Differential Alternative Splicing - an in silico detection approach

Gero Doose

Bled Presentation

Department of Computer Science and
Interdisciplinary Center for Bioinformatics,
University of Leipzig

February 19, 2016

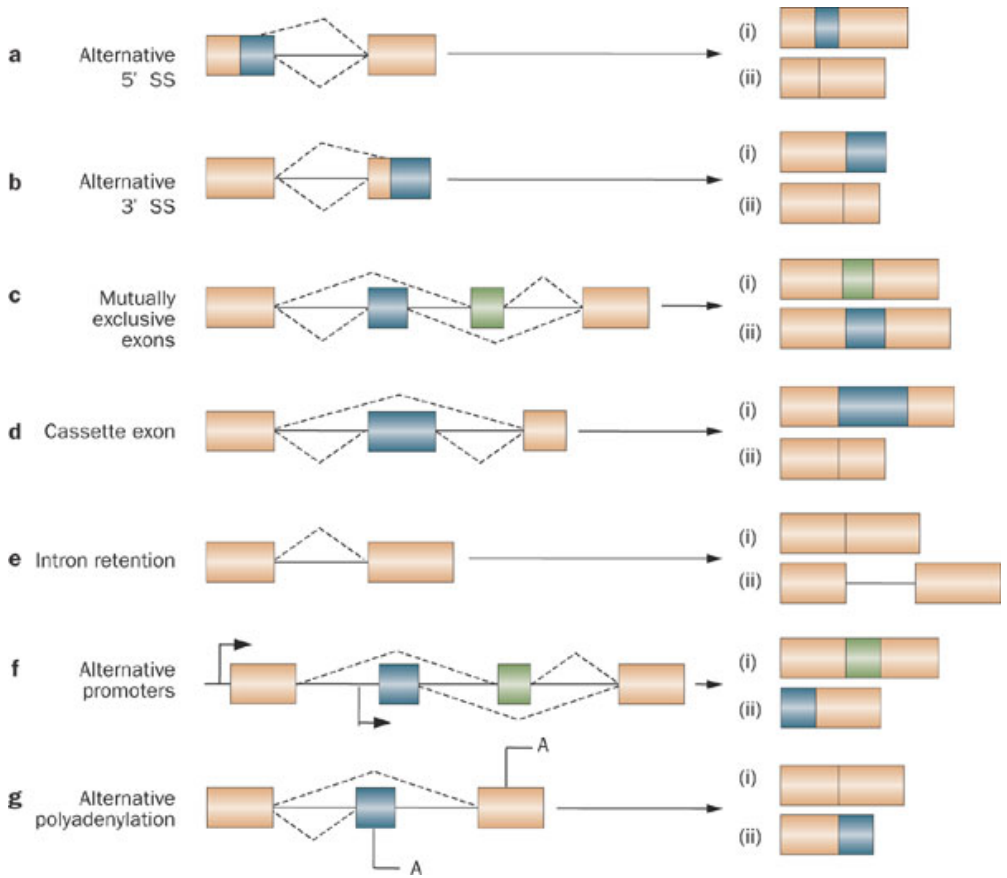
Table of Contents

Section 1

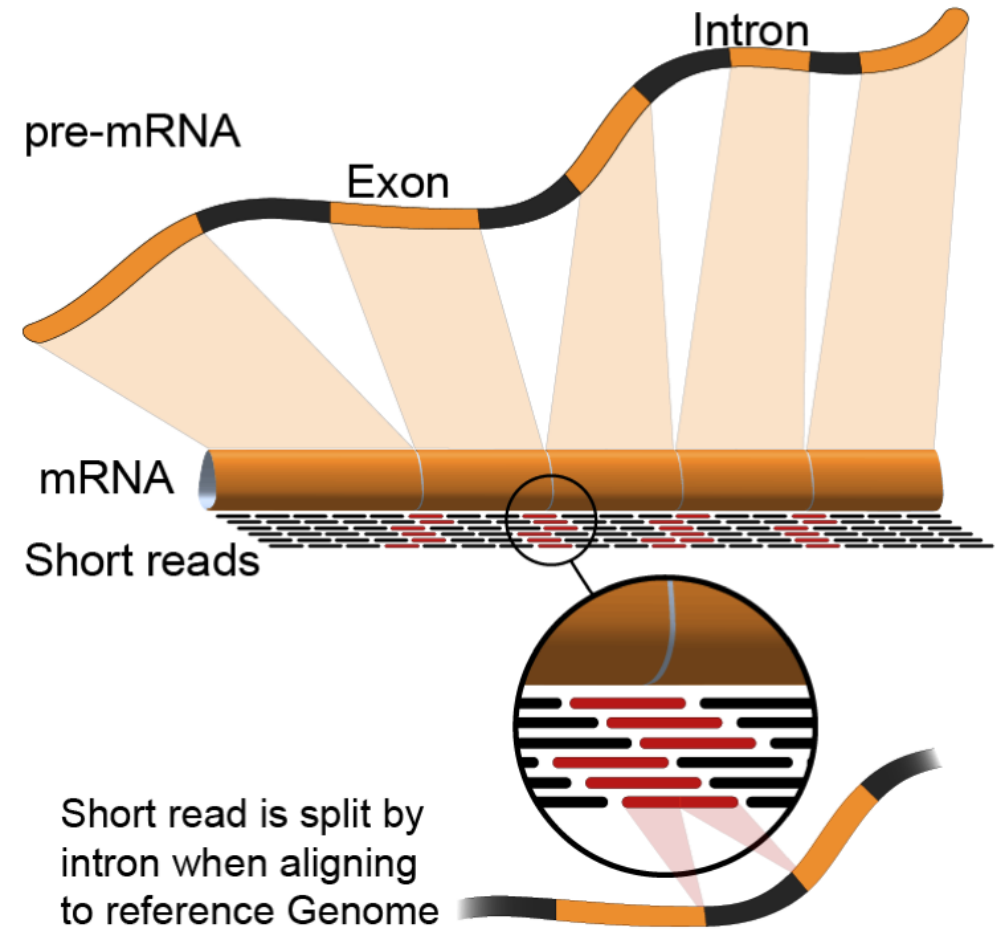
Motivation

Differential alternative splicing

Alternative splicing

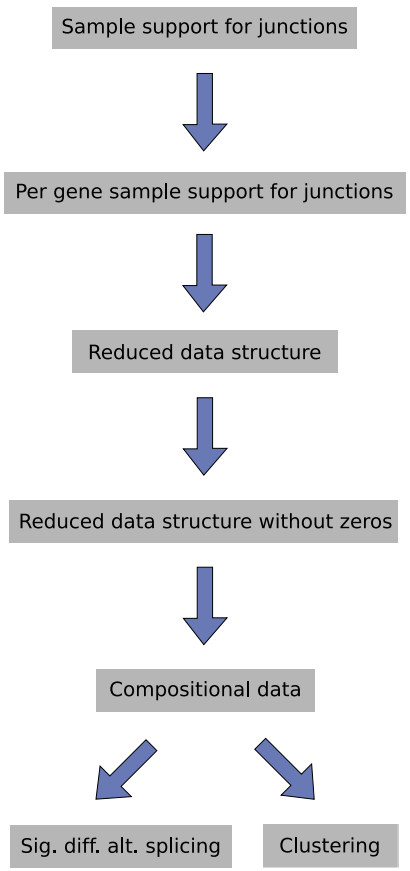


Alternative splicing



