Modeling condition-specific alternative splicing

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

* martin.strazar@fri.uni-lj.si

[1] Bioinformatics Laboratory, University of Ljubljana
Faculty of Computer and Information Science, Večna pot 113, 1000 Ljubljana
[2] The Francis Crick Institute, 1 Midland Rd, Kings Cross, London NW1 1AT, UK

Motivation

Contemporary differential exon usage statistical tests compare multiple experimental conditions to a single reference condition. The emergence of datasets including hundreds of experimental conditions calls for tailored models to detect condition-specific changes in splicing and uncover RNA binding protein-specific regulation [1].

Summary of results

We design a novel statistical model, named Condition-specific differential exon expression (csDEX), to discover changes in exon usage that occur only in a small subset of conditions. The package supports both read count- and Percent spliced-in (PSI)-based exon expression quantification. We test for alternative splicing (AS) changes on a public dataset with 189 shRNA knockdown samples of different RNA binding proteins (RBPs; including SRSF1, U2AF1/2, PTBP1, hnRNPs, TARDBP) provided by the ENCODE project [2]. We demonstrate the advantages of PSI-based quantification when seeking changes in exon usage due to AS rather than gene expression. The causal effect of RBP binding on AS is further validated by multiple independent data sources, such as RBP binding assays (eCLIP) and motif analysis, as well as successfully retrieving cryptic exons known to be TARDBP-regulated [3].

Condition-specific differential exon expression (csDEX)

csDEX-count

The read count Y_{ec} mapping to exon e upon condition c is distributed according to a negative binomial (NB) distribution:

$$Y_{ec} \sim NB(\mu_{ec}, d_e)$$

parametrized by
- mean count μ_{ec}
- dispersion d_e

csDEX-PSI

The Percent-spliced in (PSI) ψ_{ec} of e upon condition c within correspoding transcripts is distributed according to a Beta distribution:

$$\psi_{ec} \sim Beta(\mu_{ec}, \phi)$$

parametrized by
- mean percentage μ_{ec}
- precision ϕ

Parametrize μ_{ec} using a generalized linear model (GLM) with

- exon-specific factor β_e
- condition-specific factor β_c
- exon-condition interaction factor β_{ec}

For each gene, fit the GLM in order to perform analysis of variance (ANOVA), comparing the

a) null model - no interaction:

$$link(\mu_{ec}) = \beta_e + \beta_c$$

b) alternative model - interaction between exon(s) and condition(s):

$$link(\mu_{ec}) = \beta_e + \beta_c + \beta_{ec}\, \partial_{ee'}\, \partial_{cc'}$$

∂_{ij} Kronecker delta function

The link(μ) function:

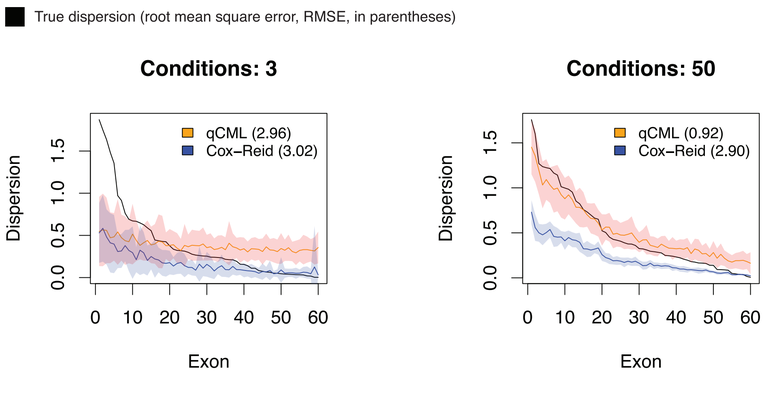
log(μ) **(csDEX-count)**

logit(μ) = μ / (1-μ) **(csDEX-PSI)**

Pairs of candidate interactions between exon e' and c' are compared using the likelihood-ratio test, resulting in a p-value. The final result is a ranked list of candidate interacting exons and conditions.

