 times the original depths (120M reads). dSreg was run using processed event counts as provided by DARTS, which is itself based on rMATS (47, 58). GLM analysis was also performed using the same event counts. MISO and BRIE were run using their own event annotation, corresponding to hg19 genome version and Ensembl annotation release 75 for all methods (21, 24). An additional $Nullmodel$ for dSreg without regulatory information, as in the simulations, was run to test the improvement in detection of splicing changes by including regulatory features. For evaluation, we selected events with at least 50 total reads in the RASL-seq experiment, and calculated the real inclusion rates as the proportion of reads supporting exon inclusion. Real AS changes were defined as those with a $|\Delta\Psi| > 0.05$ and $FDR < 0.05$ using a basic GLM in R. Then, performance was evaluated by comparing the estimation of the $\Delta\Psi$ in the down-sampled RNA-seq experiments and the ones derived from RASL-seq. We assessed the quantitative estimation of inclusion rates by calculating the Pearson coefficient with the real $\Delta\Psi$. AUROC was used to asses the ability to identify differentially spliced. The scoring function for AUROC calculation were: i) Bayes Factors for BRIE and MISO: 1 - FDR for GLM; ii) and $P(|\Delta\Psi| > 0.05|data)$ for DARTS and dSreg; and iii) MISO and BRIE were evaluated using only the subset of events that were also represented in the RASL-seq experiments.

Assessment of the ability of dSreg to identify AS regulatory drivers using ENCODE knock-down experiments.

In order to evaluate the performance of dSreg in detecting the RBPs that drive AS changes between two conditions, we used the data from systematic knock-down experiments of 206 RBPs in two different human cell lines from the ENCODE project and their corresponding binding profiles (14, 37). We downloaded the rMATS processed files available from the website and analyzed their regulatory patterns using GLM+ORA, GLM+GSEA and dSreg. Regulators were defined as differentially active if FDR<0.05 for both ORA and GSEA; or if the posterior probability of the θ_j being different from 0 was higher than 95% ($P(|\theta_j| > 0|data) > 0.95$) for dSreg. The performance was evaluated with 3 different measures. First, we analyzed the number of times the RBP that was down-regulated was found among the driver regulators of AS. This measure was normalized by the expected proportion of matches if the regulatory elements were selected randomly from the set of available regulators. Second, we measured the proportion of RBPs defined as differentially active were differentially expressed in in the knock-down experiment. Third, we sorted by absolute differential activity (or FDR for ORA and GSEA) the RBPs and calculated of RBP that was knocked-down.

Real data analysis.

GSE59383 fastq data were downloaded and mapped using vast-tools 0.2.0 (23) to identify AS events. We restricted our analysis to exon cassette events that showed at least 1 inclusion and skipping read in at least one sample. Once extracted the number of inclusion and total counts for each event and sample, we used all the methods described here (ORA, GSEA and dSreg) to analyze regulatory patterns using a compendium of CLiP-seq binding sites.

Results

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 (λ). Correlation between the estimated $\hat{\beta}_k$

and the real β_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 λ 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 ($log(\lambda) < 3$), 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 as 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 β_j 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 F1 scores and AUROC, suggesting that differences in performance are intrinsic to the method and not threshold dependent (Fig. 2C,D and 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 3 samples per condition.

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 F1 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 F1 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 (MLE) 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 (50) (see Methods section for details). We found a substantial improvement over ORA, with higher F1 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.

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 5 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 F1 score tended to decrease as the number of potential regulators increased with either ORA or GSEA, despite multiple test correction to control false discovery rate. Once more, dSreg outperformed both methods and remained unaffected by the inclusion of other inactive regulatory elements (Fig. 3B).

