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 the great transcript diversity found across mammals. AS patterns are dynamically regulated during development and in response to environmental changes. Defects or perturbations in their regulation system may lead to cardiac or neurological disorders. These regulatory mechanisms 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 between steps. Here, we examined the influence of different sequencing depths in the identification of regulatory patterns of AS using simulated data with traditional workflows, showing poor performance and high dependence on sequencing depth. We developed a bayesian model that integrates RNA-seq and regulatory elements data to simultaneously infer changes in inclusion rates and in the activity of the underlying regulators. This model pools weak evidence across AS events to increase the power to infer changes in the regulatory activity using binding sites information increasing both sensitivity and specificity on simulated data. Application to a real dataset provided new insights into the underlying regulatory mechanisms of AS changes, proving the usefulness of our approach for future studies and reanalyses. dSreg was implemented in python using stan and is freely available to the community at https://bitbucket.org/cmartiga/pydsreg/src/master/.

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 (23). This structure provides an opportunity for alternative splicing (AS) to produce different transcripts, which may encode different proteins, from the same gene (35). There is evidence of alternative mRNA processing for most mammalian genes (34, 36) and of widespread changes of AS patterns throughout brain and heart development (4, 5, 15, 18, 21, 39, 40, 52). Defects in mRNA processing of some specific genes often lead to disease (5, 24, 25) and have been associated with complex neurological disorders, such as autistic syndrome (21, 26, 39, 51), and cancer (12, 45). Therefore, understanding the regulatory mechanisms underlying physiologic and pathological changes in splicing patterns is crucial, not only to understand RNA biology better, but also to identify key therapeutic targets with a more general effect in complex diseases.

A typical two step work-flow is generally applied when studying regulatory mechanisms of AS changes between two biological conditions, e.g. disease vs control, (see Figure 1 for schematic representation). First, changes in mRNA processing 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, 38). Then, alternative mRNA processing can be studied at two different levels: 1) transcript quantification, which can be based on a prior alignment as in Cufflinks or Stringtie (38, 48), or directly estimated from fast pseudoalignment methods such as Kallisto or Salmon (9, 37); and 2) event level quantification, as performed by popular tools such as MISO, MATS, vast-tools or DEXseq (3, 21, 22, 43). Alternative splicing events can also be identified and quantified using transcript quantifications, which improves sensitivity with low sequencing depths (1, 49). Regulation is expected to take place locally, making the AS event level the preferred approach to study the regulatory mechanisms underlying changes in splicing profiles. Once AS events have been identified and quantified, different statistical tests or models are applied to find differential splicing between conditions, being a Generalized Linear Model (GLM) with binomial likelihood the most natural parametric approach (43).

The second step aims to statistically associate AS changes with features related to regulatory elements, mostly RNA binding proteins (RBPs). Such features often include nucleotide hexamers, predicted motifs, experimentally determined or predicted binding sites (14, 17, 41, 53). Over-Representation Analysis (ORA) enables finding over-represented features in the set of events showing significant changes compared with events without significant changes, assuming that they remain mostly unchanged. Therefore, a sufficiently large set of significantly changed events is required to reach sufficient power to detect enrichment of RBP-related features. Since fewer nucleotide positions can be used for estimation of inclusion rates than for gene expression, reaching enough power to detect significant differences requires higher sequencing depth, yielding many RNA-seq datasets under-powered for AS analyses and AS studies more costly overall. ORA requires the discretization of splicing changes into different categories e.g. included or skipped. Categorization according to changes not only depends strongly on typically low statistical power, but also ignores quantitative information about AS changes. There are popular methods that enable the inclusion of quantitative information in the enrichment procedure, i.e. the Gene Set Enrichment Analysis (GSEA) tool and some para-