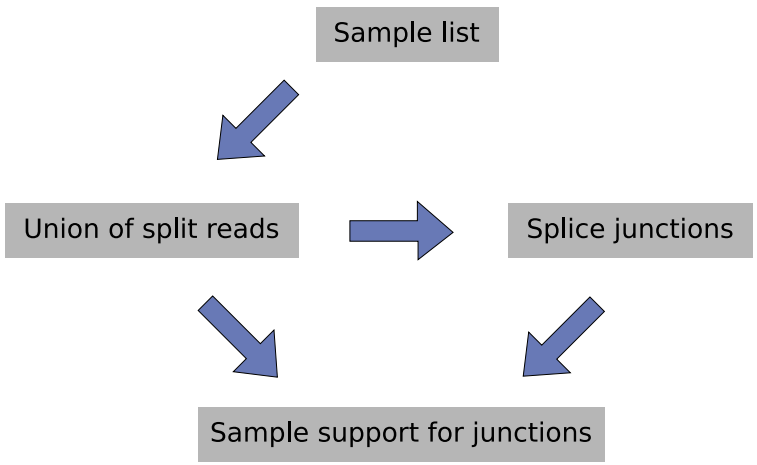
Section 2 **Method** Overview of the algorithm

Algorithmic workflow



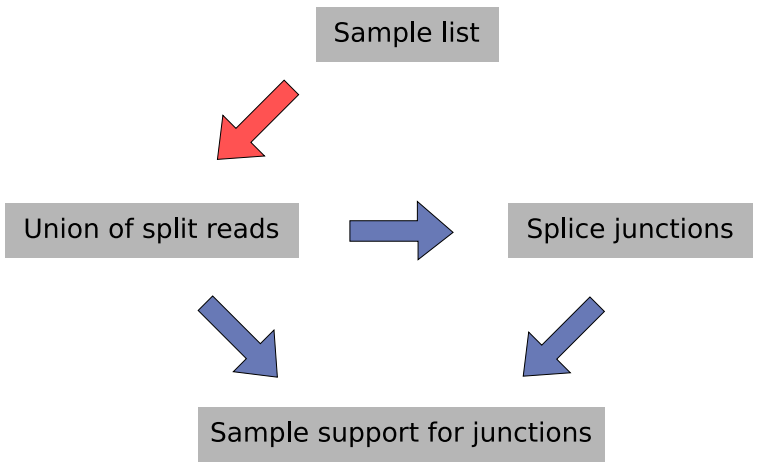
- Core algorithm is based on splice junction supports per sample
- Scanning routine for standard format (e.g. TCGA)
- Pre-calculation procedure for BAM files (segemehl)

Pre-calculation procedure for BAM files (segemehl)



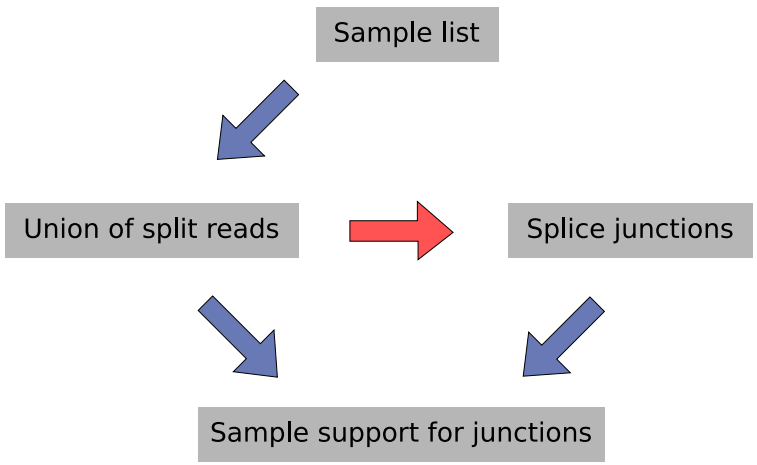
- Starting with a list of samples (location of the mapped RNA-seq files)

