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

OXFORD

Gene expression

csDEX: Condition-specific differential exon expression

Martin Stražar [1,*], Jernej Ule [2] and Tomaž Curk [1,*]

[1]Bioinformatics Laboratory, Faculty of Computer and Information Science, University of Ljubljana ,Večna pot 113, 1000 Ljubljana, Slovenia and
[2]The Francis Crick Institute,1 Midland Road, London NW1 1AT, United Kingdom

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Large-scale RNA sequencing studies with hundreds of experimental conditions allow elucidation of the alternative splicing (AS) mechanisms. A limitation of contemporary differential exon usage (DEU) tests is the comparison of multiple experimental conditions to a single reference, leading to increased probability of identifying the same exon in multiple conditions. The expression data models are based either on mapped read counts or Percent spliced-in (PSI), lacking an universal framework.
Results: We design the Condition-specific differential exon expression (csDEX) models, based on Negative Binomial (read counts) and Beta regression (PSI), identifying AS changes unique to a small subset of conditions. A low-rank approximation of the design matrix is proposed, with no increase in false positive rate and a decrease in run time. With an increasing number of conditions, csDEX improves on retrieval accuracy and hyperparameters estimation over comparable DEU methods. We evaluate csDEX on the ENCODE project shRNA knockdown RNA-seq data on 190 RBPs (e.g., SRSF1, U2AF1/2, PTBP1, hnRNPs, TARDBP) and UCSC *knownAlt* annotation. The csDEX models improve over comparable DEU methods, with precision of 98% (PSI-based) and 82% (count-based). The causal effect of RBP binding on AS is verified by independent data sources on RBP binding (eCLIP), sequence motifs, and successful retrieval of previously verified cryptic exons regulated by TARDBP.
Availability and implementation: csDEX is an open source R package, with code and examples available at `https://www.github.com/mstrazar/csDEX`.
Contact: martin.strazar@fri.uni-lj.si
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The advent of next-generation sequencing and the development of methods such as RNA-seq has lead to a major increase in mapping resolution and quantification precision of transcriptomes (Mortazavi *et al.*, 2008). The ability to quantify expression on a nucleotide resolution has driven the design of differential expression models at the gene expression and alternative splicing (AS) level, the two main aspects of transcriptome diversity.

A common experimental design assumed by AS models includes a *control* condition and a number of *case* conditions. Following the decrease in the cost of sequencing, the number of assayed experimental conditions can reach tens or thousands. For example, the Encyclopedia of DNA Elements project includes 498 human, mouse and fly RNA-seq experiments (ENCODE, Consortium (2004)), or a recent *Arabidopsis Thaliana* data set which contains 285 RNA-seq experiments (Zhang *et al.*, 2016). More similar collaborative initiatives are expected in the future (Goodwin *et al.*, 2016). The increase in the order of magnitude has strong implications on modeling in terms of efficiency and statistical power.

Here, we design a statistical package *csDEX* (condition-specific Differential Exon Expression) to detect changes in alternative splicing that are unique to a single or a few experimental conditions. Detecting condition-specific splicing changes has important practical applications, including but not limited to: identifying outliers/experimental artifacts,

retrieving exons/splice sites regulated only by a particular RNA-binding protein (RBP), using multiple conditions as background in absence of control/reference experiments. We demonstrate the utility of csDEX on a large dataset of 190 human RBP knockdown samples, identifying RBP-specific splicing changes. The identified splicing changes are ranked on statistical significance and verified with independent, external data sources on RBP binding (eCLIP) and sequence motif analysis.

Statistical models of AS have extended the models of gene expression, as both modeling scenarios have multiple data processing steps in common, including read mapping, expression quantification and differential expression analysis. A recent report empirically shows that the choice of differential expression model had the greatest effect on precision/recall of the resulting candidates (Williams *et al.*, 2017). Of the two mechanisms, alternative pre-mRNA splicing (AS) is more sensitive to the quantification measure, as changes in AS are assessed within- rather than between genes. Notable differences in AS modeling software packages therefore include representation of basic splicing unit (exon, splicing junction, exonic/intronic part, etc.), and the quantification metric, which largely determines the distributional and other modeling assumptions. In the following, we briefly review existing approaches and position csDEX within this spectrum.

Definition of basic splicing unit determines whether an AS splicing model assumes a known gene/transcript annotation. The earliest of AS models MISO (Katz *et al.*, 2010), SpliceTrap (Wu *et al.*, 2011), MATS (Shen *et al.*, 2012), rSeqDiff (Shi and Jiang, 2013), rMATS (Shen *et al.*, 2014) assume transcript quantification and compare the ratio of two isoforms, including and excluding an alternative exon of interest. A splice graph representation pioneered by FDM (Singh *et al.*, 2011) and related alternative splicing module (ASM, DiffSplice by Hu *et al.* (2013)), also used by jSplice (Christinat *et al.*, 2016), MAJIQ (Vaquero-Garcia *et al.*, 2016), SDEAP (Yang and Jiang, 2016), employ changes in probability distributions of different paths through the splice graph to discover splicing changes. Finally, the methods DEXSeq (Anders *et al.*, 2012), JunctionSeq (Hartley and Mullikin, 2016) and Alexa-Seq (Griffith *et al.*, 2010) build a non-overlapping exonic part (bin) representation from existing transcript annotation (a GTF file) and perform analysis on the newly generated GFF file. Our model csDEX assumes the former, predefined GFF-based exonic part representation, due to standardization and ability to relate to all types of alternative splicing events in the UCSC *Known Alt* annotation (Kent *et al.*, 2002) used in our case study.

Quantification of splicing unit usage is the main determinant of the modeling choices. Read-count based models (e.g. DEXSeq, JunctionSeq, DiffSplice, jSplices, Alexa-Seq, rSeqDiff) compare the fold-change of reads mapping to a splicing unit between two conditions. This approach is prone to false positives originating from changes in gene expression rather than changes in splicing, as read counts are assigned to splicing units independently from its neighbours. Methods based on isoforms (e.g. FDM, SpliceTrap) estimate the divergence between distribution of isoform usage. Similarly, the percentage-spliced in (PSI) quantity determines the ratio of isoforms including a splicing unit versus all isoforms of a gene (e.g. rMATS, MISO). Although assigned to each splicing unit individually, its computation is inherently dependent on neighbouring splicing units. Another advantage of PSI is the absence of normalization issues, present for count-based models. To our knowledge, csDEX is the first method including both count- and PSI-based expression quantification, based on generalized linear models; the Negative Binominal and the Beta regression models are designed respectively (Smithson and Verkuilen, 2006; Ferrari and Cribari-Neto, 2004; Dobson and Barnett, 2008). Here, we empirically compare count- and PSI-based modeling using the representative methods and outline the advantages of PSI for modeling AS.