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 phenomena 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 (β_k) or activity (θ_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 λ (Fig. 3C). We then tested how different numbers of total regulatory elements affected model calibration with the previous simulations using only 5 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 θ_j based on the data, since the real value is

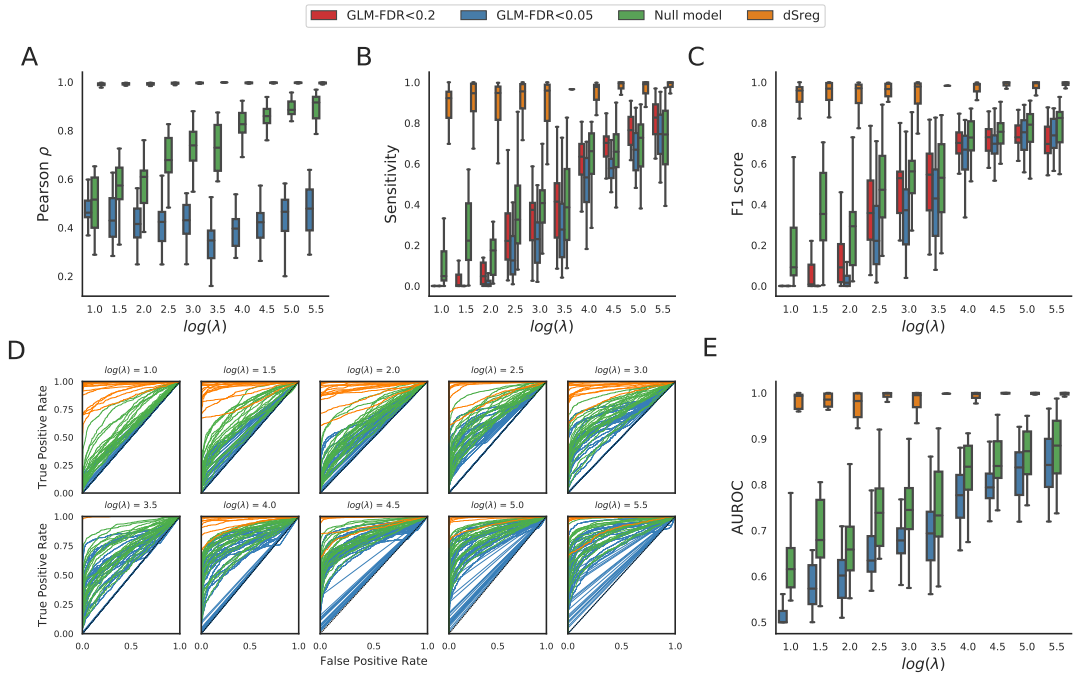

Fig. 2. Comparison of the performance for the identification of different event inclusion rates of a standard method using a single GLM per exon considering two FDR thresholds (0.05 and 0.2),a bayesian model that pools variance across all exons ($Null\ model$) and dSreg. Performance was analyzed in simulations 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** F1 score. **D, E** Receiver Operating Characteristic (ROC) curves (D) and the area under them (E)

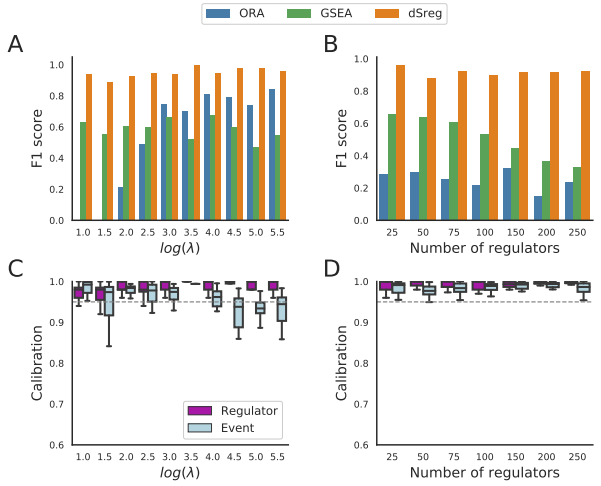

Fig. 3. Performance of methods for the detection of regulatory elements: ORA with variable FDR thresholds (0.05 and 0.2), non-parametric GSEA and dSreg. Performance was analyzed in simulations with increasing sequencing depths λ, which is the mean of the Poisson distribution used to simulate the total number of reads mapping to an exon skipping event. **A, B.** Mean F1 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).

within the 95%CI more often than expected across all tested conditions (Fig. 3C,D).