Pre-calculation procedure for BAM files (segemehl)



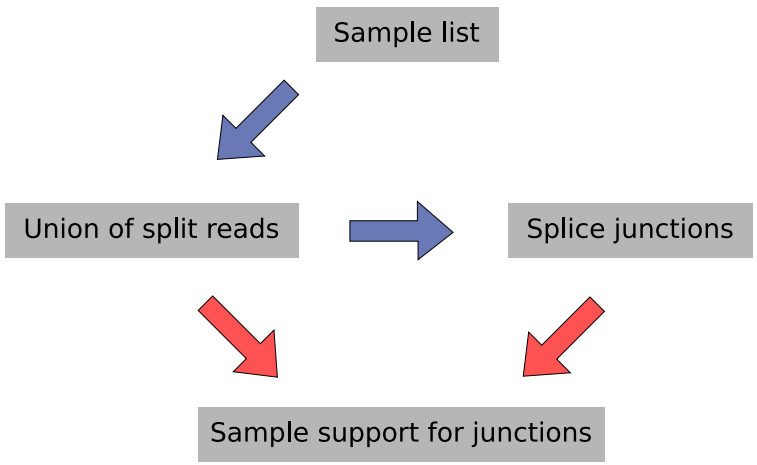
- Starting with a list of samples (location of the mapped RNA-seq files)
- Building the union of all split-mapped reads

Pre-calculation procedure for BAM files (segemehl)



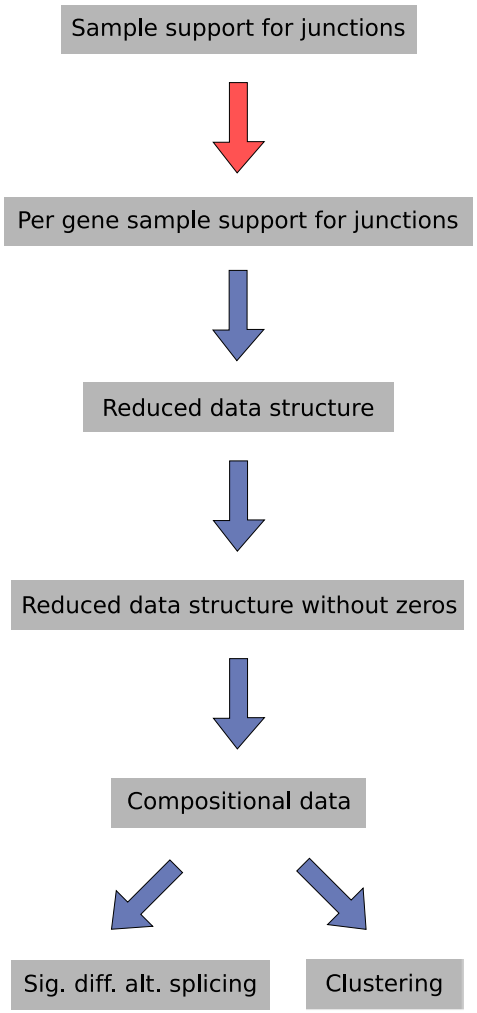
- Starting with a list of samples (location of the mapped RNA-seq files)
- Building the union of all split-mapped reads
- Calling splice junctions (cluster splice sites)

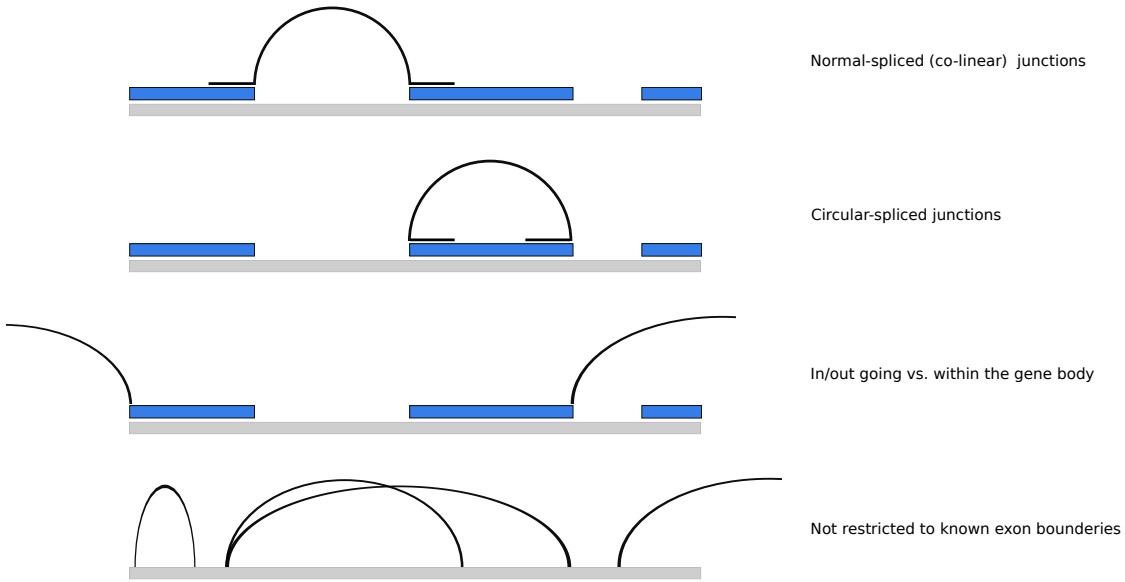
Pre-calculation procedure for BAM files (segemehl)