Modeling with multiple experimental conditions brings practical and statistical challenge. The majority of the approaches require defining a *control* condition against which all other conditions are compared. Firstly, although the modeling benefits from having more data for parameter inference, it does not enable finding changes that are specific to a small subset of conditions (condition-specificity). Secondly, performing all possible comparisons naturally leads to increase in computational complexity and degrees of freedom - less statistical power. A differential gene expression model Multi-DE tackles both problems with a low-rank approximation, however with loss of condition-specificity (Park and Wu, 2016). The csDEX model specifies exhaustive testing for all combinations of splicing units and experimental conditions. In the case of very sparse data sets (most splicing units having zero expression), we propose a model approximation based on the Wald test for parameter significance, resulting in up to three-fold execution speed-up with no increase in Type I error probability (false positive rate).

csDEX is an R package with standard documentation and guide. The user interface and input data format are extended from existing DESeq2 and DEXSeq packages, enabling interoperability and easier comparison.

2 A family of condition-specific differential expression models

The described models fall under the generalized linear model (GLM) family and are used in the analysis of variance (ANOVA) scenario. The null models assume no effect of a condition on the exonic part and hypothesis test is used to quantify the significance of observed effects. For each gene independently, we define models of read counts and percent spliced-in (PSI) related to its non-overlapping exonic parts, defined the same way as in DEXSeq.

Let e be the exonic part of interest, which includes exons, alternative splice sites, retained introns. We use c to denote an experimental condition. Both count- and PSI-based models assume the same design matrix while differing in the distributional assumptions of the observed data.

2.1 The read count model

The number of reads Y_{ec} mapping to exonic part e upon condition c is distributed according to a negative binomial (NB) distribution:

$$Y_{ec} \sim \mathcal{NB}(s_c \mu_{ec}, d_e) \tag{1}$$

where μ_{ec} denotes the expected count, s_c the size factor particular of a condition and related to depth of sequencing, and d_e the dispersion (extra-Poisson variation) of each exonic part. In order to infer regression parameters of NB-distributed data, the dispersion d_e is assumed to be known, while the mean is parametrized using the log-link function:

$$\log(\mu_{ec}) = \beta_e + \beta_c \tag{2}$$

where β_e, β_c represent exonic part- and condition- specific parameters. The alternative models

$$\log(\mu_{ec}) = \beta_e + \beta_c + \delta_{ee'}\delta_{cc'}\beta_{e'c'} \tag{3}$$

where δ is the Kronecker delta function and $\beta_{e'c'}$ represents an effect of e' on condition c'. The parameters of the alternative model are inferred once for each candidate *interaction* (pair e' and c'). Since the alternative model (Eq. 3) is a more general case of the null model (Eq. 2), the likelihood-ratio test can be used to assess the significance of the interaction paramterer $\beta_{e'c'}$ (Dobson and Barnett, 2008). Additional factors (such as library type, batch number, cell type) can be added to models 2-3.

The genes are assumed to be independent groups of exonic parts and parameters are fitted for each gene independently; for a gene with n_e exonic parts and expression in n_c conditions, we perform $n_e \times n_c$ tests.

The significance scores are used to produce a ranked-list of likely effects of conditions on the exonic parts. Instead of testing all possible pairs, there may exist a subset of exonic parts or conditions of interest to be tested while the remaining pairs are used as reference.

The hyperparameters - size factors s_c and dispersion d_e - are assumed to be known prior to estimation of $\boldsymbol{\beta}$. Various methods exists for both parameters. For size factors, we use the function `calcNormFactors` provided by the edgeR package. The discussion on dispersion estimation in NB regression models is presented in Section 3.1.1.

2.2 The Percent spliced-in model

Percent spliced-in (PSI) is a measure of expression alternative to read counts and represents the ratio of exon inclusion in all transcripts of a gene.

PSI is derived for each exonic part as a ratio of reads overlapping the exonic part versus total reads overlapping the upstream and downstream splice junctions. The division that is involved also acts as an implicit form of normalization, making the PSI values comparable across exonic parts. The PSI can be used to quantify all splicing events, such as alternative 5'/3' ends, intron retention and others.

Beta regression is used to model data in form of fractions (Ferrari and Cribari-Neto, 2004; Smithson and Verkuilen, 2006), since the support of the beta distribution is an open real interval, e.g., $(0, 1)$. The model uses a familiar parametrization; the PSI of an exonic part e in condition c is distributed according to a beta distribution with mean μ_{ec} and precision Φ, i.e., $\Psi_{ec} \sim \text{Beta}(\mu_{ec}, \Phi)$.

The link function for the mean is the $\text{logit}(x) = \log(\frac{x}{1-x})$, which is a mapping $(0, 1) \to (-\infty, \infty)$. The null model is thus given as

$$\text{logit}(\mu_{ec}) = \beta_e + \beta_c, \tag{4}$$

and the alternative model includes the interaction factor $\beta_{e'c'}$:

$$\text{logit}(\mu_{ec}) = \beta_e + \beta_c + \delta_{ee'}\delta_{cc'}\beta_{e'c'} \tag{5}$$

The definitions are analogous to model in Eqs. 2-3 and the significance of the effect of exonic part e' on condition c' is again evaluated by means of the likelihood-ratio test. Due to implicit normalization, the model uses a single precision parameter Φ. Other options such as a precision model are possible subject to additional computational cost.

2.3 Efficient parameter estimation

Closed-form solutions for likelihood-maximizing parameters for the presented generalized linear models do not exist. To fit both count and PSI models, we use the Iterative-reweighted least-squares (IRLS) algorithm with Levenberg-Marquardt damping (Press, 2007), with time complexity of $O((n_e + n_c)^3)$ per single model fit. Here, we briefly describe a model approximation scheme and refer the reader to Suppl. Section 1.1 for a more complete treatment.

Let the parameter vectors $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_a$ be the local maximizers of the log-likelihood for null and alternative models, respectively. Parameter estimation for the alternative model is repeated $n_e \times n_c$ times for each candidate interaction (pair e', c'; Suppl. Fig. 1). As the number of conditions grows with n_c, the number of model parameters and the number of fitted values grow with $n_e + n_c$ and $n_e \times n_c$, respectively. This leads in an increased number of measurements associated to an exon factor. Consequently, an increasing proportion of parameter values tends to be approximately equal in the null and the alternative models. In other words, the mean-squared error (the distance) between corresponding components of $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_a$ decreases with n_c (Supplementary Fig. 2a-2b).