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 120M reads per sample) for which a subset of AS events were quantified using RASL-seq and can be used as gold standard (58). 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 (14). 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 (21, 24, 58). 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 (Figure 4A,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 (Figure 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 (37). 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 (Figure 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 (Figure 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 (Figure 4E and F, respectively). Altogether, these results suggest that dSreg also outperforms previous methods in the identification of regulatory elements using real data.

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 3 samples per condition as in our simulated scenario. Binding sites for a number of RNA binding proteins were obtained from CLiP-seq experiments and only those located in the upstream and downstream intronic flanking 250bp were used (see Extended Methods section for details). We run the 3 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, 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 (54) and mutations in one of these genes (EFTUD2) have been associated with congenital heart defects, among other phenotypes (31).

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 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 (1, 53), our model works directly at the event level, reducing the dependency on the transcript annotation (59). In contrast to previous approaches, including bayesian methods like MISO (24), 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 $Null\ model$ (21, 58), except when data is 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

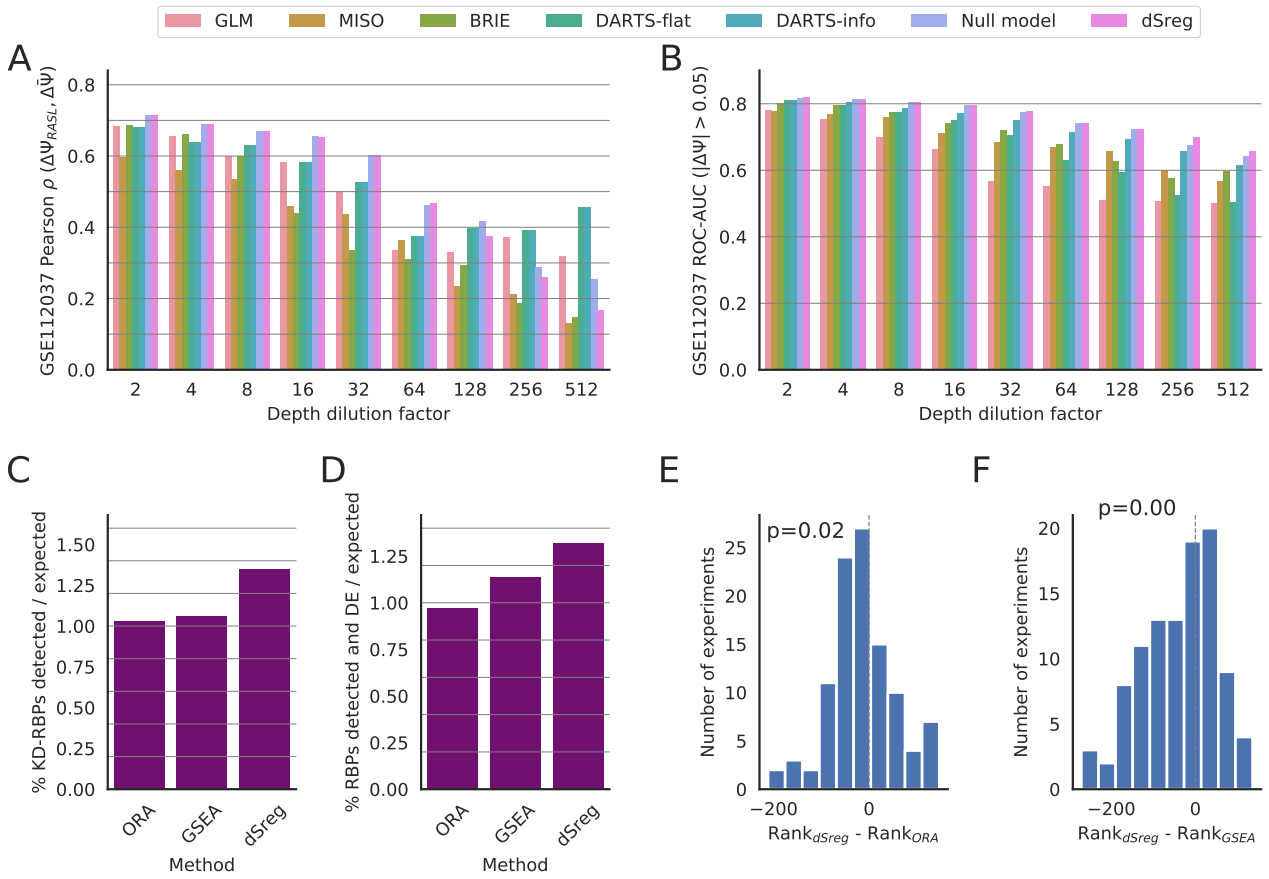

Fig. 4. Evaluation of the performance of dSreg with other methods on real data. **A,B** Performance of differential splicing methods using RASL-seq quantification as true values, measured by Pearson correlation of $\Delta\Psi$ (A) and Area Under the ROC (AUROC) for exons significantly changed, defined 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 (DE) compared to expectation. Expectations were calculated by 20000 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).

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 (6, 29, 51). 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 (22, 32, 34). Moreover, dSreg is 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.

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 MCMC 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 (23, 46), among others, for more accurate detection of AS