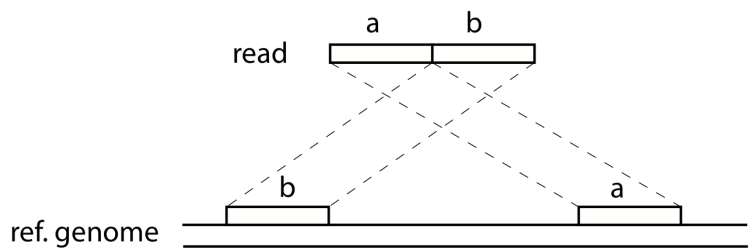


- Starting with a list of samples (location of the mapped RNA-seq files)
- Building the union of all split-mapped reads
- Calling splice junctions (cluster splice sites)
- Calculating table of individual sample supports

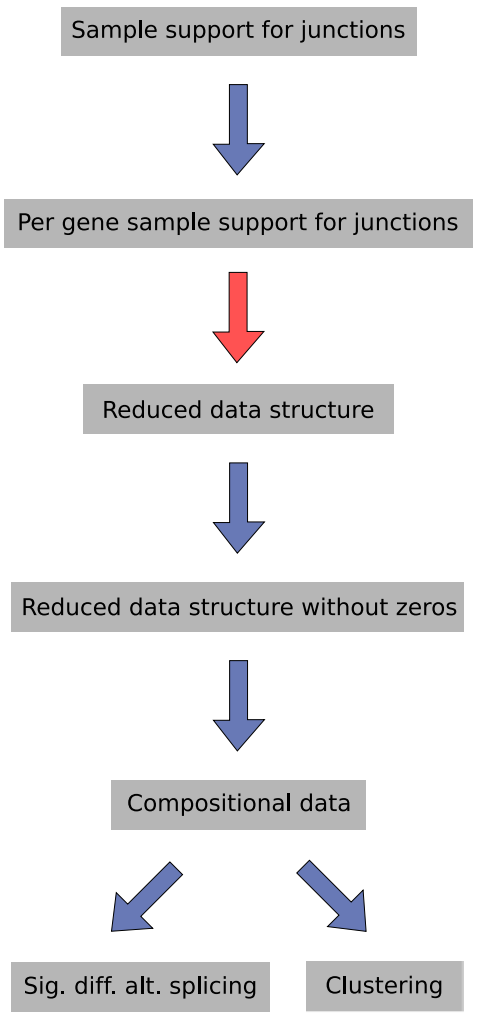
Intersecting with gene annotation





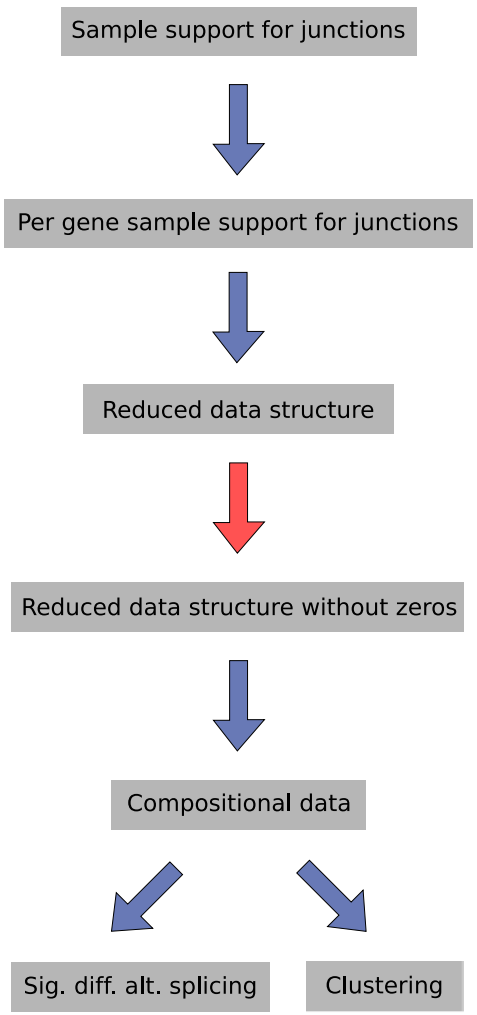


Matrix reduction

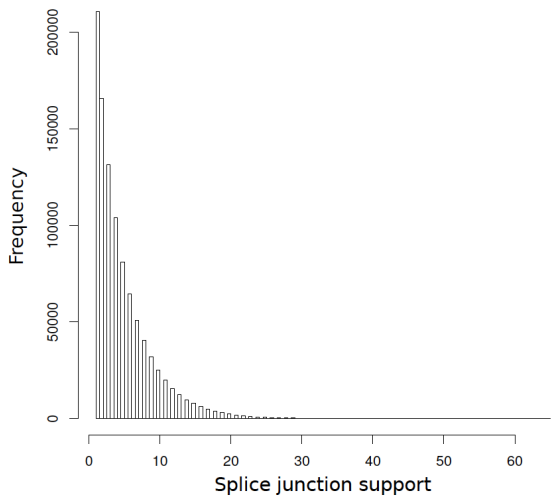


- minimum splice junction support value (J)
- minimum samples amount (S)
  - Removing ***samples*** not showing (J) in at least one splice junction
  - Removing ***junctions*** with less than (S) samples showing (J)
  - Removing ***genes*** not containing at least 2 junctions and (S) samples per condition