We propose an approximation scheme based on the Wald test for parameter significance to decrease the number of parameters for each alternative model fit. Briefly, having obtained a maximum likelihood estimate of $\boldsymbol{\beta}_0$, the value of each component $\boldsymbol{\beta}_{0,i}$ is compared to its estimated variance σ_i^2, leading to a test statistic

$$w_i = \frac{\beta_{0,i}^2}{\sigma_i^2} \tag{6}$$

which follows a χ^2 distribution with one degree of freedom. This way, the parameters significantly different from zero can be identified, for a user-defined significance level α. The factors corresponding to the remaining parameters - non-significantly different from zero - are collapsed into a single *residual* factor ϵ_{ec}. The new design matrices for the null and the alternative models are defined by

$$\begin{aligned} \text{link}(\mu_{ec}) &= \beta_e + \beta_c + \epsilon_{ec} \\ \text{link}(\mu_{ec}) &= \beta_e + \beta_c + \epsilon_{ec} + \delta_{ee'}\delta_{cc'}\beta_{e'c'} \end{aligned} \tag{7}$$

where β_e or β_c are nonzero if and only if they have been identified as significant by the Wald test (Eq. 6). For the null model, such definition does not affect the likelihood of the original null model. On the other hand, alternative model has a reduced capacity, retaining only the significant parameters. The maximum likelihood of the reduced model is upper bounded by that of original alternative models (Eq. 3 and Eq. 5), which affects statistical power, but does not increase the probability of Type I error, which is desired for our application. The conditions for applying the likelihood-ratio test are still met as the reduced null model is a special case of the reduced alternative model. This can be seen as a low-rank approximation of the design matrix where non-significant columns are combined into one. The effect of model reduction on timing properties, retrieval accuracy and the model capacity is investigated in Section 3.1.3.

3 Results and discussion

3.1 Experiments on generated data

To assess the accuracy of retrieval of differentially used exonic parts, and investigate the effect of increasing number of conditions, we perform experiments with artificially generated data for count- and PSI-based models.

The number of exons is set to $n_e = 60$ and the number of conditions varies, $n_c \in [3, 5, 10, 30, 50]$. The sampling distributions for size factors, exonic part, condition, exon-condition interaction parameters and size factors are determined to assume comparable values to real-world datasets (data not shown), and are defined as

$$\begin{aligned} s_c &\sim \Gamma_1(1, 2) \\ \beta_e, \beta_c &\sim \mathcal{U}(-1, 2) \\ \beta_{e'c'} &\sim \mathcal{U}([-8, -6] \cup [1, 2]), \end{aligned} \tag{8}$$

where $\Gamma_1(a, b)$ is the gamma distribution with one degree of freedom, and $\mathcal{U}(a, b)$ is the uniform distribution on $[a, b]$. The number of ground-truth interacting pairs e', c' is set to 5% of the total number of exons and selected at random. Finally, the observable values Y_{ec} and Ψ_{ec} are sampled according to the models in Eq. 3 and Eq. 5, respectively. To compute standard deviations, 30 replicate datasets are generated with two replicates per condition.

3.1.1 Dispersion estimation

The variance of negative-binomial distributed data is dependent on the mean and dispersion as $\text{var}[Y_{ec}] = \mu_{ec} + d_e\mu_{ec}^2$. The values of exonic part-specific hyperparameters d_e are required prior to model fitting. The

Cox-Reid dispersion estimate, a method proposed for small number of conditions (or replicates), used by the DEXSeq package, provides a conservative estimate, taking the maximum of model fit and individual exon estimate. The advantage of large number of conditions provides greater statistical power when estimating hyperparameters.

Due to unequal library sizes across conditions, the counts Y_{ec} are sampled from distributions with different means. We use a quantile-adjusted conditional ML (qCML) to generate identically distributed pseudodata and derive a common estimate (Robinson *et al.*, 2010).

Suppl. Fig. 4 shows the results on generated data, which confirm qCML is the least biased and outperforms the Cox-Reid estimate with increasing number of conditions (at $n_c = 50$; qCML RMSE 0.92, Cox-Reid RMSE 2.90). The qCML model overestimates the low dispersion values, as shown on Suppl. Figs. 4a-d. This favours Type II error (false negative) over Type I error, which is a suitable trade-off for differential expression tests in genomic studies.

3.1.2 Retrieval of condition-specific differentially expressed exons

Next, we evaluate the accuracy of the differential exon expression retrieval by each method. The statistical significance scores produced by each method are used to rank the exons and receiver-operating characteristics (ROC) are estimated with respect to the ground truth. The area under the ROC curve (AUC) is used as a measure of retrieval accuracy.

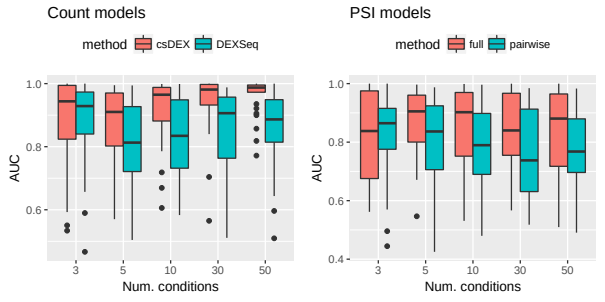
To this end, we demonstrate the advantage of modeling all experimental conditions jointly versus exhaustive pairwise *condition vs. control* comparisons. For count data, DEXSeq is run once per each condition, versus an arbitrarily chosen control condition (which is the same for all comparisons). For PSI data, csDEX is run jointly for all conditions (labeled *full*) and, similarly, once per each condition versus control (labeled *pairwise*). The statistical significance scores are reported using scientific notation (i.e., 'aE-x' = $a \times 10^{-x}$).

The results are shown in Fig. 1. When detection of condition-specific changes in exon expression are desired, csDEX performs equally or better than DEXSeq for an increasing number of conditions, e.g. at $n_c = 5$, csDEX mean AUC=0.865 and DEXSeq AUC=0.813 (p-value=6.6E-2, Student t-test), and for $n_c = 50$, the mean AUC scores are 0.962 and 0.859, respectively (p-value=8.6E-5). For PSI-based models which have an upper bounded variance, the differences are in mean AUC are less drastic, but still statistically significant as overall mean AUC=0.847 (full model) and AUC=0.789 (the pairwise model, p-value=2.eE-4, Student t-test).

The retrieval accuracy varies greatly between the number of conditions due to different, independently generated datasets with a low number of replicates. To disambiguate the effect of differences in sampling, AUC scores for the same dataset replicates are compared pairwise. For each comparison, csDEX consistently produces a higher AUC than DEXSeq (Wilcoxon signed rank test, p-value=3.4E-13, Fig. 1b, label all). This property is preserved when the tests are confined to datasets with the same numbers of conditions. The number of times csDEX outperforms DEXSeq increases with n_c, e.g., for $n_c = 3$, p-value=5.7E-02, and for $n_c = 50$, p-value=1.9E-07. The same effect is observed for PSI-based models, where for $n_c = 3$, p-value=6.3E-01, and for $n_c = 30$, p-value=2.0E-05, and overall p-value=5.7E-08 (Fig. 1c). Together, these results argue that joint modeling of multiple conditions increases the statistical power of the test to detect condition-specific changes.