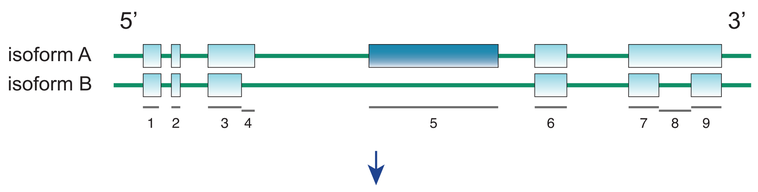
Large number of conditions improves hyperparameter estimation

In case of a large number of samples, dispersion can be estimated using maximum likelihood (ML). Due to unequal library sizes across conditions, we use a quantile-adjusted conditional ML (qCML) to generate identically distributed pseudodata and derive a common estimate [4]. Results on simulated data confirm qCML is the least biased in large sample cases and outperforms small sample-based methods such as the Cox-Reid dispersion estimate.

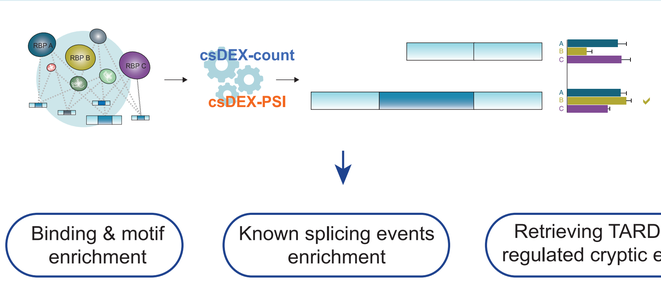


Experimental setup

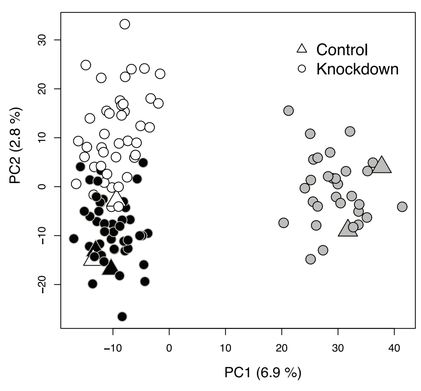
Select human genes containing at least 1 alternative (cassette) exon and between 5 and 15 exonic parts (annotation **hg19/5-15**) or between 13 and 66 exonic parts (**hg19/13-66**).



Differential exon usage analysis, integrating ENCODE RNA-seq data and 189 RBP knockdowns. Identify RBP-specific changes in exonic part usage.



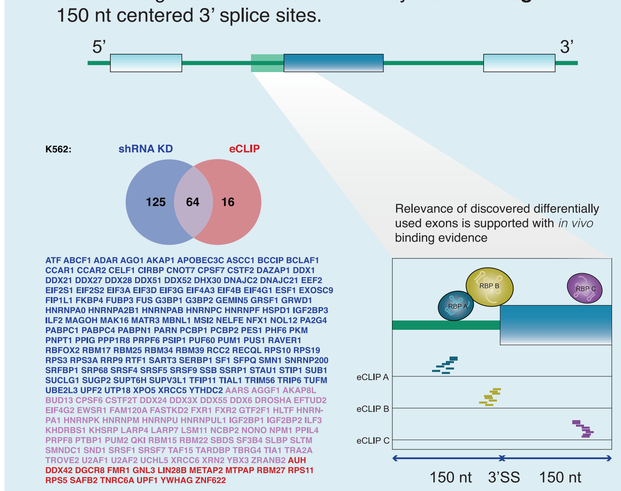
Non-negligible amount of variance in the shRNA+RNASeq dataset due to batch effects



Principal component analysis (PCA) of RNA-seq samples, where the PSI values are mapped to the hg19/13-66 annotation. Three experimental batches with the most samples are shown for clarity: black batch (48 samples, dated 16. 10. 2014), white (44, 17. 12. 2014), gray (32, 16. 3. 2016). PCA of the whole dataset reveals higher explained variance (PC1: 10.7 %, PCA2: 5.9 %) and average within-batch distance (60.9 ± 43.1) significantly lower than between-batch distance (198.8 ± 108.3).