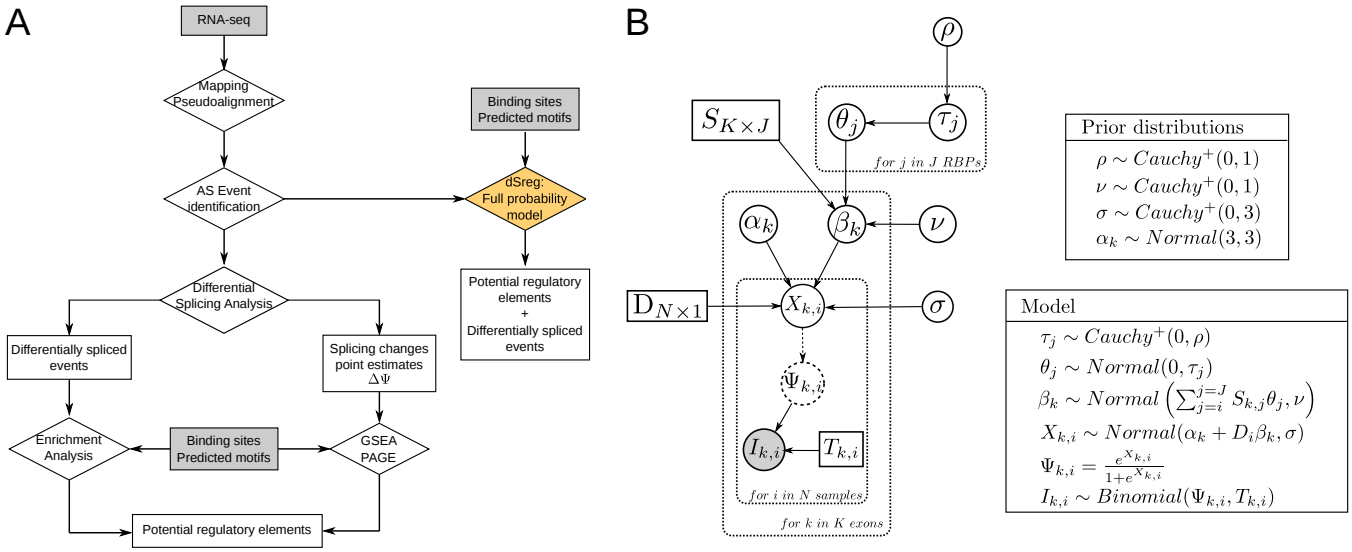

Fig. 1. General and proposed work-flows for AS regulation analysis. **A**. Diagram representing the different steps required for a classical analysis of regulation of alternative splicing using RNA-seq data and the proposed model in dSreg. **B**. Directed Acyclic Graph (DAG) representing the full probabilistic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

metric versions such as Parametric Analysis of Gene Set Enrichment (PAGE) (44, 46). Although these tools were designed for functional analysis, they can also be used to perform enrichment of known targets of regulatory elements. Such GSEA-like methods have been applied to analyze AS regulation using sequence motifs (42, 49). Even if there is no systematic evaluation of the performance these approaches, 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. Therefore, methods that take into account the uncertainty of the estimations would be expected to provide better results. Moreover, an additional limitation affecting both ORA and GSEA approaches lies on the high number of different features or binding sites and potential co-linearities among them, resulting in a high false positive rate from confounding effects.

In this work, we first studied the performance and limitations of the classical enrichment approaches (ORA and GSEA) for the detection of regulatory elements driving AS changes using simulated data. 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 regulatory effects of the RBPs binding to each event. We used a hierarchical shrinkage prior for the changes in the activity of RBPs with predefined binding sites to promote sparsity i.e. AS are due only to changes the activity of a limited number of regulators, and therefore to limit false positive rate due to co-linearities in binding profiles of different RBPs. This model showed improved identification of both regulatory mechanisms and AS changes in simulated data under these assumptions. We further tested the usefulness of this model in a real RNA-seq dataset obtained during cardiomyocytes differentiation, which enabled us to considerably reduce the number of potential AS regulators and unveiled a regulatory role for some of the core components of the spliceosome in the differentiation process.

Methods

dSreg: a mechanistic probability model for differential splicing.

dSreg models the AS changes between two different conditions as a function of changes in the activity of a few of the existing RBPs acting through their known binding sites. As a result of primary processing of RNA-seq data, our data will consist on a total of K AS events detected across N samples. For each event k and sample i, we observe $I_{k,i}$ reads supporting exon inclusion out of a total of $T_{k,i}$ reads mapping to the exon skipping event, which depends on the unknown probability of inclusion $\Psi_{k,i}$. The binomial distribution enables the calculation of the conditional probability of observing $I_{k,i}$ reads given $T_{k,i}$ and $\Psi_{k,i}$.