3.1.3 Comparison of full and reduced models

Finally, we compare the effect of design matrix approximation scheme, described in Section 2.3. To assess the retrieval accuracy measured by AUC, a synthetic dataset with $n_c = 10$ is used. The significance threshold α affects Type I error probability, but did not significantly affect the retrieval accuracy (change in AUC < 0.0008 for PSI model and < 0.005 for the count model between $\alpha = 1$ and α=1E-10), as shown on Suppl. Fig. 3). On the other hand, decreasing α leads to a significant reduction in the number of model parameters (30 down to an average of 5.2 and 2.0 for the PSI model and the count model, respectively). Consequently, the running time decreases accordingly (3.48-fold reduction for the PSI model, and 3.21-fold for the count model). Together, these results show the savings with usage of reduced models on large-scale biological datasets.



n_c	csD.	DEXS.	p_W	p_t	full	pairw.	p_W	p_t
3	0.88	**0.89**	5.7E-02	5.7E-1	0.82	**0.83**	6.3E-1	6.7E-1
5	**0.87**	0.81	8.7E-03	6.6E-2	**0.87**	0.80	1.2E-3	1.9E-2
10	**0.93**	0.81	8.0E-06	1.1E-3	**0.86**	0.78	3.6E-4	1.4E-2
30	**0.94**	0.85	1.2E-06	2.1E-3	**0.85**	0.76	2.0E-5	6.4E-3
50	**0.96**	0.86	1.9E-07	8.6E-5	**0.84**	0.78	1.3E-2	6.7E-2
all	**0.91**	0.85	3.4E-13	2.0E-6	**0.85**	0.79	5.7E-8	2.3E-4

Fig. 1. Accuracy of retrieval for known interactions on generated data is measured by Area under ROC curve (AUC) for increasing number conditions. Tables show mean scores for csDEX and DEXSeq (for count data) and full versus pairwise models (for PSI data). The statistical significance of differences in means is performed with one-sided Wilcoxon signed rank test (p_W) and one-sided Student t-test (p_t).

3.2 Experiments on ENCODE RNA-seq datasets

To evaluate the methods, we use publicly available RNA-seq data from the ENCODE project (Consortium, 2004). We perform differential analysis of non-overlapping exonic parts in a defined subset of genes. The discovered differentially used exonic parts are analysed for enrichment of known splicing events, RBP binding events and RNA motifs.

3.2.1 Experimental setup

We obtained 208 RNA-seq experiments on 189 individual RNA-binding protein knockdowns (shRNA interference, denoted shRNA+RNA-seq) and 19 controls in a human immortalised myelogenous leukemia line K562, provided by the ENCODE consortium (Suppl. Table 4). The number of reads mapping to each exonic part is extracted from BAM files aligned to the hg19 genome using the script `dexseq_count.py` (package DEXSeq). Provided transcript quantifications files are used to compute PSI values, as a ratio of isoforms including an exonic part versus all isoforms (Suppl. Table 5). Similarly, we downloaded BAM files for 244 experiments aligned to the hg38 genome in order to compare the results with a study of TARDBP-regulated cryptic exons (Suppl. Table 6).

A subset of the human gene annotation file (GTF, Ensembl hg19) was prepared to evaluate the precision of annotated splicing events retrieval. The selected genes contain at least 5 and up to 15 unique exons, contain at least one *cassetteExon* and do not overlap with any other gene (Suppl. Table 2). The associated exons are split into non-overlapping exonic parts (GFF file) using the script `dexseq_prepare_annotation.py`. The final GFF file contains 1,073 genes and 11,047 unique exonic parts.

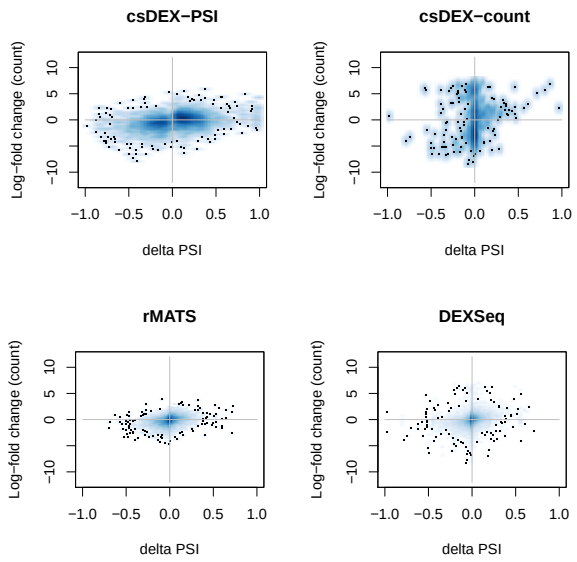
This smaller dataset is used to efficiently evaluate differential exon usage methods.

Annotated alternative splicing events are extracted from the UCSC Genome Browser track *knownAlt*, consisting of eight types of events: *altFinish*, *altFivePrime*, *altPromoter*, *altThreePrime*, *bleedingExon*, *cassetteExon*, *retainedIntron*, *strangeSplice*. We assign an exonic part to any of the above eight *knownAlt* event categories based on partial overlap. Since *knownAlt* annotation is not complete, we additionally denote an exonic part *alternative* if it is not a part of every known transcript for a given gene.

Similarly, we create a larger annotation with genes containing at least 13 and up to 66 unique exons (Suppl. Table 3). The bounds are selected to represent the distribution of the number of unique exons per gene; for genes with at least one *cassetteExon*, the distribution of the number of exons (in $\log_{10}$ scale) follows a shape close to the normal distribution, with the interval $[13, 66]$ a one standard deviation away from the mean of 29.88 exons (Suppl. Fig. 5). The annotation contains 139,759 exonic parts in 4,789 genes. This larger dataset is used to examine RBP binding and motif enrichment with greater statistical power.

We obtained 89 experiments on genome-wide RBP binding, determined with ENCODE cross-linking and immunoprecipitation (eCLIP) protocol in same cell line K562 (Suppl. Table 7). In total, there is 69 RNA-binding proteins with both shRNA+RNA-seq and eCLIP experiments available. For each exonic part, we count the number of overlapping eCLIP reads.

Data from the Catalog of Inferred Sequence Binding Preferences of RNA binding proteins[1] (CISBP) is included to corroborate the direct binding evidence (Ray *et al.*, 2013). The RNA motifs are obtained with the *in vitro* assay RNAcompete and are represented as position-probability matrices of 4-11 nt in length. Out of 69 RBPs with shRNA knockdowns and binding evidence (eCLIP), there are 18 with known, direct motif evidence in the CISBP.



	N	ρ_P	ρ_S	P(s)	P(w)
csDEX-PSI	5188	**0.34**	0.33	**66%**	**55%**
csDEX-count	452	0.14	0.11	50%	15%
rMATS	7851	0.31	**0.35**	60%	10%
DEXSeq	17102	0.24	0.31	56%	8%

Fig. 2. Agreement between actual change in PSI ($\Delta\Psi$) and change in read counts (ΔY) for interactions (pairs e, c) retrieved by the compared methods (at FDR$< 10\%$). Legend: N number of retrieved interactions; ρ_P / ρ_S Pearson/Spearman correlation; $P(s)$ percentage of interactions equal up to sign, $\text{sign}(\Delta\Psi) = \text{sign}(\Delta Y)$; $P(w)$ percentage of significant changes in $\Delta\Psi$.