RBP binding and motif enrichment increases with complex design

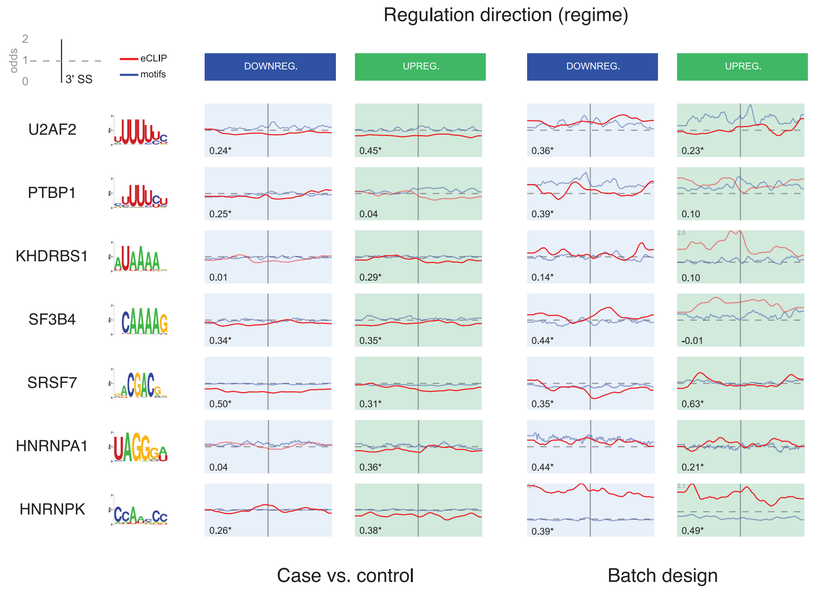
RBP binding and motif enrichment analyses within **regions** of 150 nt centered 3' splice sites.



Two modeling designs are compared: **case vs. control** (one diff. analysis per each knockdown) and **batch design** (joint diff. analysis for all RBPs in a batch).

Regulated alternative exonic parts (hg19/13-66, FDR<10%) are used as foreground and 20,000 non-significantly regulated as a background set. 7 shown RBPs display binding enrichment of at least 1.5-fold in at least one regime. No comparable enrichment is observed when using case vs. control design.

Inset values display the values of cross-correlation (CC) between eCLIP and motif signal. Highlighted are cases where CC > 0.15.



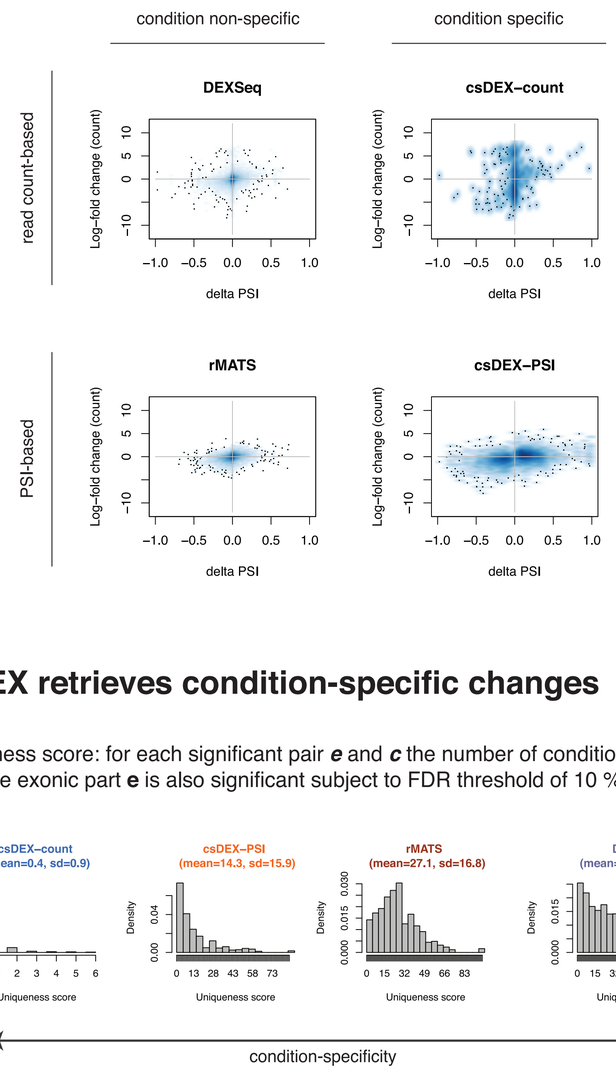

```
$ R
> require(devtools)
> install_github("mstrazar/csDEX")
```

