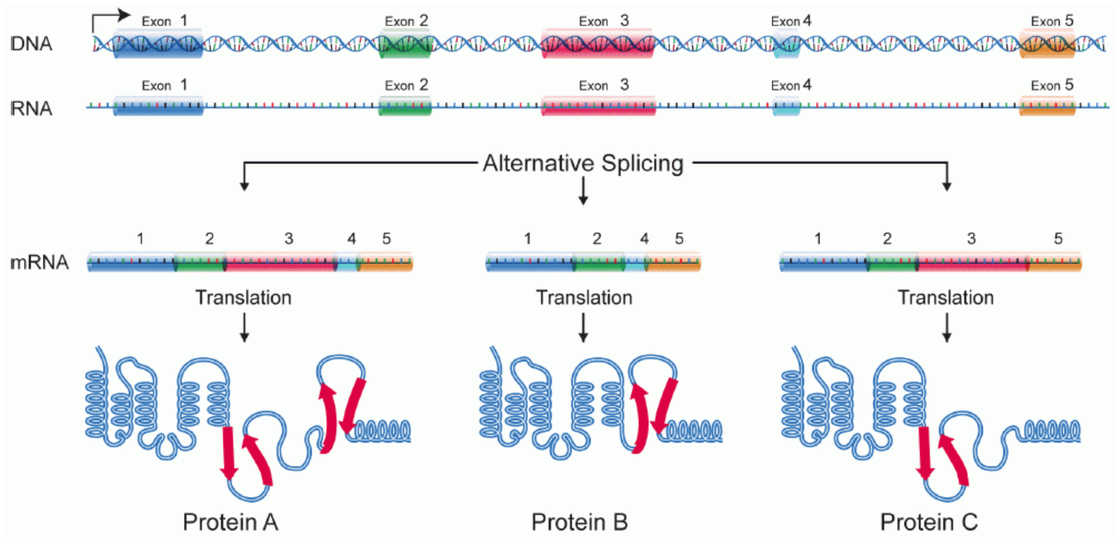
Improving power to detect differential exon usage by L1-regularization (lasso) model selection

Lukas M. Weber, Charlotte Soneson, Mark D. Robinson
Institute of Molecular Life Sciences, University of Zurich
lukas.weber@imls.uzh.ch

1. Background

Alternative splicing: during gene expression, coding regions of the genome (exons) can be spliced together into different messenger RNA (mRNA) sequences (transcripts), resulting in varying isoforms of the final protein.



Wikipedia: https://en.wikipedia.org/wiki/Alternative_splicing

Differential analysis compares gene expression (mRNA abundance) between two or more conditions, for example cancer versus healthy.

- **Differential gene expression**: differences in total expression level of all mRNA transcripts from a gene.
- **Differential transcript expression** (DTE): differences in expression level of individual transcripts.
- **Differential transcript usage** (DTU): differences in proportional expression of the set of transcripts from a gene, which can occur as a result of **differential splicing**.
- **Differential exon usage** (DEU): surrogate for DTU used for quantification.

2. Statistical methods for DEU

DEXSeq [1] is a popular R/Bioconductor package used to perform statistical tests for DEU for RNA-seq data.

- DEXSeq methods begin with a table of read counts for each exon in each sample.
- Exon-level tests are summarized into gene-level q-values to rank all genes in the data set by evidence for differential splicing.

Example read count table

		Condition 1		Condition 2	
Gene	Exon	Sample 1	Sample 2	Sample 3	Sample 4
1	1	300	310	150	150
	2	400	410	195	210
	3	100	100	55	50
2	1	210	200	100	100
	2	110	100	55	50
	3	40	35	40	40
	4	150	140	140	150

voom-diffSplice [2] can be used for microarrays and RNA-seq data.

- Microarray data are continuous intensity values, while RNA-seq data are discrete read counts.

3. Model selection approach

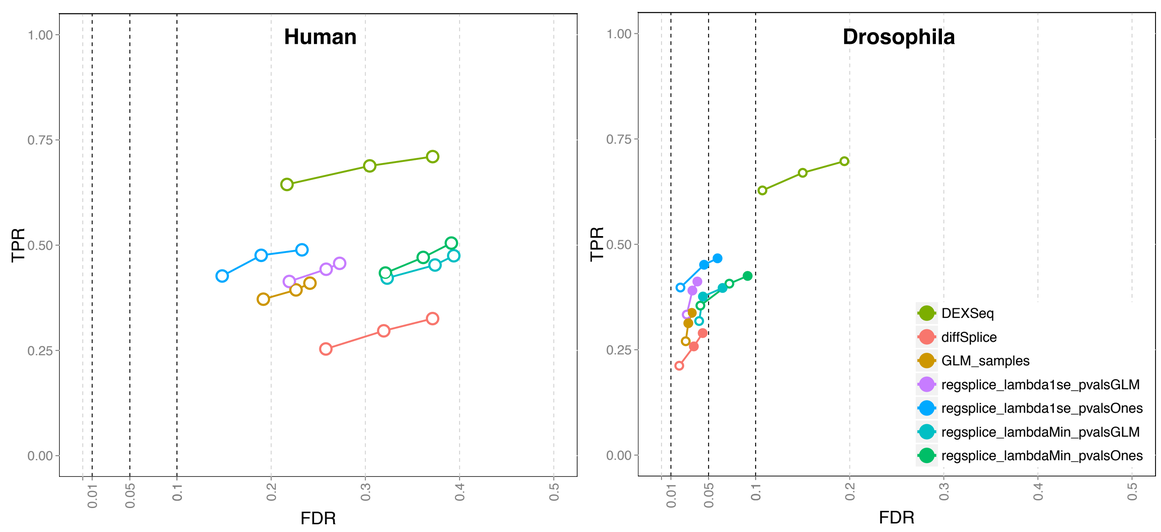
- In the voom-diffSplice framework, linear models with interaction terms for every exon are used to test each gene for differential splicing.
- However, biology suggests many interaction terms are redundant – differential splicing often involves only a few exons.
- We proposed using automated model selection techniques to select a subset of interaction terms for each gene. This reduces the complexity of the models and increases statistical power.
- The lasso (or L1-regularized regression) [3] is an efficient method to perform variable selection while fitting a linear model.