3.2.2 Effects of quantification method on retrieved interactions

The variability in read counts can arise both from differences in expression on a gene level and from splicing of individual exonic parts. The disambiguation between the sources of variability is made harder if expression of exonic parts is quantified independently (of other exonic parts). Conversely, the quantification based on PSI implicitly includes all exonic parts of a transcript. Fig. 2 shows the exonic part and condition pairs with FDR$< 10\%$ retrieved by csDEX-count, DEXSeq (count-based models) and csDEX-PSI, rMATS (PSI-based models). To assess the dependence between actual percent-spliced in Ψ and read count Y, each significant interaction - exonic part and condition pair (e, c) - is placed in a plane spanned by

- $\Delta\Psi$; difference between Ψ_{ec} and average $\Psi_{ec'}$ (x-axis), and
- ΔY; $\log_2$ fold difference between read count Y_{ec} and average $Y_{ec'}$ (y-axis),

where the averages are computed over all conditions c'. Unsurprisingly, the two quantities are proportional, as seen by positive correlation coefficients, which are significant for all four methods (p<2E-14%). Using the PSI-based models csDEX-PSI and rMATS results significantly stronger correlation compared to count-based models. It confirms that significant change in PSI implies a perceived change in read counts, but not vice versa. This is supported by counting the interactions having the same sign in both $\Delta\Psi$ and ΔY: 66% and 60% for PSI-based; 50% and 56% for count-based models, with csDEX-PSI reporting the strongest agreement between perceived changes in read counts and PSI.

The Wald Test is used to assess the degree of condition-specific changes; the magnitude of $\Delta\Psi$ for each exonic part is compared to its standard deviation across all conditions. The largest percentage (55%) of significant changes is reported by csDEX-PSI. Surprisingly, the percentage of csDEX-count (15%) is slightly larger than rMATS (10%) and almost twice as large as DEXSeq (8%), supporting the advantage of including condition-specific model parameters.

From a similar perspective, each of the retrieved interaction lists is used to estimate a distribution of *uniqueness score* (Suppl. Fig. 6). For each significant pair (e, c), the count of conditions c' where the same feature e is also significant under the same FDR threshold of 10%. Expectedly, the proposed condition-specific models are heavily skewed towards zero (csDEX-PSI: 14.3 ± 15.9; csDEX-count: 0.4 ± 0.9) compared to non-condition specific models (rMATS: 27.1 ± 16.8; DEXSeq: 30.1 ± 22.0).

Together, this comparison illustrates the robustness of PSI-based models for retrieving changes in alternative splicing as the count values can be affected by gene expression and read coverage effects. Finally, the inclusion of condition-specificity in model definition enables the retrieval of unique changes particular to a small subset of conditions.

3.2.3 Precision of annotated splicing events retrieval

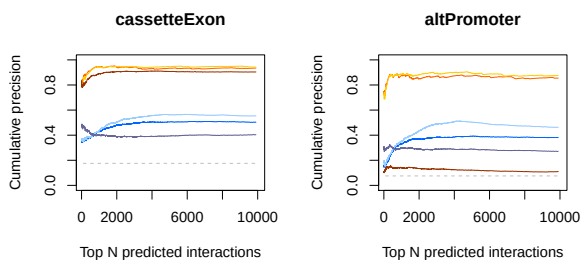
Obtaining ground truth information presents a major challenge in evaluation of differential expression methods. To quantitatively evaluate the lists of retrieved candidate interactions by each method, we use the UCSC *knownAlt* annotation as ground truth for independent validation. Note that the *knownAlt* annotation is not presented to any of the methods, nor does it influence annotation of non-overlapping exonic parts (the GFF file), but it is used only for initial selection of genes.

[1] `http://cisbp-rna.ccbr.utoronto.ca/`

For each method, we select the top 10,000 most significant interactions. For each of the nine AS event types listed in Section 3.2.1, we compute the cumulative precision for each possible significance cut-off; precision is defined as the number of exonic parts annotated with the particular AS event (positives) versus constitutive exonic parts (negatives). We assume that if an exonic part is deemed differentially spliced, the agreement with an existing, independently annotated AS events supports the case (true positive). Conversely, constitutive exonic parts (false negatives) may still be identified as differentially spliced due differences in read coverage/gene expression, incomplete annotation, or due to inherent uncertainty associated to model fitting.

Not surprisingly, all six evaluated methods display an average cumulative precision greater than the overall probability of the particular AS event type, confirming that the retrieved lists are non-random (Fig. 3). For example, there is 72% of exonic parts annotated with any AS event (i.e. of category *alternative*), while all of the methods score above 80%. For all nine categories, the PSI-based models (shown in warm colors) outperform the count-based models by a large margin. This finding is supported by the results in Section 3.2.2, showing that the variance in counts does not imply variance in PSI. The quantification with PSI is explicitly dependent on whether a feature is present in all transcripts, since constitutive features will have PSI of either 0 or 1. Nevertheless, the condition-specificity and a larger set of reference conditions implied in csDEX-PSI (prec.=98%) improves the precision also over rMATS (prec.=91.5%). The csDEX-count model improves over DEXSeq in all nine categories, confirming that a difference in count should be compared to a large set of conditions to be deemed significant. The two most abundant AS event categories are shown on Fig. 3, bottom panel. The cumulative precision first exhibits a rapid rise, indicating higher abundance of true positives towards the top of the lists, reaching the prior probability as the list length approaches the full dataset size. The cumulative precision plots for all AS event types are shown in Suppl. Fig. 7. The Wald Test-based model approximations to csDEX-PSI and csDEX-count, presented in Section 2.3 and indicated with a '*' exhibit a comparable performance to the corresponding exact models, while slightly improving precision, confirming that this type of model approximation can further decrease the probability of Type I error. The reduced models additionally decrease running time over 3-fold (Suppl. Fig. 2c-2d), with median time of 0.63 seconds per model fit. In summary, the condition-specific csDEX-PSI and csDEX-count improved precision over its *condition vs. control* counterparts and retrieved plausible (exonic part, condition) pairs representing annotated AS events by an independent source.

	csDEX-PSI	csDEX-PSI*	rMATS	csDEX-count	csDEX-count*	DEXSeq
alternative	**.98**	**.98**	.915	.806	.817	.807
altFinish	**.14**	.13	.004	.024	.023	.019
altFivePrime	.49	**.53**	.217	.084	.094	.056
altPromoter	.86	**.88**	.122	.368	.453	.285
altThreePrime	.47	**.51**	.256	.086	.203	.047
bleedingExon	.77	**.80**	.350	.238	.252	.226
cassetteExon	.93	**.94**	.900	.487	.530	.399
retainedIntron	.54	**.58**	.185	.101	.104	.069
strangeSplice	.13	**.14**	.000	.021	.173	.007



Best viewed in color.

Fig. 3. Precision of recovering annotated AS events from UCSC knownAlt track asessing the top 10,000 of most significant interactions (pairs e', c') retrieved by each method. The cumulative precision for events cassetteExon and altPromoter is shown in detail. (dash: prior probability, *: reduced model).

3.2.4 Condition-specific regulated exonic parts are enriched in RBP binding and motifs

We analyze the condition-specific, regulated exonic parts in context of two independent data sources: i) the RBP binding protein occupancy (eCLIP) and ii) motif analysis (CISBP data), obtained as described in Section 3.2.1. For the experiments in this section, we use the larger annotation consisting of genes with 13 to 66 unique exons (Suppl. Table 3).