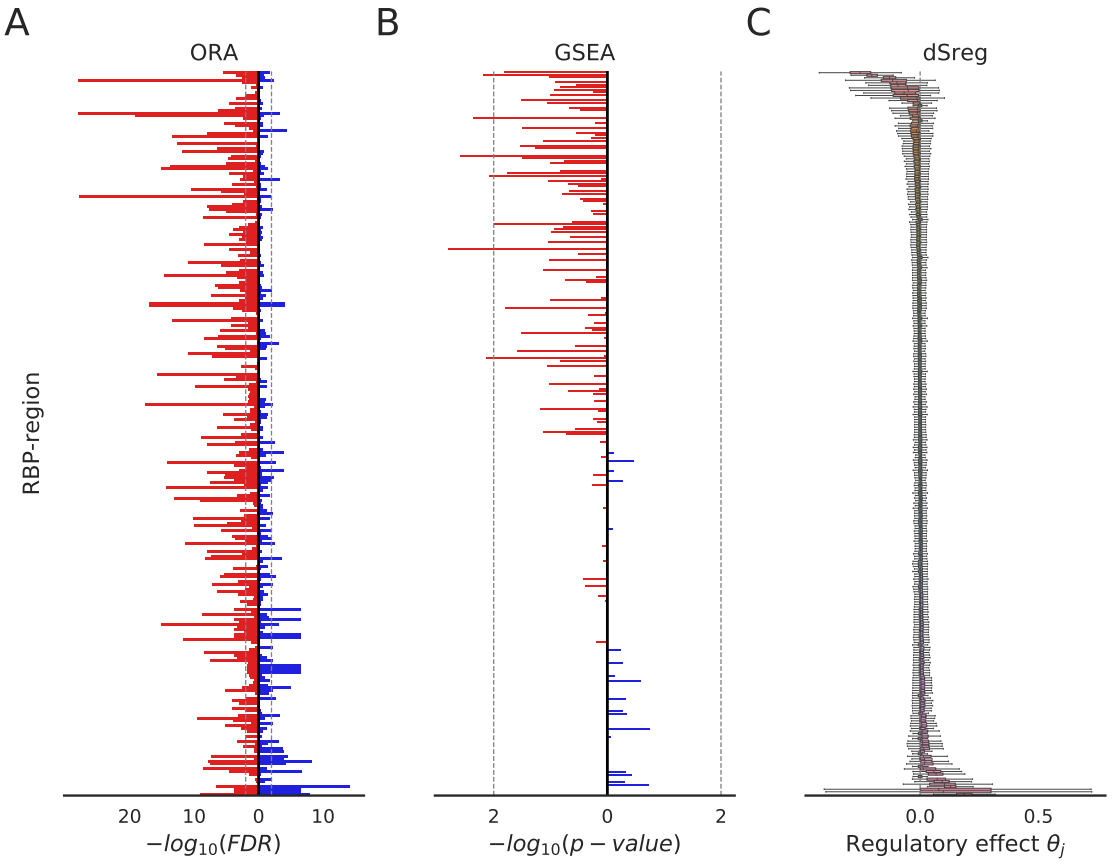

Fig. 5. Comparison of ORA, GSEA and dSreg using a real RNA-seq dataset from a cardiomyocyte differentiation experiment. RBPs on the y-axis are sorted for the three panels according to the posterior mean of the regulatory effect θ_j inferred by dSreg. **A**. Candidate regulatory proteins derived from 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.

regulatory mechanisms using RNA-seq 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 Instituto de salud Carlos III (ISCIII) [CPII14/00027, RD012/0042/0066]; 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 Spanish Ministry of Economy, Industry and Competitiveness and the Pro-CNIC Foundation and is a Severo Ochoa Center of Excellence (MEIC award SEV-2015-0505).

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.

References

1. G. P. Alamancos, A. Pagès, J. L. Trincado, N. Bellora, and E. Eyras. SUPPA: a super-fast pipeline for alternative splicing analysis from RNA-Seq. *bioRxiv*, page 008763, 2014.
2. B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838, aug 2015.
3. S. Anders, S. Anders, A. Reyes, and W. Huber. Detecting differential usage of exons from RNA-Seq data, apr 2012.
4. K. Auinash and T. a. Cooper. Functional consequences of developmentally regulated alternative splicing. *Nature Reviews Genetics*, 12(10):715–729, 2012.