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

$\Psi_{k,i}$ 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 on the group to which the sample belongs, we take the logit transformation $X_{k,i}$,

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

and assume that it is drawn from a normal distribution with different means per condition: α_k and $\alpha_k + \beta_k$, such that β_k represents the difference between the two conditions; and with a certain standard deviation σ_k. For simplicity, we assumed here that the standard deviation $\sigma_k = \sigma$ is the same across all K AS events:

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

where D_i is a variable that takes the value 1 when the sample belongs to condition 2, and 0 when belongs to condition 1:

$$D_i = \begin{cases} 1 & \text{if sample } i \text{ in group 2} \\ 0 & \text{if sample } i \text{ in group 1} \end{cases}$$

Up to this point, this model is a simple logistic regression for each event with the only assumption that the sample variance is common across events. However, we can imagine that changes in the probability of inclusion of exon k between two conditions, indirectly modeled by β_k, depend on the change in the activity θ_j of a particular regulatory RBP j and on whether it can bind to a specific region of exon k e.g. the upstream or downstream intron. This information is encoded in a matrix $\mathbf{S}_{K\times J}$, with value 1 whenever the protein binds j to the exon k and 0 otherwise. The matrix $\mathbf{S}$ could also contain continuous values such as the probability of binding, its affinity, scores given by Position Weighted Matrices (PWMs) (41) or any other predictive tool (2, 32).

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

Under this model, we assume that regulatory elements have additive and independent effects on the inclusion rate of a given exon. Incorporation of synergistic or competitive effects to the model would require to add interaction terms that would greatly increase the number of parameters to be estimated so they were left out in dSreg. Under these assumptions, we model β_k, the change in the logit-transformed inclusion rate of exon k, as a normal distribution with a linear combination of regulatory effects $\vec{\theta}$ and S_k (the binding profile of exon k) as mean, and a certain standard deviation ν. Adding variance to the distribution of β_k allows to have some changes in AS not necessarily due to the regulatory features included in the model.

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

A large number of regulatory proteins are usually tested in this type of analyses. However, only the binding of a few RBPs may have an effect on the inclusion rates of target exons. We formalize this prior belief setting a horseshoe prior for 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.

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

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

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

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 and α_k is the logit transformation of the inclusion rate in condition 1, 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 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 data and Θ, since the 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.

$$p(\Theta, \mathbf{I}) = p(\mathbf{I},\ \mathbf{T},\ X, \alpha, \beta, \nu, \theta, \tau, \rho, D, S) = \quad \textbf{(11)}$$

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

where,

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

Once the full posterior distribution is completely specified, it can be approximated using Markov Chain Monte Carlo (MCMC) algorithms. In particular, it was implemented using stan (10) in dSreg, a small python library to fit this model to analyze AS changes between two conditions and their regulation in any RNA-seq dataset. The model is represented a as Directed Acyclic Graph (DAG) to show dependencies among parameters in Fig. 1B.

Data simulation.

Data can be simulated by setting fixed values of the parent nodes of the DAG representing the probabilistic model (Fig. 1B) and drawing samples from the corresponding distributions for each parameter. We therefore needed to have fixed values for the parent nodes σ, α_k, θ_k, $T_{k,i}$, and S. We simulated 20 datasets per 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 1 ($\Psi_{k,a}$) 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 distribtuion,

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

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 discrete 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.

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

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

Next, we needed to simulate changes in the activity of a few RBPs. For that, we randomly draw 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 2, 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 2 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-mean 2 ($log(\lambda) = 2$) 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.

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 (46). 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.

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 (19), 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).

Real data analysis.

GSE59383 fastq data were downloaded and mapped using vast-tools 0.2.0 (21) 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 to find regulatory patterns using a compendium of CLiP-seq binding sites from several databases in BED format (8, 14, 28, 53). 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 analysis.

Results

Traditional differential splicing analysis shows poor performance and threshold dependence.

First, we assessed the performance of a classical GLM approach for its ability to detect significant changes in inclusion rates and how it was influenced by sequencing depth λ using simulated data (20 datasets per condition; see Methods section for details). λ represents the mean of a Poisson distribution used to simulate the total number of counts arisen from a particular event. We found that, at low sequencing depths λ, the sensitivity at 5% FDR is very low ($< 10\%$ with $log(\lambda) \geq 1$) when using a simple GLM. This is expected since there is very little information to estimate the inclusion rate in each exon and sample.