To obtain a high-quality list of regulated exonic parts for each condition, we infer parameters of the csDEX-PSI for each batch of experiments. A batch is defined by a common control experiment (mock shRNA with a non-specific target) and performed on the same date (Suppl. Table 1). Principal component and Multi-dimensional scaling analyses (PCA and MDS) of samples reveal a higher similarity of samples within the same batch comparing to average similarity between batches (Suppl. Figs. 8-10). This suggests a strong possibility of *batch effects* (Leek *et al.*, 2010), prompting us to perform separate csDEX tests per each batch (batch design). A batch effect factor could in principle be included into the model definition. However, ENCODE datasets includes only repeats of control experiments within each batch. The repeats of RBP knockdowns in different batches are not included, limiting the ability to infer batch-specific model parameters. We compare the results obtained by the batch design to a classic *condition vs. control (pairwise) design* for each surveyed factor individually.

We fit a reduced csDEX-PSI model to each gene (parameter significance cut-off $\alpha = 0.05$, see Section 2.3). For each of the 18 conditions (RBP knockdowns with provided eCLIP and motif data), we select significantly regulated *alternative* exonic parts (FDR<10%). For each condition, we sample an additional 20,000 non-significantly regulated exonic parts (FDR=1) to serve as a background set (backg.). We refer to an exonic part as *upregulated*, when its expression is *reduced* in the RBP knockdown, i.e. its expression is lower than predicted by the null model. The definition of *downregulated* exonic parts is analogous.

We examine the fold enrichment in binding and motif probability at 300 nt regions centered at the 3' splice sites (3' SS), whose recognition by the spliceosome is affected by many surveyed RBPs, such as U2AF2, PTBP1, SR proteins or hnRNPs. The differences in binding were very subtle, and hardly detectable by Proportion or Hypergeometric enrichment tests when looked at in a *position non-specific* manner (data not shown), presumably due to previously observed high noise and sparsity in of the eCLIP tags (Haberman *et al.*, 2017). Therefore, we examined *position-specific* enrichment at individual nucleotide positions within the defined regions. The exact derivation of the enrichment scores is described in Suppl. Section 2.4. Briefly, the binding enrichment at nucleotide position i is defined as a ratio of probabilities (odds) of an eCLIP tag presence at regulated (reg.) versus background (backg.) set of regions. Similarly, the motif enrichment is defined as a ratio of motif scores (agreement of the sequence with a given motif position-probability matrix). The alignment between binding and enrichment signals (vectors of fold enrichment within 300 nt regions) is scored using *cross-correlation*, defined as a maximal

Pearson correlation when one of the signals is allowed to be shifted by at most 50 nt.

The results for a subset of binding and motifs patterns is presented on Fig. 4, with the complete set of RBPs is shown on Suppl. Figs. 11-12 and Suppl. Tables 12-13. Out of the 18 surveyed RBPs, nine display high cross correlation (>0.15) for both up- and down- regulated exonic parts, four only in upregulated and five only for downregulated exonic parts. Comparing the results of *batch design* in to *condition vs. control* design, both eCLIP and motifs signals display higher positive enrichment. Twelve RBPs display an eCLIP signal enrichment of at least 1.5-fold in at least one regime, whereas no comparable enrichment is in witnessed in the condition vs. control design.

The U2 auxiliary factor (U2AF2) is a core spliceosomal component that plays a role in 3' splice site recognition (Fu *et al.*, 2014). It binds polypyrimidine tracts and interacts with the U1 protein for initial exon definition. We observe highest enrichment in neat 3' splice sites of the 1641 downregulated parts (N=1641; max. 1.59-fold eCLIP signal enrichment, 70 nt downstream of 3' SS) and cross-correlation of 0.36 with motif enrichment signal. Thus, binding of U2AF2 at the pre-mRNA defines a splice site, causing the directly bound region to likely be excluded from the transcript. The specifically upregulated exonic parts (N=925) display a somewhat lesser 1.44-fold binding enrichment and cross-correlation with motif signal of 0.23.

The polypyrimidine tract-binding protein (PTBP1) antagonizes U2AF2 binding, interfering with functional recognition of 3' splice sites and preventing spliceosome assembly (Sharma *et al.*, 2008). Our results agree with this proposition; for the 1,114 downregulated exonic parts, we observe 1.8-fold motif and 1.46-fold binding enrichment -15 nt upstream of the 3' splice sites, with 0.39 cross-correlation. A similar binding and motif enrichment is observed at 1,286 upregulated exonic parts, however with a much smaller cross-correlation of 0.1. The highest motif enrichment is observed at -8 nt of the 3' SS, suggesting a possible successful recognition of the intron 3' end by U2AF2 and consequently a successful inclusion into the end transcript.

The *KH domain containing, RNA binding, signal transduction associated 1 protein* (KHDRBS1), also identified as Sam68, is generally associated to splicing activation by i) binding exonic splicing enhancers or ii) enhancing the binding of U2AF2 to alternatively spliced pre-mRNA (Matter *et al.*, 2002; Tisserant and Konig, 2008). For the 634 upregulated exonic parts, this is confirmed by a peak 2.51-fold binding enrichment at -8 nt around the 3' splice site and an average of 1.78-fold enrichment within the [-150, +150] nt region proximal to the 3' SS, with somewhat weaker motif enrichment. The 730 downregulated exonic parts display a weaker binding enrichment pattern downstream of the 3' SS. Together, this confirms the role of KHDRBS1 as an splicing activator in both intronic and exonic regions.

The various roles of *SF3B complex* (SF3B) include recognition of branch point adenosine. This enables SF3B to present a temporary steric barrier to branch point sequence prior to activation, preventing pre-mature splicing and promotion of stable interaction for U2 and U11/U12 di-snRNP to pre-mRNA (Lardelli *et al.*, 2010; Rakesh *et al.*, 2016). For the 1,086 downregulated exonic parts, the binding enrichment pattern expectedly mimics that of U2AF2, binding at the 3' splice site motif CAAAG (cross-correlation 0.44) and exhibiting largest enrichment at the starts of exonic parts, 38 nt downstream from the 3' SS.

The *Serine and arginine rich splicing factors* (SRSFs) largely bind exonic splicing enhancers, hence its binding and alignment with the signal is present mainly in the exonic regions (Fu *et al.*, 2014). This is exemplified by e.g. the SRSF7 protein, with 1.41-fold binding enrichment in intronic and exonic parts around the 3'SS of the 549 upregulated exonic parts, and 0.63 cross-correlation of binding and motif signal. Conversely, the downregulated exonic parts display a slight depletion, with the binding probability of 0.78-fold from the expected.