5. F. E. Baralle and J. Giudice. Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology*, 18(7):437–451, 2017.
6. Y. Barash, J. a. Calarco, W. Gao, Q. Pan, X. Wang, O. Shai, B. J. Blencowe, and B. J. Frey. Deciphering the splicing code. *Nature*, 465(7294):53–9, may 2010.
7. M. J. Betancourt and M. Girolami. Hamiltonian Monte Carlo for Hierarchical Models. 2013.
8. K. Blin, C. Dieterich, R. Wurmus, N. Rajewsky, M. Landthaler, and A. Akalin. DoRiNA 2.0–upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Research*, 43(D1):D160–D167, 2015.
9. N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, 2016.
10. B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan : A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1), 2017.
11. C. M. Carvalho, N. G. Polson, and J. G. Scott. Handling sparsity via the horseshoe. *Journal of Machine Learning Research*, 5:73–80, 2009.
12. H. Ctor Climente-Gonzá Lez, E. Porta-Pardo, A. Godzik, E. E. Correspondence, and E. Eyras. The Functional Impact of Alternative Splicing in Cancer. *CellReports*, 20:2215–2226, 2017.
13. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, jan 2013.
14. D. Dominguez, P. Freese, M. S. Alexis, A. Su, M. Hochman, T. Palden, C. Bazile, N. J. Lambert, E. L. Van Nostrand, G. A. Pratt, G. W. Yeo, B. Graveley, and C. B. Burge. Sequence,

Structure and Context Preferences of Human RNA Binding Proteins. *Molecul*, 70:854–7, jun 2018.