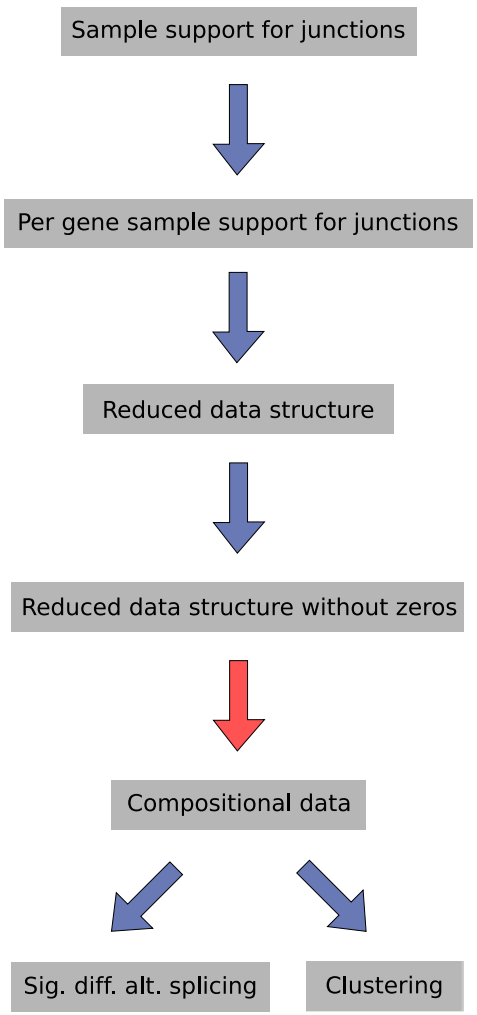
Zero-replacement



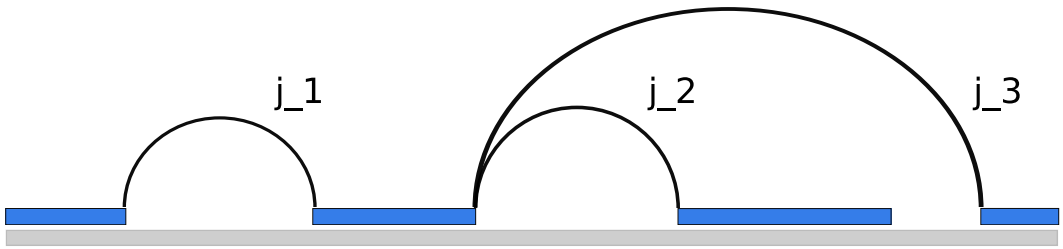
- Accounting for technical and biological variance
- Sampling junction supports from a negative binomial distribution
- Tracking junctions where more than 50% of a condition are replaced



Compositional data approach



- Normalizing raw counts to ratios per gene



$C_1 = [40, 30, 10]$ $\quad$ $C_1 = [0.5, 0.375, 0.125]$

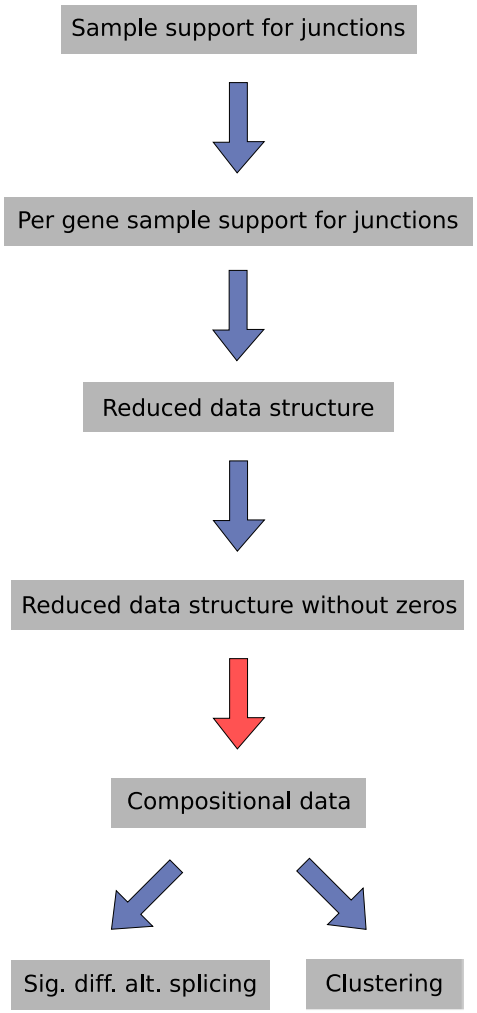
$C_2 = [120, 30, 90]$ $\quad$ $C_2 = [0.5, 0.125, 0.375]$

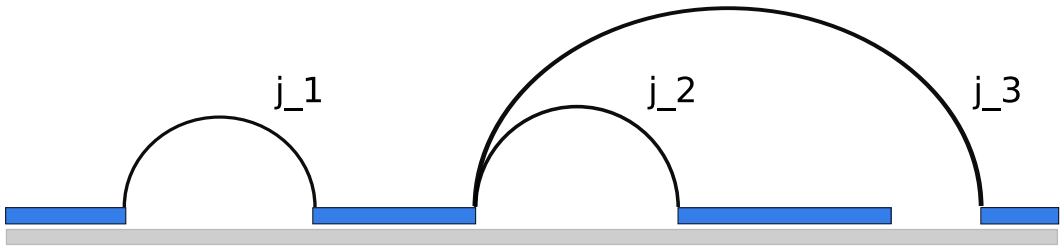
- Simplex as the appropriate sample space:

$$\mathcal{S}^D = \{[x_1, ..., x_D] : x_i \geq 0 \text{ for } i = 1, ..., D \text{ and } \sum_{i=1}^{D} x_i = 1\}$$