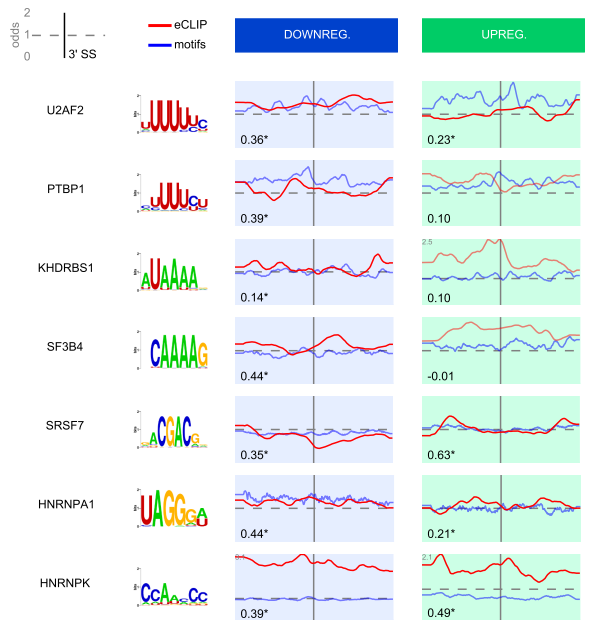


Fig. 4. Fold enrichment (odds) of binding and motif score probabilities when comparing up- and down- regulated exonic parts, against the background set. The plots show [-150, 150] nt regions centered at 3' splice sites (3' SS), represented by a black vertical line. The gray dashed line represents fold enrichment of 1 (i.e. no enrichment). The numbers represent values of cross-correlation with maximum allowed displacement of 50 nt. Line plots are shown in darker color when the cross-correlation > 0.15.

The *Heterogeneous nuclear ribonucleoproteins A1 and K* (HNRNPA1 and HNRNPK) affect splicing by binding both exonic and intronic regions. Both RBPs display binding enrichment in the upstream and downstream of the 3'SS. Data for both proteins report more downregulated exonic parts (1261 for HNRNPA1; 1938 for HNRNPK) than upregulated (934 for HNRNPA1; 1164 for HNRNPK). For HNRNPA1 we observe a higher binding enrichment of 1.35-fold at the downregulated exonic parts, with cross-correlation of 0.44; For HNRNPK, we observe a stronger 3.1-fold enrichment and cross-correlation of 0.39. This is consistent with the knowledge that HNRNPs can cause both exon inclusion and exclusion when binding to upstream intronic splicing silencers.

Together, the concordance with known binding patterns present a positive control, and present an indirect, but relevant evidence to the quality of retrieved exonic parts.

3.2.5 csDEX successfully retrieves TARDBP cryptic splicing events

A further validation of csDEX ability to discover condition-specific splicing changes was performed vis-a-vis a thoroughly investigated unannotated exons arising upon silencing of the TARDBP gene. The long length and reduced evolutionary conservation of intronic sequences contribute to an increased probability of emergence of novel 5' and 3' splice site pairings, resulting in the so-called *cryptic exons* or *pseudoexons* (Ling *et al.*, 2015). In a recent study, TARDBP-regulated cryptic splicing was investigated within nine human and mouse RNA-seq datasets, including the K562 samples obtained from ENCODE (Humphrey *et al.*, 2016). The authors report on a confident list of 84 cryptic exons that undergo increased expression upon TARDBP depletion. In contrast, the changes in expression of the same cryptic exons are not observed upon depletion of FUS or hnRNPC, the two proteins also related to ALS and cryptic exons,

suggesting TARDBP-specific regulation. The causal relationship between TARDBP binding and direct effect on splicing was further supported by RBP-RNA interactions (iCLIP and eCLIP) as well as the enrichment of characteristic GU-rich RNA motifs.

We used a provided GFF annotation file with a total of 204,961 exonic parts, out of which 11,919 *cryptic exonic parts* were not part of standard Ensembl annotation (Suppl. Table 8). Using the read alignment (BAM) data for the hg38 annotation, we run the csDEX-count model with 244 different RBP knockdowns (Suppl. Tables 6). The list of 84 high confidence cryptic exons was used as a positive control (Humphrey *et al.* (2016), Suppl. Table 9, retrieved on 28. 3. 2017). To verify whether the cryptic exons would be detected by csDEX, we test only for TARDBP specific changes, while using the data for all 244 conditions for parameter inference (i.e. the condition c' always corresponds to TARDBP in alternative models, Eq. 3). The predicted TARDBP-specific changes are ranked by statistical significance and selected subject to a FDR$< 5\%$ threshold.

First, we verified that the 11,919 cryptic exonic parts are detected upon TARDBP depletion. Among all tested 204,961 exonic parts, the cryptic exonic parts tend to be enriched at the top of the ranked list with probability of 30.2% among the selected 331 exonic parts with FDR$< 5.8\%$, comparing to 5% overall probability (Suppl. Fig. 13).

Furthermore, we examine the subset of 11,416 testable cryptic exonic parts to evaluate the retrieval accuracy of 74 testable cryptic exons (positive control). There are 46 predicted exonic parts (FDR<5%), where 42 are downregulated in the wild-type conditions, confirming that TARDBP generally silences the cryptic exons. Out of 46, there are 20 correctly retrieved cryptic exons (true positives, TP). The fraction of true cryptic exons within the selected ones is significantly higher with 43.3% comparing to 0.4% in the remaining positions (p-value<1E-31, Hypergeometric test; Suppl. Fig. 14). All the true positive (TP) and false negative (FN) examples nevertheless tend to be ranked higher than expected by chance (p-value<1E-13, Wilcoxon rank sum test, blue and red points respectively, Suppl. Fig. 14). Together, the results support the ability of csDEX to detect biologically meaningful, sparsely present cryptic exons, regulated specifically by the TARDBP gene.

4 Conclusion

Post-transcriptional gene regulation has a comparable effect to gene expression regulation on transcriptome and proteome diversity. With ongoing refinement of experimental protocols, the ability to precisely detect differential expression on a sub-transcript level is ever increasing.

Here, we present csDEX, a statistical modeling package for detection of condition-specific alternative splicing. By comparing count- and Percent spliced-in (PSI) based expression quantification, we highlight the proneness of count-based models to detect changes in gene expression rather than alternative splicing. csDEX provides both PSI- and read count-based models within the family of generalized linear models, focusing on condition-specific changes. The retrieved exonic parts from a case study involving more than 200 RNA-seq samples are compared to multiple, independent positive controls, enabling the quantification of retrieval accuracy of related models. When compared to annotation of known alternative splicing events, the csDEX model based on PSI quantification proves to perform with close to 90% accuracy, while both types of csDEX models retrieve splicing changes with lowest overlap between conditions. The predictions are further validated with indirect data sources, such as RBP binding evidence and motif analysis, as well as known cryptic exons arising upon silencing the TARDBP gene.

As the cost of sequencing and computational resources steadily decreases, large-scale multi-factor models can be used to underpin experimental pipelines. The current ENCODE RNA-seq+shRNA KD experiments do not provide sufficient control over sources of variability, such as batch effects, to infer model parameters using all the available data on hundreds of conditions. Careful experimental design is needed to isolate condition-specific AS regulation.