15. B. L. Fogel, E. Wexler, A. Wahnich, T. Friedrich, C. Vijayendran, F. Gao, N. Parikshak, G. Konopka, and D. H. Geschwind. RBFOX1 regulates both splicing and transcriptional networks in human neuronal development. *Human Molecular Genetics*, 21(19):4171–4186, 2012.
16. A. Gelman, J. B. B. Carlin, H. S. S. Stern, and D. B. B. Rubin. *Bayesian Data Analysis, Third Edition (Texts in Statistical Science)*. 2014.
17. G. Giudice, F. Sánchez-Cabo, C. Torroja, and E. Lara-Pezzi. ATtRACT-a database of RNA-binding proteins and associated motifs. *Database*, 2016(November):1–9, 2016.
18. J. Giudice, Z. Xia, E. T. Wang, M. a. Scavuzzo, A. J. Ward, A. Kalsotra, W. Wang, X. H. T. Wehrens, C. B. Burge, W. Li, and T. a. Cooper. Alternative splicing regulates vesicular trafficking genes in cardiomyocytes during postnatal heart development. *Nature communications*, 5:3603, 2014.
19. B. R. Graveley. Alternative splicing: Increasing diversity in the proteomic world. *Trends in Genetics*, 17(2):100–107, 2001.
20. M. D. Hoffman and A. Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. 15:1351–1381, 2011.
21. Y. Huang and G. Sanguinetti. BRIE: transcriptome-wide splicing quantification in single cells. *Genome Biology*, 18(1):123, dec 2017.
22. C. Iannone, A. Pohl, P. Papasaikas, D. Soronellas, G. P. Vicent, M. Beato, and J. Valcárcel. Relationship between nucleosome positioning and progesterone-induced alternative splicing in breast cancer cells. pages 360–374, 2015.
23. M. Irimia and S. W. Roy. Origin of spliceosomal introns and alternative splicing. *Cold Spring Harbor perspectives in biology*, 6(6), jun 2014.
24. Y. Katz, E. T. Wang, E. M. Airoldi, and C. B. Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, 7(12):1009–1015, 2010.
25. E. Kim, A. Magen, and G. Ast. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Research*, 35(1):125–131, jan 2007.
26. E. Lara-Pezzi, M. Desco, A. Gatto, and M. V. Gómez-Gaviro. Neurogenesis: Regulation by Alternative Splicing and Related Posttranscriptional Processes. *The Neuroscientist*, 23(5):466–477, oct 2017.
27. E. Lara-Pezzi, J. Gómez-Salinero, A. Gatto, and P. García-Pavía. The alternative heart: Impact of alternative splicing in heart disease. *Journal of Cardiovascular Translational Research*, 6(6):945–955, 2013.
28. J. A. Lee, A. Damianov, C. H. Lin, M. Fontes, N. N. Parikshak, E. S. Anderson, D. H. Geschwind, D. L. Black, and K. C. Martin. Cytoplasmic Rbfox1 Regulates the Expression of Synaptic and Autism-Related Genes. *Neuron*, 89(1):113–128, jan 2016.
29. M. K. K. Leung, H. Y. Xiong, L. J. Lee, and B. J. Frey. Deep learning of the tissue-regulated splicing code. *Bioinformatics (Oxford, England)*, 30(12):i121–9, jun 2014.
30. J.-H. Li, S. Liu, H. Zhou, L.-H. Qu, and J.-H. Yang. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Research*, 42(D1):D92–D97, 2014.
31. M. A. Lines, L. Huang, J. Schwartzentruber, S. L. Douglas, D. C. Lynch, C. Beaulieu, M. L. Guion-Almeida, R. M. Zechi-Ceide, B. Gener, G. Gillessen-Kaesbach, C. Nava, G. Baujat, D. Horn, U. Kini, A. Caliebe, Y. Alanay, G. E. Utine, D. Lev, J. Kohlhase, A. W. Grix, D. R. Lohmann, U. Hehr, D. Böhm, F. C. FORGE Canada Consortium, J. Majewski, D. E. Bulman, D. Wieczorek, and K. M. Boycott. Haploinsufficiency of a spliceosomal GTPase encoded by EFTUD2 causes mandibulofacial dysostosis with microcephaly. *American journal of human genetics*, 90(2):369–77, feb 2012.
