University of Zurich[UZH]

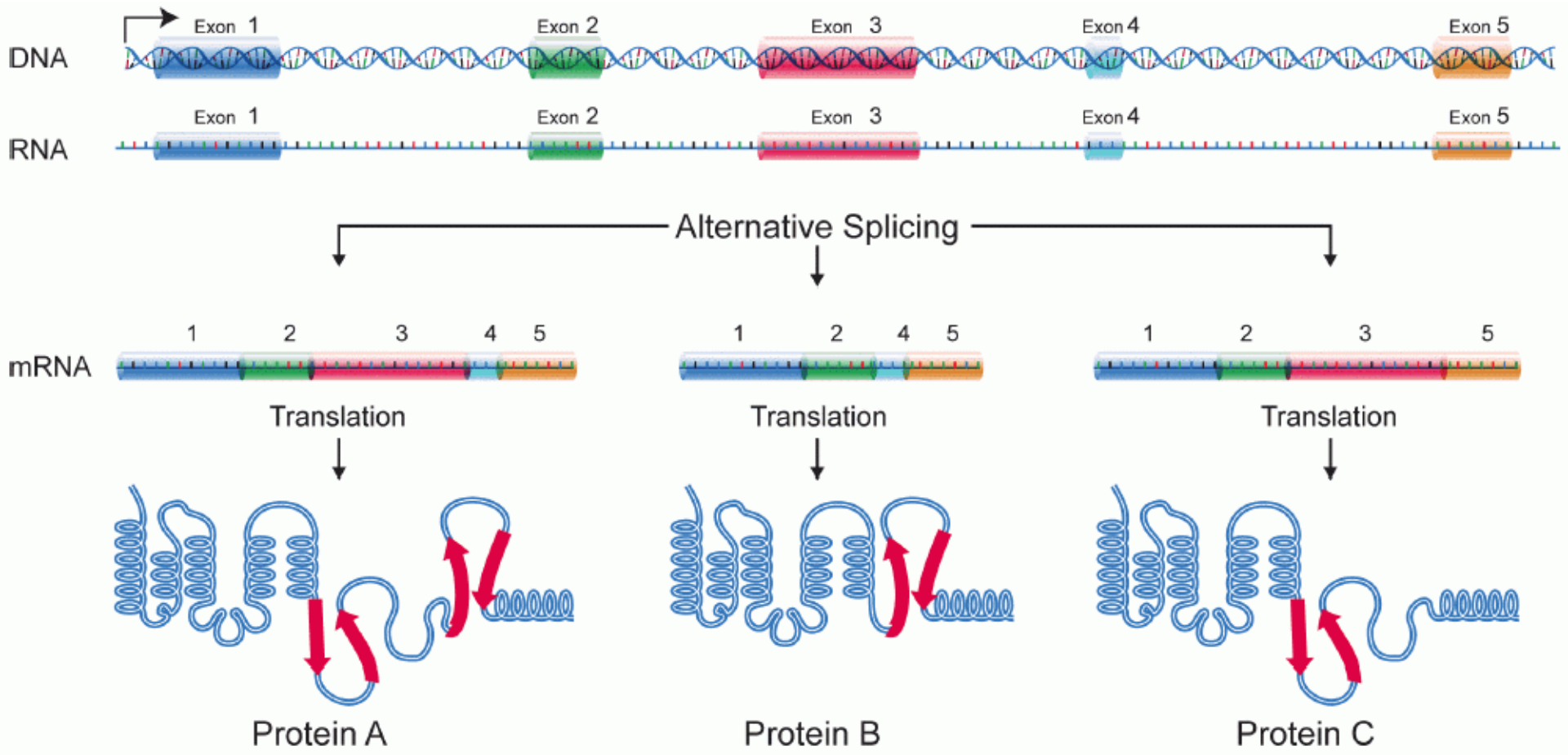
regsplice: Lasso-based model selection for improved detection of differential exon usage