Nevertheless, joint modelling of multiple conditions within a batch improved found binding and RNA motifs patterns over case vs. control design. Together, our experimental findings favour csDEX as the differential exon expression method of choice when condition-specific changes are desired.

References

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome research*, **22**(10), 2008–17.

Christinat, Y., Pawlowski, R., and Krek, W. (2016). jSplice: a high-performance method for accurate prediction of alternative splicing events and its application to large-scale renal cancer transcriptome data. *Bioinformatics*, pages btw145–.

Consortium, E. P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**(5696), 636–40.

Dobson, A. J. and Barnett, A. (2008). *An introduction to generalized linear models*. CRC press.

Ferrari, S. and Cribari-Neto, F. (2004). Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics*, **31**(7), 799–815.

Fu, X.-D., Ares, M., and Ares Jr, M. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics*, **15**(August), 689–701.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, **17**(6), 333–351.

Griffith, M., Griffith, O. L. O., Mwenifumbo, J., Goya, R., Morrissy, a. S., Morin, R. D., Corbett, R., Tang, M. J., Hou, Y.-C., Pugh, T. J., Robertson, G., Chittaranjan, S., Ally, A., Asano, J. K., Chan, S. Y., Li, H. I., McDonald, H., Teague, K., Zhao, Y., Zeng, T., Delaney, A., Hirst, M., Morin, G. B., Jones, S. J. M., Tai, I. T., and Marra, M. a. (2010). Alternative expression analysis by RNA sequencing. *Nature methods*, **7**(10), 843–7.

Haberman, N., Huppertz, I., Attig, J., König, J., Wang, Z., Hauer, C., Hentze, M. W., Kulozik, A. E., Le Hir, H., Curk, T., Sibley, C. R., Zarnack, K., and Ule, J. (2017). Insights into the design and interpretation of iCLIP experiments. *Genome Biology*, **18**(1), 7.

Hartley, S. W. and Mullikin, J. C. (2016). Detection and Visualization of Differential Splicing in RNA-Seq data with JunctionSeq. *Nucleic acids research*, pages 1–38.

Hu, Y., Huang, Y., Du, Y., Orellana, C. F., Singh, D., Johnson, A. R., Monroy, A., Kuan, P. F., Hammond, S. M., Makowski, L., Randell, S. H., Chiang, D. Y., Hayes, D. N., Jones, C., Liu, Y., Prins, J. F., and Liu, J. (2013). DiffSplice: The genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Research*, **41**(2), 1–18.

Humphrey, J., Emmett, W., Fratta, P., Isaacs, A. M., and Plagnol, V. (2016). Quantitative analysis of cryptic splicing associated with TDP-43 depletion. *bioRxiv*, pages 1–21.

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), 1009–15.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome research*, **12**(6), 996–1006.

Lardelli, R. M., Thompson, J. X., Yates, J. R., and Stevens, S. W. (2010). Release of SF3 from the intron branchpoint activates the first step of pre-mRNA splicing. *RNA (New York, N.Y.)*, **16**(3), 516–528.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, **11**(10), 733–739.

Ling, J. P., Pletnikova, O., Troncoso, J. C., and Wong, P. C. (2015). TDP-43 repression of nonconserved cryptic exons is compromised in ALS-FTD. *Science*, **349**(6248), 650–655.

Matter, N., Herrlich, P., and König, H. (2002). Signal-dependent regulation of splicing via phosphorylation of Sam68. *Nature*, **420**(6916), 691–695.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, **5**(7), 621–628.

Park, Y. and Wu, H. (2016). Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics*, **32**(10), 1446–1453.

Press, W. H. (2007). *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press.

Rakesh, R., Joseph, A. P., Bhaskara, R. M., and Srinivasan, N. (2016). Structural and mechanistic insights into human splicing factor SF3b complex derived using an integrated approach guided by the cryo-EM density maps. *RNA Biology*, **0**(0), 1–16.

Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., Na, H., Irimia, M., Matzat, L. H., Dale, R. K., Smith, S. a., Yarosh, C. a., Kelly, S. M., Nabet, B., Mecenas, D., Li, W., Laishram, R. S., Qiao, M., Lipshitz, H. D., Piano, F., Corbett, A. H., Carstens, R. P., Frey, B. J., Anderson, R. a., Lynch, K. W., Penalva, L. O. F., Lei, E. P., Fraser, A. G., Blencowe, B. J., Morris, Q. D., and Hughes, T. R. (2013). A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, **499**(7457), 172–7.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, **26**(1), 139–40.

Sharma, S., Kohlstaedt, L. A., Damianov, A., Rio, D. C., and Black, D. L. (2008). Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome. *Nature structural & molecular biology*, **15**(2), 183–91.

Shen, S., Park, J. W., Huang, J., Dittmar, K. A., Lu, Z. X., Zhou, Q., Carstens, R. P., and Xing, Y. (2012). MATS: A Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Research*, **40**(8), 1–13.

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 of the United States of America*, **111**(51), E5593–601.

Shi, Y. and Jiang, H. (2013). rSeqDiff: Detecting differential isoform expression from RNA-Seq data using hierarchical likelihood ratio test. *PLoS ONE*, **8**(11).

Singh, D., Orellana, C. F., Hu, Y., Jones, C. D., Liu, Y., Chiang, D. Y., Liu, J., and Prins, J. F. (2011). FDM: A graph-based statistical method to detect differential transcription using RNA-seq data. *Bioinformatics*, **27**(19), 2633–2640.

Smithson, M. and Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, **11**(1), 54.

Tisserant, A. and Konig, H. (2008). Signal-Regulated Pre-mRNA Occupancy by the General Splicing Factor U2AF. *PloS one*, **101**(1), e1418.

Vaquero-Garcia, J., Barrera, A., Gazzara, M. R., González-Vallinas, J., Lahens, N. F., Hogenesch, J. B., Lynch, K. W., and Barash, Y. (2016). A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife*, **5**, 1–30.

Williams, R., Baccarella, A., Parrish, J. Z., Kim, C. C., Francisco, S., and Francisco, S. S. (2017). Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. *BMC Bioinformatics*, pages 1–29.

Wu, J., Akerman, M., Sun, S., McCombie, W. R., Krainer, A. R., and Zhang, M. Q. (2011). Splice Trap: A method to quantify alternative splicing under single cellular conditions. *Bioinformatics*, **27**(21), 3010–3016.

Yang, E.-W. W. and Jiang, T. (2016). SDEAP: a splice graph based differential transcript expression analysis tool for population data. *Bioinformatics*, **32**(23), 3593–3602.

Zhang, R., Calixto, C. P. G., Marquez, Y., Venhuizen, P., Tzioutziou, N. A., Guo, W., Spensley, M., Frei dit Frey, N., Hirt, H., James, A. B., Nimmo, H. G., Barta, A., Kalyna, M., and Brown, J. W. S. (2016). AtRTD2: A Reference Transcript Dataset for accurate quantification of alternative splicing and expression changes in Arabidopsis thaliana RNA-seq data. *bioRxiv*, **0**(June), 051938.