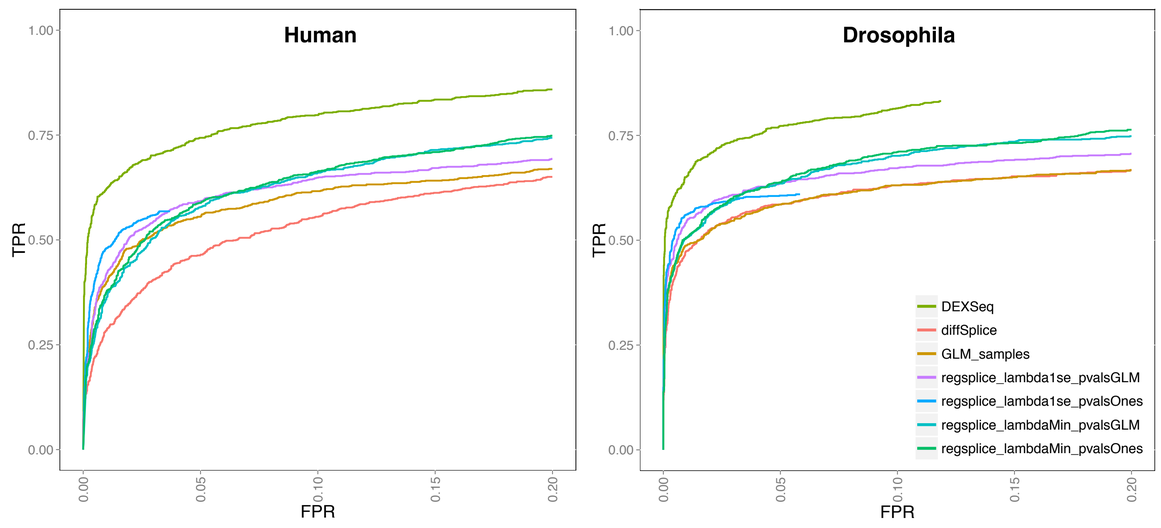
solve for $\boldsymbol{\beta}$ *that minimizes:*

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|$$

4. Results

- Simulated data sets: human, fruit fly (*drosophila melanogaster*).
- Model selection approach implemented as a new R package ("regsplice") built around core fitting functions from R package "glmnet" [4].
- Compare methods using true positive rate (TPR) vs. false discovery rate (FDR) plots and receiver operating characteristic (ROC) curves.





5. Discussion

Model selection approach improves performance of voom-diffSplice testing framework, but DEXSeq still performs best for RNA-seq data.

Inclusion of sample effect terms (e.g. GLM_samples) also improves performance.

Model selection approach:

- can be used with continuous data, e.g. microarrays or voom [5] transformed RNA-seq data (DEXSeq requires discrete counts, i.e. RNA-seq only)
- choice of method for genes where lasso selects zero interaction terms (p-value = 1, full GLM)
- fast computational speed (<10 min for human data set with 4 CPU cores; much faster than DEXSeq but slower than voom-diffSplice)

Next steps:

- Test on experimental microarray and RNA-seq data
- Bioconductor package (regsplice) and paper

References

1. Anders S., Reyes A., and Huber W. (2012). *Detecting differential usage of exons from RNA-seq data*. Genome Research, 22:2008. R package: DEXSeq, version 1.14.1.
2. Function "diffSplice" in R package: limma, version 3.24.10.
3. Tibshirani R. (1996). *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society Series B, 58(1), 267–288.
4. Friedman J., Hastie T., and Tibshirani R. (2010). *Regularization paths for generalized linear models via coordinate descent*. Journal of Statistical Software, 33(1), 1–22. R package: glmnet, version 1.9-8.
5. Law C.W., Chen Y., Shi W., and Smyth G.K. (2014). *Voom: precision weights unlock linear model analysis tools for RNA-seq read counts*. Genome Biology, 15, R29.