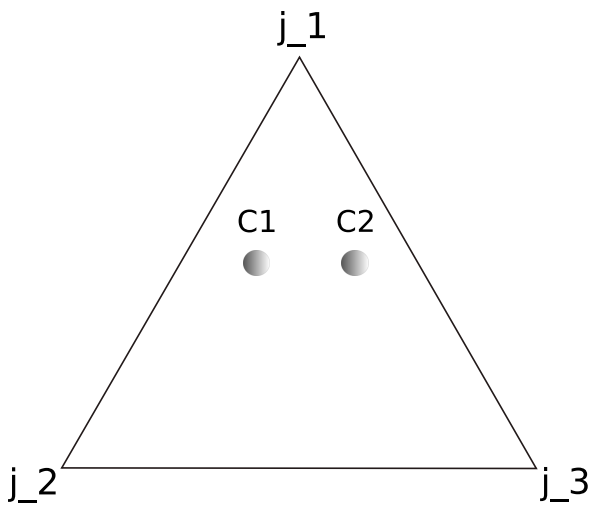
Aitchison (1986)

Compositional data approach



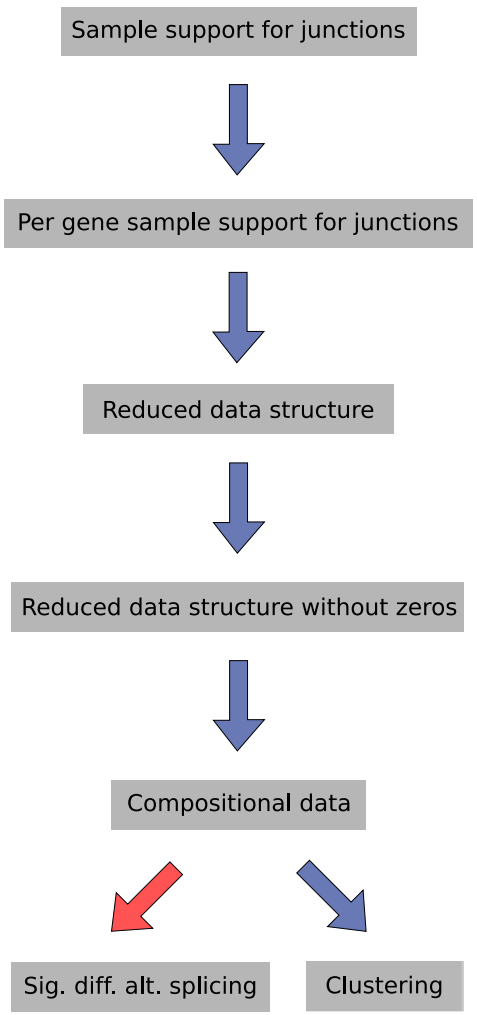


$$C_1 = [0.5, 0.375, 0.125]$$
$$C_2 = [0.5, 0.125, 0.375]$$



Ternary Diagram

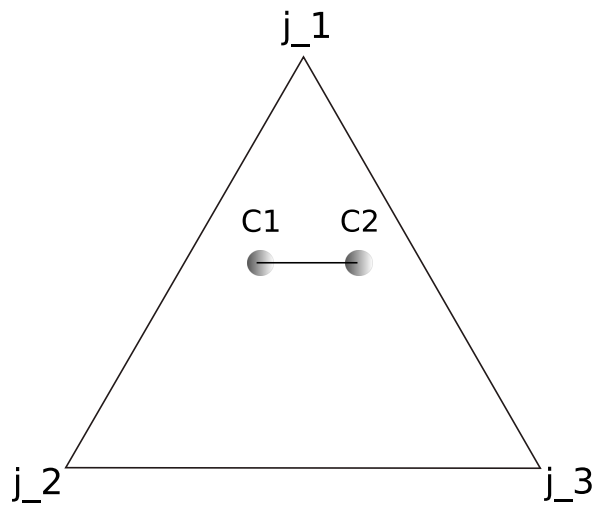
Compositional data approach



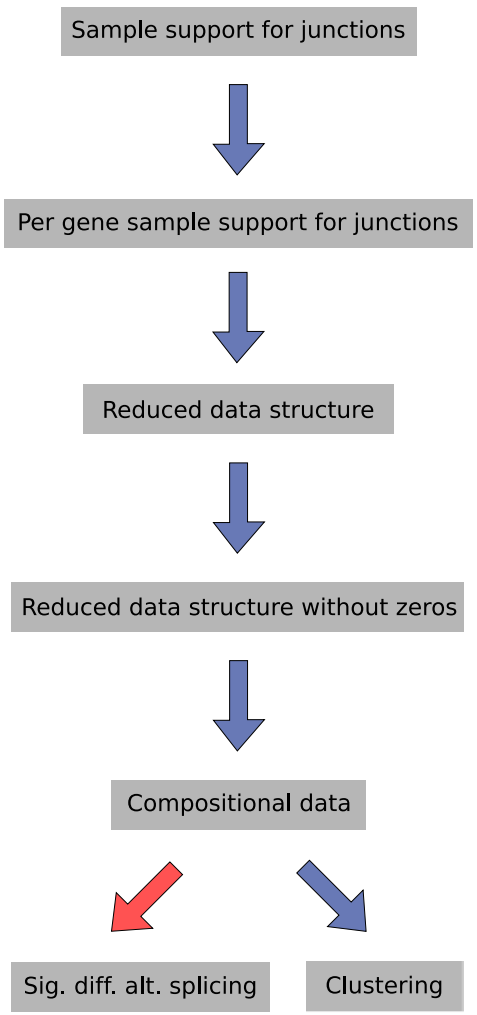
- Aitchison distance

$$d(x_a, x_b) = \left[\sum_{i=1}^{D} \left(\log\left(\frac{x_{ai}}{g(x_a)}\right) - \log\left(\frac{x_{bi}}{g(x_b)}\right) \right)^2 \right]^{\frac{1}{2}}$$