European Bioconductor Developers' Meeting 2015
Cambridge, UK

Lukas Weber, University of Zurich

8 December 2015

Alternative splicing



Wikipedia

Differential analysis

Compare gene expression (mRNA abundance) between groups of samples in different conditions, e.g. 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 mRNA transcripts
- Differential transcript usage (DTU): differences in proportional expression of transcripts from a gene (e.g. due to differential splicing)
- Differential exon usage (DEU): surrogate for DTU used for quantification and testing

Current methods to test for DEU

DEXSeq:

- table of RNA-seq read counts at exon-level (counting bins)
- exon-level statistical tests
- gene-level q-values to rank genes 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:

- gene-level linear models with interaction terms for exons
- RNA-seq and microarrays

[other approaches]:

- DRIMSeq (Dirichlet-multinomial models): Gosia Nowicka
- transcript abundance, e.g. kallisto/sleuth
- comparison of approaches for counting reads: Charlotte Soneson (bioRxiv)

Model selection approach

- based on voom-diffSplice approach: gene-level linear models with interaction terms for exons
- biology suggests many interaction terms are redundant, since differential splicing may only involve some exons
- lasso-based model selection to select a subset of interaction terms for each gene

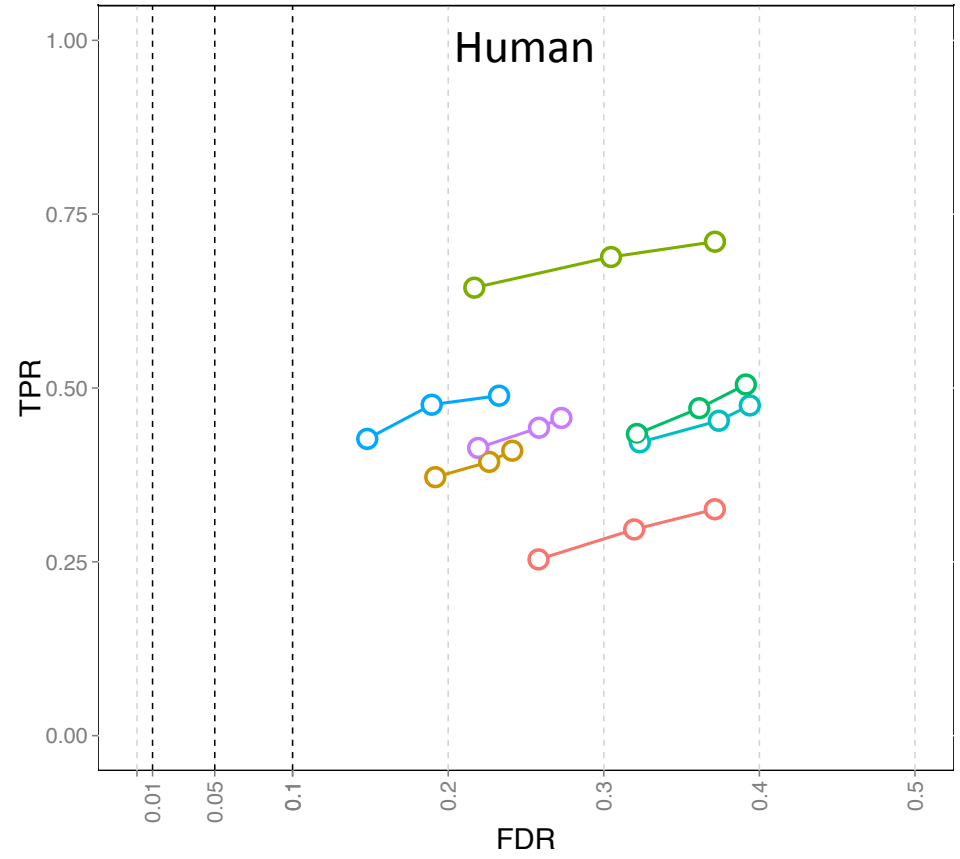
solve for $\boldsymbol{\beta}$ *that minimizes:*