As λ increases, so does the sensitivity of GLM (Fig. 2A). However, the specificity also tends to decrease: as the inclusion rate is better estimated in each sample, differences arisen by chance from the selection of only 3 samples per condition accumulate more evidence and become significant. As expected, the specificity is lower when relaxing the threshold on the FDR (Fig. 2B). Interestingly, the F1 score, which integrates both sensitivity and specificity, is higher with higher FDR thresholds (Fig. 2C). To avoid the need to select an arbitrary threshold to assess the performance of the different methods, we calculated the Receiver Operating Characteristic (ROC) curves for each simulated dataset and the area under them (AUROC, Fig. 2D and E). These results show that, at low sequencing depths ($log(\lambda) < 3$), the performance is rather poor, with AUROC values of 0.7 at most.

Incorporation of information about regulatory elements increases the power to detect AS changes.

We then run dSreg to evaluate if there were improvements in the identification of splicing changes. Additionally, we also fitted 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). This was to check whether potential improvements were due to the inclusion of binding sites and changes in RBPs activity in the model or just to variance pooling. We selected significantly changed events as those with a posterior probability higher than 95% of having a difference β_k larger than 0. dSreg but not the Null model showed increased sensitivity, even at very low sequencing depths, when there was practically no information from individual events (Fig. 2A). This increased sensitivity did not come with a decrease in specificity as could be expected, but it was 1 for all the simulations performed (Fig. 2C), with consequently very high F1 scores and area under the ROC curves, 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 variance pooling across events only marginally improves inference of splicing changes, at least with the low variance used in these simulations. Therefore, dSreg effectively used information about the underlying regulatory mechanisms to correct differences that easily arise by chance in datasets with sample size, as simulations were done with only 3 samples per condition.

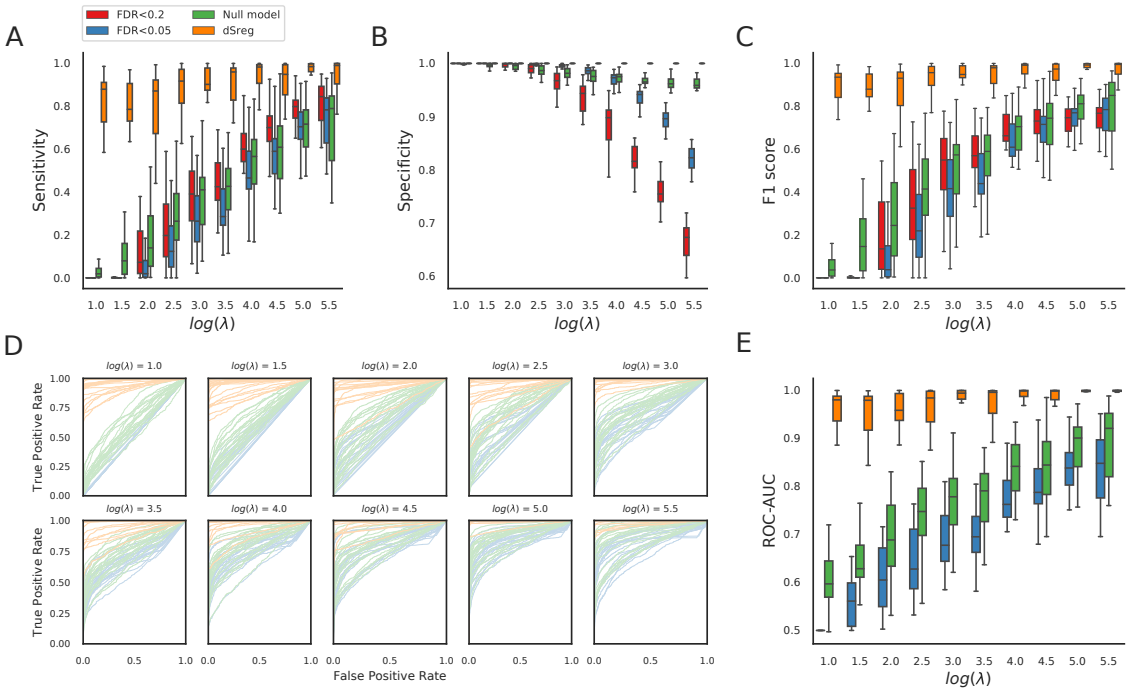

Fig. 2. Comparison of the performance for the identification of different event inclusion rates of a standard method using a single GLM per exon considering two FDR thresholds (0.05 and 0.2),a bayesian model that pools variance across all exons (Null model) and dSreg. Performance was analyzed in simulations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