32. R. F. Luco, Q. Pan, K. Tominaga, B. J. Blencowe, O. M. Pereira-Smith, and T. Misteli. Regulation of alternative splicing by histone modifications. *Science (New York, N.Y.)*, 327(5968):996–1000, feb 2010.
33. D. Maticzka, S. J. Lange, F. Costa, and R. Backofen. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome biology*, 15(1):R17, 2014.
34. J. Merkin, P. Chen, M. Alexis, S. Hautaniemi, and C. Burge. Origins and Impacts of New Mammalian Exons. *Cell Reports*, 10(12):1992–2005, 2015.
35. J. Merkin, C. Russell, P. Chen, and C. B. Burge. Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. *Science*, 338(December):1593–1600, 2012.
36. T. W. Nilsen and B. R. Graveley. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280):457–463, jan 2010.
37. E. L. V. Nostrand, P. Freese, G. A. Pratt, X. Wang, X. Wei, R. Xiao, S. M. Blue, D. Dominguez, N. A. L. Cody, S. Olson, L. Zhan, C. Bazile, L. Philip, B. Bouvrette, M. O. Duff, K. E. Garcia, C. Gelboin-burkhart, A. Hochman, N. J. Lambert, H. Li, T. B. Nguyen, T. Palden, I. Rabano, R. Stanton, J. Bergalet, B. Zhou, A. Su, R. Wang, A. Brian, A. L. Louie, S. Aigner, X.-d. Fu, E. Lecuyer, B. Christopher, E. Lecuyer, and G. Yeo. A Large-Scale Binding and Functional Map of Human RNA Binding Proteins Correspondence and requests for materials should be addressed to Brenton Graveley (graveley@uchc.edu), Chris Burge (cburge@mit.edu), Xiang-dong Fu (xdfu@ucsd.edu),. *bioRxiv*, pages 1–74, 2017.
38. M. I. Nuno L. Barbosa-Morais, Q. Pan, H. Y. Xiong, S. Gueroussov, L. J. Lee, V. Slobodeniuc, C. Kutter, S. Watt, R. Çolak, T. Kim, C. M. Misquitta-Ali, M. D. Wilson, P. M. Kim, D. T. Odom, B. J. Frey, and B. J. Blencowe1. Research articles. *Science*, 338(December):1587–1594, 2012.
39. R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419, apr 2017.
40. M. Pertea, D. Kim, G. M. Pertea, J. T. Leek, and S. L. Salzberg. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*, 11(9):1650–1667, aug 2016.
41. J. Piironen and A. Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051, 2017.
42. M. Quesnel-vallières, M. Irimia, S. P. Cordes, and B. J. Blencowe. Essential roles for the splicing regulator nSR100 / SRRM4 during nervous system development. *Genes and Development*, pages 746–759, 2015.
43. B. Raj, M. Irimia, U. Braunschweig, T. Sterne-Weiler, D. O'Hanlon, Z. Y. Lin, G. I. Chen, L. E. Easton, J. Ule, A. C. Gingras, E. Eyras, and B. J. Blencowe. A global regulatory mechanism for activating an exon network required for neurogenesis. *Molecular Cell*, 56(1):90–103, 2014.
44. D. Ray, H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, H. Na, M. Irimia, L. H. Matzat, R. K. Dale, S. a. Smith, C. a. Yarosh, S. M. Kelly, B. Nabet, D. Mecenas, W. Li, R. S. Laishram, M. Qiao, H. D. Lipshitz, F. Piano, A. H. Corbett, R. P. Carstens, B. J. Frey, R. a. Anderson, K. W. Lynch, L. O. F. Penalva, E. P. Lei, A. G. Fraser, B. J. Blencowe, Q. D. Morris, and T. R. Hughes. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–7, 2013.
