Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Gene expression

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 (CNIC), Melchor Fernández Almagro 3, Madrid, Spain

[*] Corresponding authors: Tel: +34 914531200; Fax: +34 914531265; Email: fscabo@cnic.es, elara@cnic.es; equal contribution

Abstract

Motivation: Alternative splicing (AS) is an important mechanism in the generation of transcript diversity across mammals. AS patterns are dynamically regulated during development and in response to environmental changes. Defects or perturbations in its regulation may lead to cancer or neurological disorders, among other pathological conditions. The regulatory mechanisms controlling AS in a given biological context are typically inferred using a two step-framework: differential AS analysis followed by enrichment methods. These strategies require setting rather arbitrary thresholds and are prone to error propagation along the analysis.
Results: To overcome these limitations, we propose dSreg, a Bayesian model that integrates RNA-seq with data from regulatory features, e.g. binding sites of RNA binding proteins (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.
Availability: dSreg was implemented in python using stan and is freely available to the community at https://bitbucket.org/cmartiga/dsreg.
Contact: fscabo@cnic.es, elara@cnic.es,
Supplementary information: Supplementary data are available at *Bioinformatics* online.

Extended Methods

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.

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

$$\Psi_{k,a} \sim \begin{cases} Beta(10,1)\ if\ u_k > 0.2 \\ Uniform(0,1)\ if\ u_k < 0.2 \end{cases} \tag{2}$$

$$\alpha_k = logit(\Psi_{k,a}) = log\left(\frac{\Psi_{k,a}}{1-\Psi_{k,a}}\right) \tag{3}$$

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,

$$\Sigma_{J\times J} \sim InvWishart\left(J+1, \frac{1}{J}\mathbb{I}_J\right) \tag{4}$$

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.

$$\vec{M}_k \sim MvNormal(-2.5, \Sigma) \tag{5}$$

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.

$$P_{k,j} = InvLogit(M_{k,j}) = \frac{e^{M_{k,j}}}{1+e^{M_{k,j}}} \tag{6}$$

$$S_{k,j} \sim Bernoulli(P_{k,j}) \tag{7}$$

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.

$$\theta_j \sim \begin{cases} Uniform(-2.5, 2.5)\ if\ j \in A \\ 0\ otherwise \end{cases} \tag{8}$$

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.

$$\beta_k \sim Normal\left(\sum_{j=1}^{j=J} S_{k,j}\theta_j,\ \nu\right) \tag{9}$$

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,

$$X_{k,i} \sim Normal(\alpha_k + D_{k,i}\beta_k, \sigma) \tag{10}$$

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

$$T_{k,i} \sim Poisson(\lambda) \tag{11}$$

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 $X_{k,i}$.

$$I_{k,i} \sim Binomial\left(T_{k,i}, InvLogit(X_{k,i})\right) \tag{12}$$

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 (Carpenter *et al.*, 2017) using non-centered parametrization, whenever it was possible, to improve sampling efficiency (Betancourt and Girolami, 2013). The joint posterior distributions of the parameters were approximated using No-U Turn Sampler (NUTS) as implemented in Stan (Hoffman and Gelman, 2011), 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 ($\hat{R}$) (Gelman *et al.*, 2014).

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 (Subramanian *et al.*, 2005). 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 (Blin *et al.*, 2015; Li *et al.*, 2014; Yang *et al.*, 2015; Dominguez *et al.*, 2018). 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 (Zhang *et al.*, 2019). 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 (Shen *et al.*, 2014; Zhang *et al.*, 2019). 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 (Katz *et al.*, 2010; Huang and Sanguinetti, 2017). 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 (Nostrand *et al.*, 2017; Dominguez *et al.*, 2018). 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 (Irimia and Roy, 2014) 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.

References

Betancourt, M. J. and Girolami, M. (2013). Hamiltonian Monte Carlo for Hierarchical Models.

Blin, K., Dieterich, C., Wurmus, R., Rajewsky, N., Landthaler, M., and Akalin, A. (2015). DoRiNA 2.0–upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Research*, **43**(D1), D160–D167.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). <i>Stan</i> : A Probabilistic Programming Language. *Journal of Statistical Software*, **76**(1).

Dominguez, D., Freese, P., Alexis, M. S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N. J., Van Nostrand, E. L., Pratt, G. A., Yeo, G. W., Graveley, B., and Burge, C. B. (2018). Sequence, Structure and Context Preferences of Human RNA Binding Proteins. *Molecul*, **70**, 854–7.

Gelman, A., Carlin, J. B. B., Stern, H. S. S., and Rubin, D. B. B. (2014). *Bayesian Data Analysis, Third Edition (Texts in Statistical Science)*.

Hoffman, M. D. and Gelman, A. (2011). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. **15**, 1351–1381.

Huang, Y. and Sanguinetti, G. (2017). BRIE: transcriptome-wide splicing quantification in single cells. *Genome Biology*, **18**(1), 123.

Irimia, M. and Roy, S. W. (2014). Origin of spliceosomal introns and alternative splicing. *Cold Spring Harbor perspectives in biology*, **6**(6).

Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, **7**(12), 1009–1015.

Li, J.-H., Liu, S., Zhou, H., Qu, L.-H., and Yang, J.-H. (2014). 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.