dSreg improves the detection of changes in the activity of RBPs.

We then focused on the identification of regulatory elements that drive splicing changes in our simulated datasets comparing 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 downstream enrichment analyses. The dependency of ORA on the detection of significant changes led to low F1 scores at any tested FDR threshold, specially 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 (46) (see Methods section for details). We found a substantial improvement over ORA, with higher F1 scores, specially at low sequencing depths. Interestingly, GSEA did not seem to benefit from higher sequencing depths, which would help improve the quality of estimated changes in inclusion rates (Fig. 3A). However, dSreg outperformed both ORA and GSEA, as it takes into account both quantitative information (opposed to mere ranking) and the uncertainty of the estimations, showing much higher F1 scores for any of the sequencing depths tested (Fig. 3A). Furthermore, it uses information about the regulatory mechanisms to infer single event changes (Fig. 2). 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.

dSreg is robust when testing high numbers of regulatory elements.

We had so far explored the effect of sequencing depth on results using simulations. We then wanted to assess how a higher number of potential regulatory elements J influenced the results, since the number of false positives is expected to increase. In addition, co-linearities among binding profiles of different RBPs might hinder the identification of the real regulatory elements. With this aim, 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 outperforms both methods and remained unaffected by the inclusion of other inactive regulatory elements, at least up to the 250 regulators that were tested here (Fig. 3B).

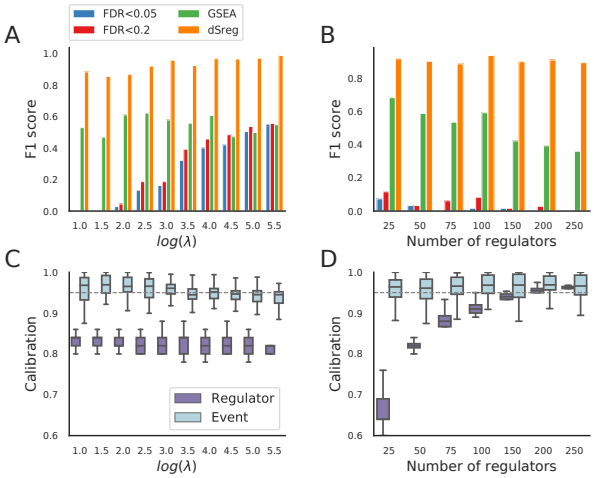

Fig. 3. Performance of methods for the detection of regulatory elements: ORA with variable FDR thresholds (0.05 and 0.2), non-parametric GSEA and dSreg. Performance was analyzed in simulations with increasing sequencing depths λ, which is the mean of the Poisson distribution used to simulate the total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

Model calibration depends on the proportion 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 underestimated, given that 95%CI included the real values less 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 active 5 out of an increasing number of candidates RBPs. Calibration did depend on the number of total regulatory elements to be tested, such that best calibration was reached with 2-2.5% active RBPs (5 out of 150-200 regulators, Fig. 3C). The prior distribution on the global shrinkage parameter ρ may need to be adjusted according to our expectation of the proportion of independently active regulatory elements contributing to splicing changes in a particular experiment.

AS regulation in cardiomyocyte differentiation by core-spliceosomal factors.