with $g(x) = \left(\prod_{i=1}^{D} x_i \right)^{\frac{1}{D}}$



Compositional data approach

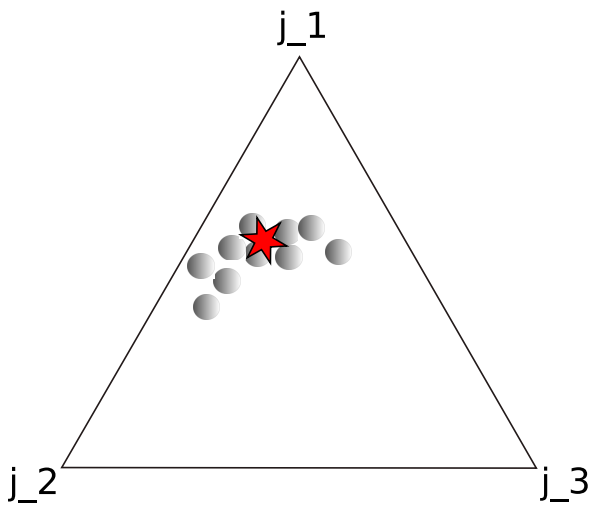


- Central tendency

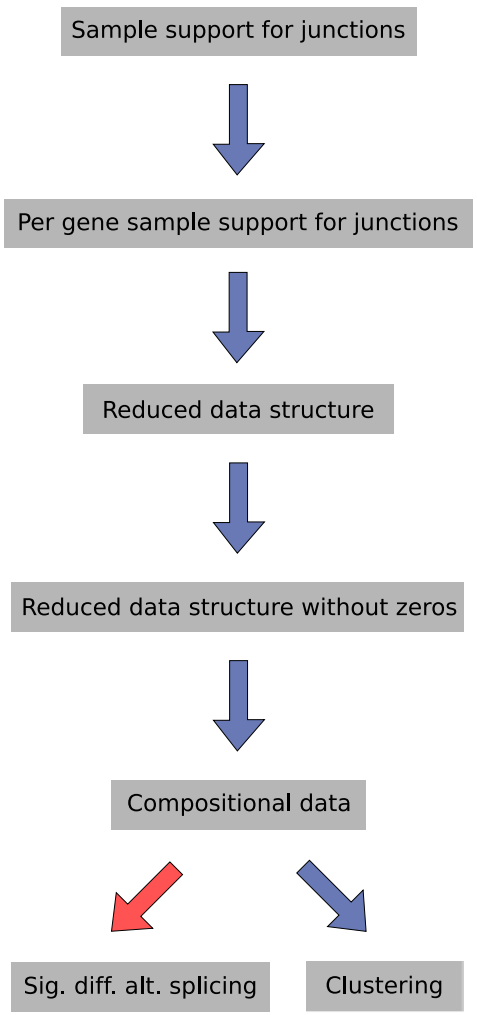
  Let $C = (x_{ij}, ..., x_{ND})$ be a set of N compositional vectors with D components:

$$cen(C) = \left[\frac{g(x_{i1})}{\sum_{j=1}^{D} g(x_{ij})}, ..., \frac{g(x_{iD})}{\sum_{j=1}^{D} g(x_{ij})} \right]$$