github.com/mstrazar/csDEX

Perceived change in PSI implies change in read counts, but not *vice versa*

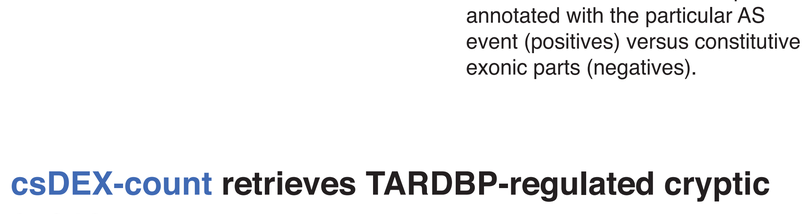
Comparing for each diff. used exonic part e in condition c:
- $\Delta\Psi$; difference between Ψ_{ec} and actual average $\Psi_{ec'}$ over all c' (x-axis)
- ΔY; $\log_2$-fold difference between read count Y_{ec} and average $Y_{ec'}$ over c' (y-axis)

Exonic parts retrieved by PSI-based models show strongest agreement (rMATS, Pearson corr. = 0.31; csDEX-PSI, corr. = 0.34).



csDEX retrieves condition-specific changes

Uniqueness score: for each significant pair e and c the number of conditions c' where the same exonic part e is also significant subject to FDR threshold of 10 %.



csDEX-PSI retrieves known splicing events with highest precision

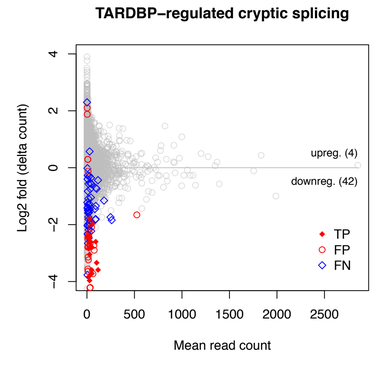
	csDEX-PSI	csDEX-PSI*	rMATS	csDEX-count	csDEX-count*	DEXSeq
alternative	**.98**	**.98**	.915	.806	.817	.807
altFinish	**.14**	.13	.004	.024	.023	.019
altFivePrime	.49	**.53**	.217	.084	.094	.056
altPromoter	.86	**.88**	.122	.368	.453	.285
altThreePrime	.47	**.51**	.256	.086	.203	.047
bleedingExon	.77	**.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

UCSC knownAlt annotation is used as ground truth for validation. For each method, we select the top 10,000 most significant interactions. For each of the nine AS event types, we compute the **cumulative precision** for each possible significance cut-off.

Precision: number of exonic parts annotated with the particular AS event (positives) versus constitutive exonic parts (negatives).

csDEX-count retrieves TARDBP-regulated cryptic exons

Previously unnanotated splice site pairings - **cryptic exons** - can emerge upon TARDBP knockdown. We test for TARDBP-specific changes with csDEX-count (as cryptic exons are not part of the annotation) [5]. Known cryptic exons are heavily enriched in the 46 significantly diff. used exons (FDR<5% shown in red; p-value<1E-31, hypergeometric test).



diff. used	cryptic	non-cryptic	total
YES	20	26	46
NO	51	11319	11370
total	71	11345	11416

References

[1] Fu, X.D. et al (2014). Context-dependent control of alternative splicing by RNA-binding proteins. Nat. Rev. Genet., 15(August), 689–701. **[2]** The Encode Consortium (2011), A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 9, e1001046. **[3]** Ling, J.P. et al (2015). TDP-43 repression of nonconserved cryptic exons is compromised in ALS-FTD. Science, 349(6248), 650–655. **[4]** Robinson, M.D. et al (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1), 139–40. **[5]** Humphrey, J. et al (2017). Quantitative analysis of cryptic splicing associated with TDP-43 depletion. BMC Medical Genomics, 10(1), 38.