We then tested our model on a real 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 Cross Linking and immunoprecipitation followed by sequencing (CLiP-seq) experiments and only those located in the upstream and downstream intronic flanking 250bp were used, reaching a total 286 binding profiles to test (see Methods section for details). We run the 3 approaches explored in this work and found that ORA results in a high number of significantly enriched candidates, most of which are likely to represent false positives as in our simulation analysis (Fig. 4A). GSEA, on the other hand, showed no significant enrichment at $FDR < 0.05$, and only a few at nominal $p - values < 0.05$, which suggest that these p-values can easily arise by chance. Indeed, there is little concordance with results from the over-representation analysis (Fig. 4A and B). dSreg showed an overall agreement with ORA results, as top hits showed differential activity in dSreg. However, dSreg provided a reduced number of RBPs whose binding site profiles helped explain the observed AS changes, suggestive of higher specificity (Fig. 4, Table 1). Our results highlighted the role of PTBP1 in cardiac myocytes differentiationm which has been recently suggested (30) and unveiled an additional role of its paralog PTBP2. Interestingly, a great deal of the identified regulatory RBPs are considered members of the core spliceosome (BUD13, EFTUD2, PRPF8, SF3A3, SF3BA4), 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 (50) and mutations in one of these genes (EFTUD2) have been associated with congenital heart defects, among other phenotypes (29).

Discussion

Here we propose dSreg, a new method that integrates the analysis of differential splicing 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, specially 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 the small sample size. 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 constant. 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 up the possibility to analyze AS more accurately using RNA-seq data with low sequencing depth, both for re-analysis of previously sequenced samples or for more cost-effective new RNA-seq experiments. Whereas transcript-based methods also lower the requirements on sequencing depths (1, 49), our model works directly at the event level,