45. E. Sebestyén, B. Singh, B. Miñana, A. Pagès, F. Mateo, M. A. Pujana, J. Valcárcel, and E. Eyras. Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome research*, 26(6):732–44, jun 2016.
46. S. Shen, J. W. Park, J. Huang, K. A. Dittmar, Z.-x. Lu, Q. Zhou, R. P. Carstens, and Y. Xing. MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic acids research*, 40(8):e61, apr 2012.
47. S. Shen, J. W. Park, Z.-x. Lu, L. Lin, M. D. Henry, Y. N. Wu, Q. Zhou, and Y. Xing. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences*, 111:E5593–E5601, 2014.
48. C. Simillion, R. Liechti, H. E. Lischer, V. Ioannidis, and R. Bruggmann. Avoiding the pitfalls of gene set enrichment analysis with SetRank. *BMC Bioinformatics*, 18(1):151, dec 2017.
49. T. P. Stricker, C. D. Brown, C. Bandlamudi, M. McNerney, R. Kittler, V. Montoya, A. Peterson, R. Grossman, and K. P. White. Robust stratification of breast cancer subtypes using differential patterns of transcript isoform expression. *PLOS Genetics*, 13(3):e1006589, mar 2017.
50. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. a. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, 102(43):15545–15550, 2005.
51. J. M. Taliaferro, N. J. Lambert, P. H. Sudmant, M. S. Alexis, C. A. Bazile, C. B. Burge, J. M. Taliaferro, N. J. Lambert, P. H. Sudmant, D. Dominguez, J. J. Merkin, M. S. Alexis, C. A. Bazile, and C. B. Burge. RNA Sequence Context Effects Measured In Vitro Predict In Vivo Protein Binding and Regulation. *Molecular Cell*, pages 1–13, 2016.
52. C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology*, 31(1):46–53, 2013.
53. J. L. Trincado, J. C. Entizne, G. Hysenaj, B. Singh, M. Skalic, D. J. Elliott, and E. Eyras. SUPPA2: Fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biology*, 19(1):40, dec 2018.
54. L. Vigevani, J. Valca, P. Papasaikas, and J. Ramo. Extensive Regulatory Potential of the Core Spliceosomal Machinery. pages 1–16, 2015.
55. J. L. Wagnon, M. Briese, W. Sun, C. L. Mahaffey, T. Curk, G. Rot, J. Ule, and W. N. Frankel. CELF4 Regulates Translation and Local Abundance of a Vast Set of mRNAs, Including Genes Associated with Regulation of Synaptic Function. *PLoS Genetics*, 8(11), 2012.
56. S. M. Weyn-Vanhentenryck, H. Feng, D. Ustianenko, R. Duffié, Q. Yan, M. Jacko, J. C. Martinez, M. Goodwin, X. Zhang, U. Hengst, S. Lomvardas, M. S. Swanson, and C. Zhang. Precise temporal regulation of alternative splicing during neural development. *Nature Communications*, 9(1):2189, dec 2018.
57. Y.-C. T. Yang, C. Di, B. Hu, M. Zhou, Y. Liu, N. Song, Y. Li, J. Umetsu, and Z. Lu. CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics*, 16(1):51, 2015.
58. Z. Zhang, Z. Pan, Y. Ying, Z. Xie, S. Adhikari, J. Phillips, R. P. Carstens, D. L. Black, Y. Wu, and Y. Xing. Deep-learning augmented RNA-seq analysis of transcript splicing. *Nature Methods*, 16(4):307–310, apr 2019.
59. S. Zhao and B. Zhang. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC Genomics*, 16(1):97, 2015.