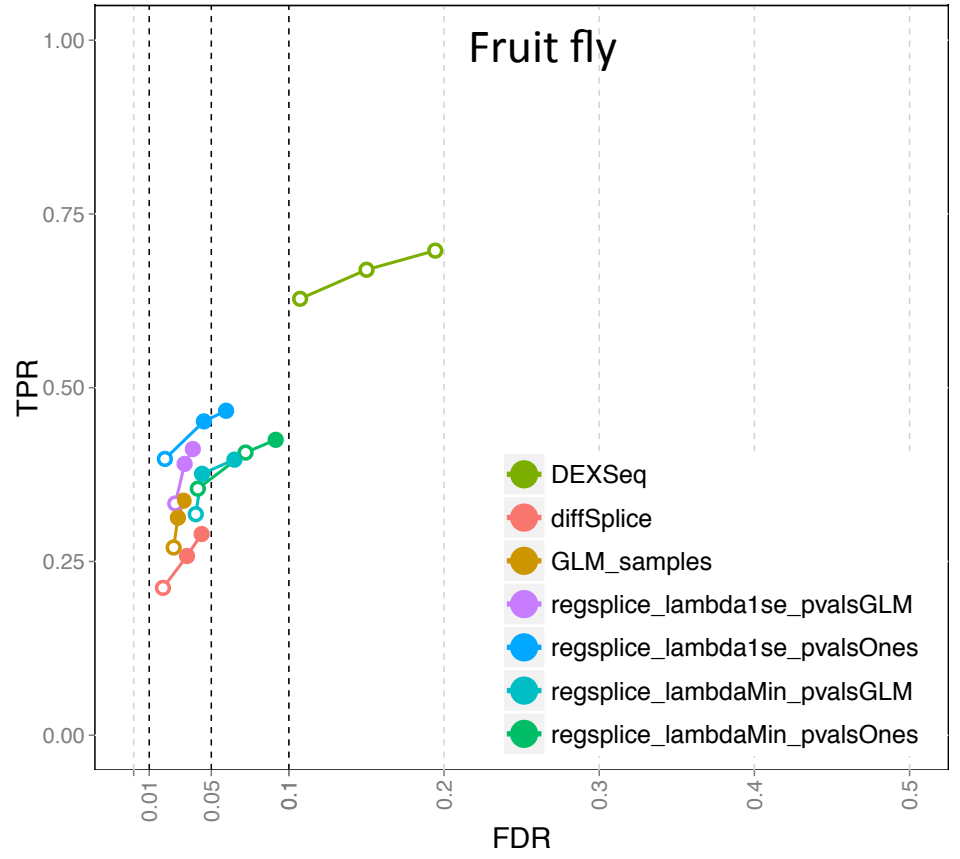
$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|$$

- reduces complexity of gene-level linear models, which may improve statistical power

Results

- methods implemented as R package regsplice, using lasso fitting functions from glmnet
- simulated data for human, fruit fly
- iCOBRA package for evaluation plots (Charlotte Soneson)





Discussion

- Lasso-based model selection approach improves performance of voom-diffSplice testing framework, but DEXSeq still performs best for RNA-seq data
  - Inclusion of terms for sample effects also improves performance
  - genes where lasso selects zero interaction terms: two options (p-value = 1, or fit full GLM)

- *regsplice* methods work with continuous data (microarrays or voom-transformed RNA-seq)

- fast computational speed
  - <10 min for human data set on a standard MacBook Air laptop
  - faster than DEXSeq, slower than voom-diffSplice

Next steps

- test using more data sets: RNA-seq, microarrays
- submit *regsplice* package to Bioconductor

Acknowledgments

- Mark Robinson
- Charlotte Soneson
- Robinson lab (UZH)

Additional slides

Additional results: ROC curves

- receiver operating characteristic curves (*iCOBRA* package)