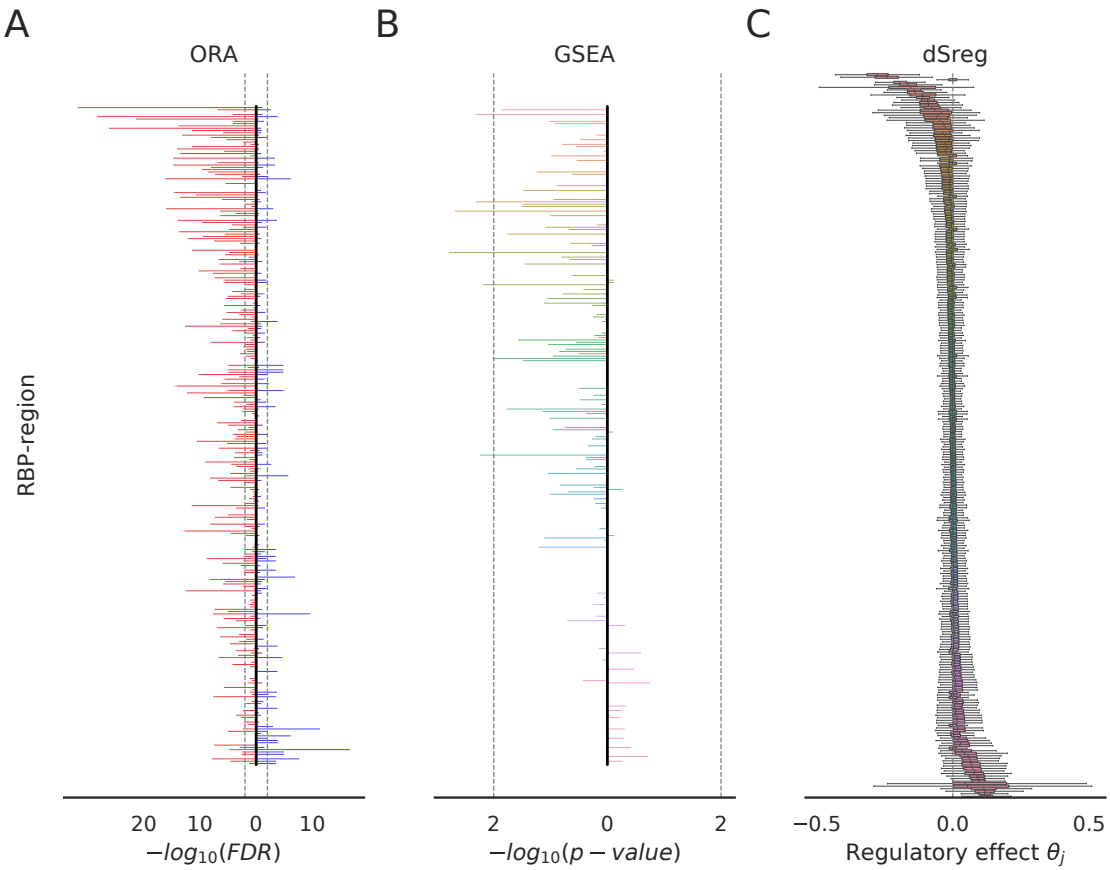

Fig. 4. Comparison of ORA, GSEA and dSreg using a real RNA-seq dataset from a cardiomyocyte differentiation experiment. RBPs on the y-axis are sorted for the three panels according to the posterior mean of the regulatory effect θ_j inferred by dSreg. **A**. Candidate regulatory proteins derived from 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.

reducing the dependency on the transcript annotation (54). In contrast to previous approaches, including bayesian methods like MISO (22), our model is motivated by how splicing changes arise between two biological conditions rather than on how inclusion and skipping reads are generated from the inclusion rate (Ψ) in a particular sample. In any case, integration of these two types of models is not only possible, but also a clear way for future improvements.

Our good results on simulations are, however, restricted to those cases in which splicing changes are mediated only by a subset of differentially active RBPs binding to completely known sites. Although we have included a high number of RBPs without any effect on the changes in inclusion rates between the two 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, we found that in the cardiomyocyte differentiation experiment, GSEA lacks sufficient power to detect regulators unlike the other two approaches. In contrast, in the simulations, GSEA showed better performance than ORA on included and skipped exons, which indicates that our model is somehow incomplete and that there are other factors contributing to changes in AS. This is expected. as previous results 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, 27, 47). Since our aim is to identify regulatory mechanisms in a particular scenario rather than predicting splicing patterns, our model is more appropriate than black box neural networks. Yet, we expect that careful modeling of additional AS regulatory mechanisms will improve the results, e.g. nucleosome positioning and histone modifications, which would require also more layers of information (20, 31, 33). Moreover, this model 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. Integrative modeling of clustering methods with enrichment would be an interesting way forward for more complex experimental designs.

Conclusions

Our model provides an example of how joint modeling of interdependent phenomena can improve results compared with completely separated analysis relying on discretization according to rather arbitrary thresholds. Bayesian inference

through MCMC methods provide 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 alternative splicing data, but these models are flexible enough to easily include information about regulators expression, post-transcriptional modifications or any other 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 to be taken into account in the analysis i.e. the binding sites. Therefore, dSreg adds a valuable statistical framework to existing software aimed at identifying AS events, such as rMATS, vast-tools, (21, 43), for more accurate identification of AS 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.

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. In *International Conference on Artificial Intelligence and Statistics*, pages 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. M. D. Hoffman and A. Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. 15:1351–1381, 2011.
20. 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.
21. M. Irimia and S. W. Roy. Origin of spliceosomal introns and alternative splicing. *Cold Spring Harbor perspectives in biology*, 6(6), jun 2014.
22. 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.
23. E. Kim, A. Magen, and G. Ast. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Research*, 35(1):125–131, jan 2007.
24. 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.
25. 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.
26. 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.
27. 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.
28. 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.
29. 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.
30. Z. Liu, L. Wang, D. J. Welch, H. Ma, Y. Zhou, R. H. Vaseghi, S. Yu, B. J. Wall, S. Alimohamadi, M. Zheng, C. Yin, W. Shen, F. J. Prins, J. Liu, and L. Qian. Single cell transcriptomics reconstructs fate conversion from fibroblast to cardiomyocyte. *Nature*, 551(7678):100–104, 2017.
31. 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.
32. D. Maticzka, S. J. Lange, F. Costa, and R. Backofen. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome biology*, 15(1):R17, 2014.
33. 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.
34. 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.
35. T. W. Nilsen and B. R. Graveley. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280):457–463, jan 2010.
36. 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.
37. 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.
38. 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.
39. 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.
40. 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.

41. 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.
42. 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.
43. 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.
44. 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.
45. 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.
46. 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.
47. 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.
48. 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.
49. 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.
50. L. Vigevani, J. Valca, P. Papasaikas, and J. Ramo. Extensive Regulatory Potential of the Core Spliceosomal Machinery. pages 1–16, 2015.
51. 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.
52. 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.
53. 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.
54. 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.