Nostrand, E. L. V., Freese, P., Pratt, G. A., Wang, X., Wei, X., Xiao, R., Blue, S. M., Dominguez, D., Cody, N. A. L., Olson, S., Zhan, L., Bazile, C., Philip, L., Bouvrette, B., Duff, M. O., Garcia, K. E., Gelboin-burkhart, C., Hochman, A., Lambert, N. J., Li, H., Nguyen, T. B., Palden, T., Rabano, I., Stanton, R., Bergalet, J., Zhou, B., Su, A., Wang, R., Brian, A., Louie, A. L., Aigner, S., Fu, X.-d., Lecuyer, E., Christopher, B., Lecuyer, E., and Yeo, G. (2017). 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.

Shen, S., Park, J. W., Lu, Z.-x., Lin, L., Henry, M. D., Wu, Y. N., Zhou, Q., and Xing, Y. (2014). rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences*, **111**, E5593–E5601.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. a., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**(43), 15545–15550.

Yang, Y.-C. T., Di, C., Hu, B., Zhou, M., Liu, Y., Song, N., Li, Y., Umetsu, J., and Lu, Z. (2015). CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics*, **16**(1), 51.

Zhang, Z., Pan, Z., Ying, Y., Xie, Z., Adhikari, S., Phillips, J., Carstens, R. P., Black, D. L., Wu, Y., and Xing, Y. (2019). Deep-learning augmented RNA-seq analysis of transcript splicing. *Nature Methods*, **16**(4), 307–310.

Supplementary Figures

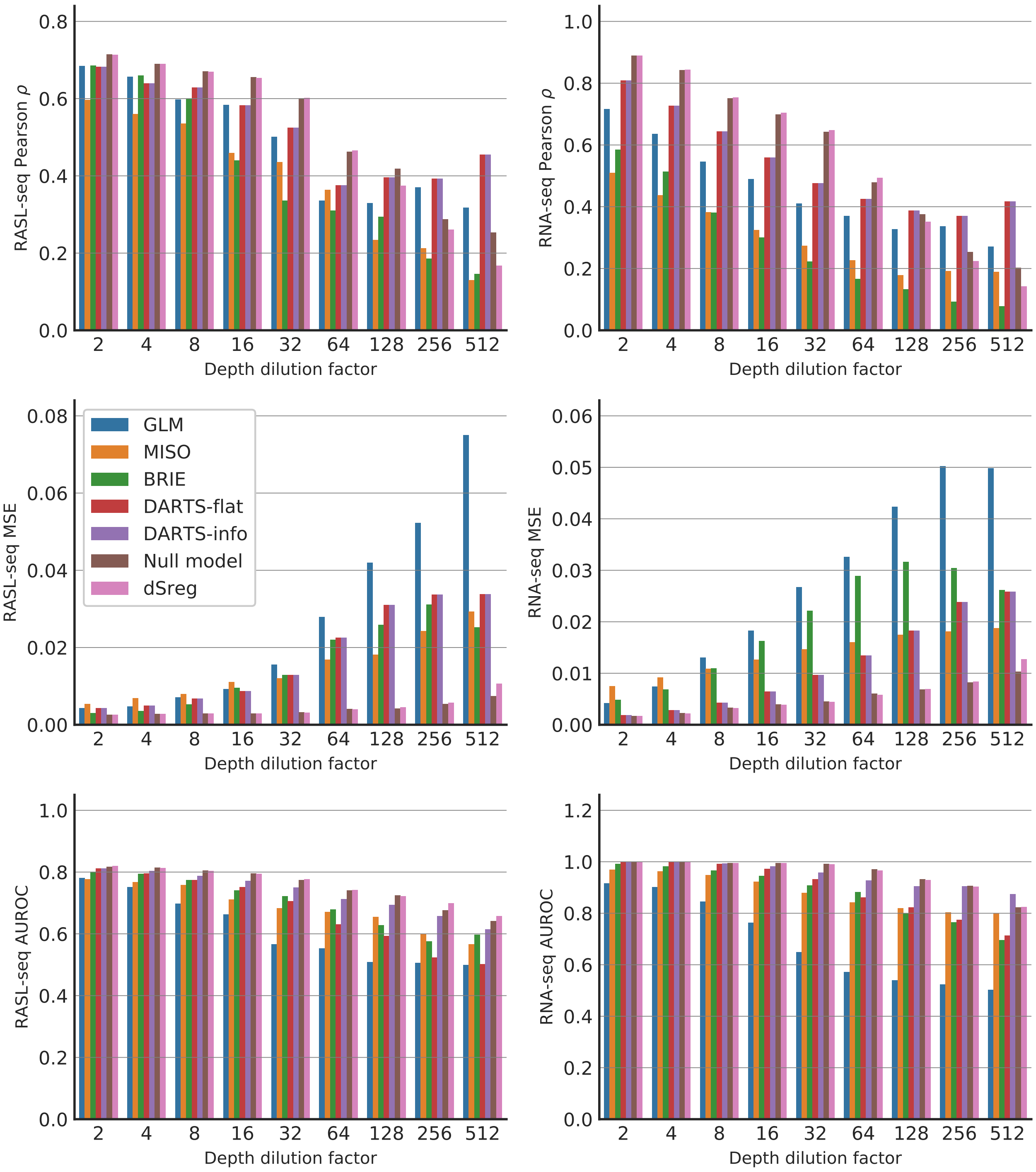

Fig. 1. Evaluation of the identification of AS changes of dSreg with other methods on real data. Performance of differential splicing methods using RASL-seq quantifications (left column) and full coverage RNA-seq (right column) as true values, measured by Pearson correlation of $\Delta\Psi$, MSE, and AUROC. The different measures are represented in different rows