with $g(x_{ij}) = \left(\prod_{i=1}^{N} x_{ij} \right)^{\frac{1}{N}}$

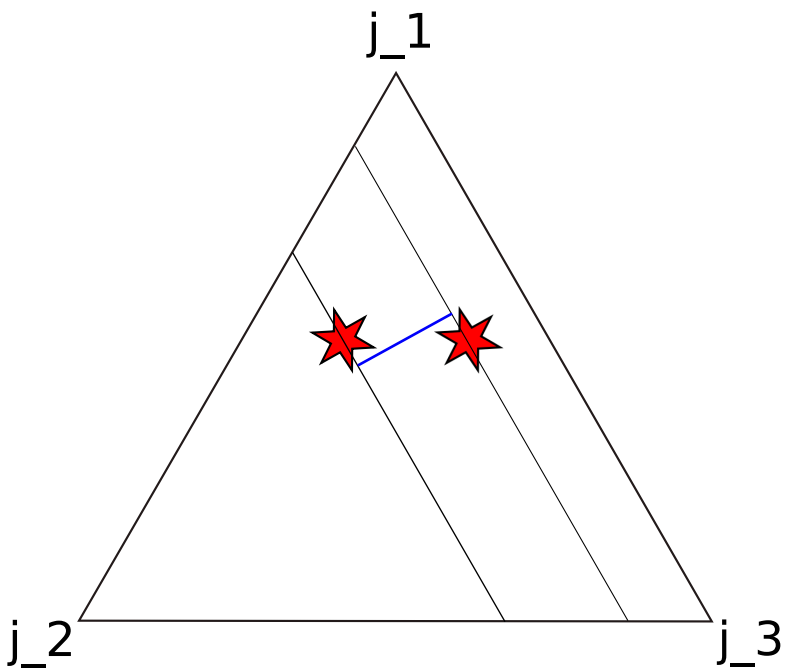


Compositional data approach



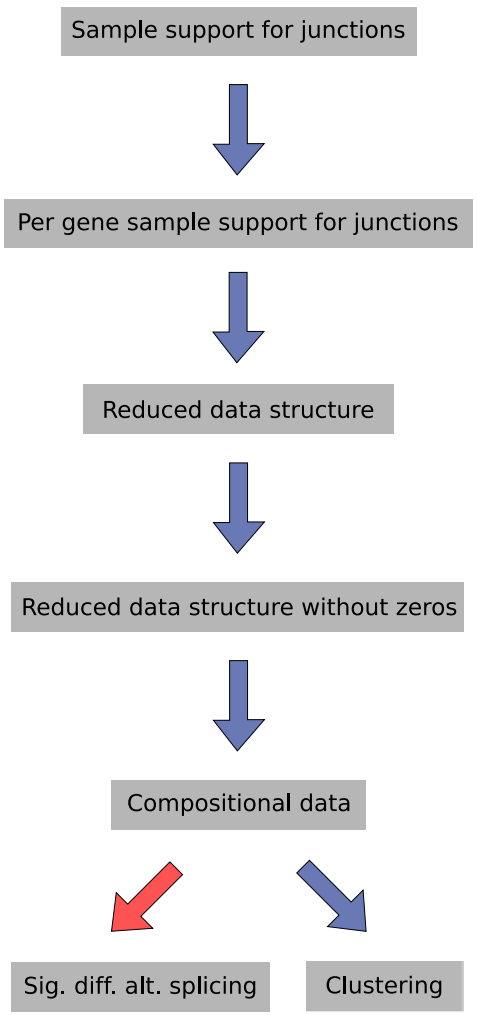
- Abundance change

$$abc(x_i) = d(cen(C_{base})_i, cen(C_{compare})_i)$$



$$C_1 = [0.5, 0.375, 0.125]$$
$$C_2 = [0.5, 0.125, 0.375]$$

Detection of differential alternative splicing



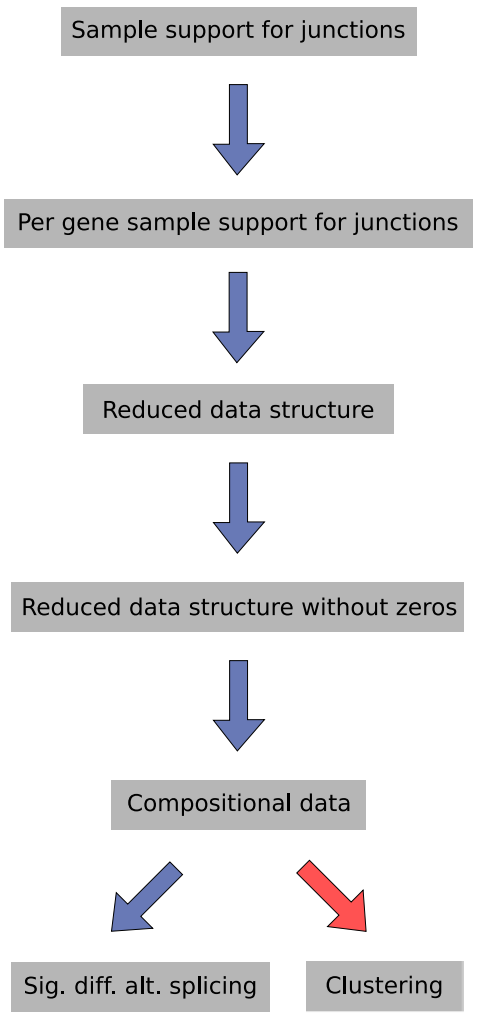
- For all components with $|abc| \geq 1$:
- Centered log-ratio transformation (clr):

$$clr(x) = \left[\log \frac{x_1}{g(x)}, ..., \log \frac{x_D}{g(x)}\right]$$

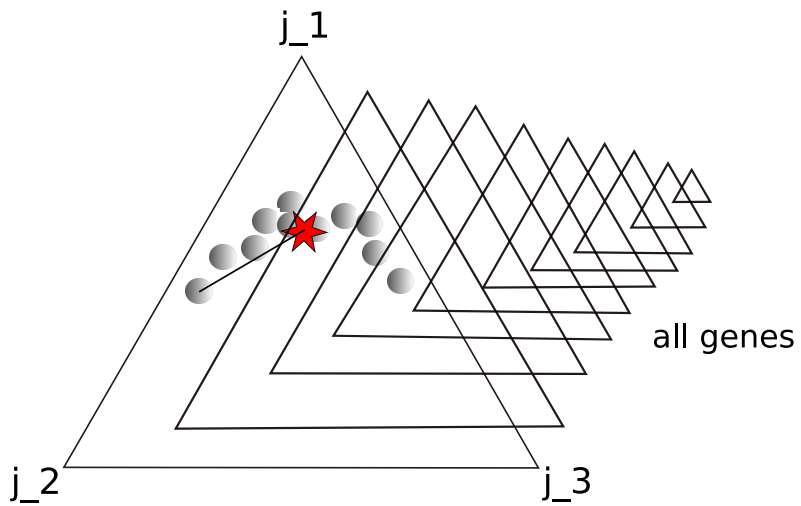
with $g(x) = \left(\prod_{i=1}^{D} x_i\right)^{\frac{1}{D}}$

- non parametric test statistic
  (Wilcoxon rank-sum)
- Multiple testing correction
  (Benjamini Hochberg)

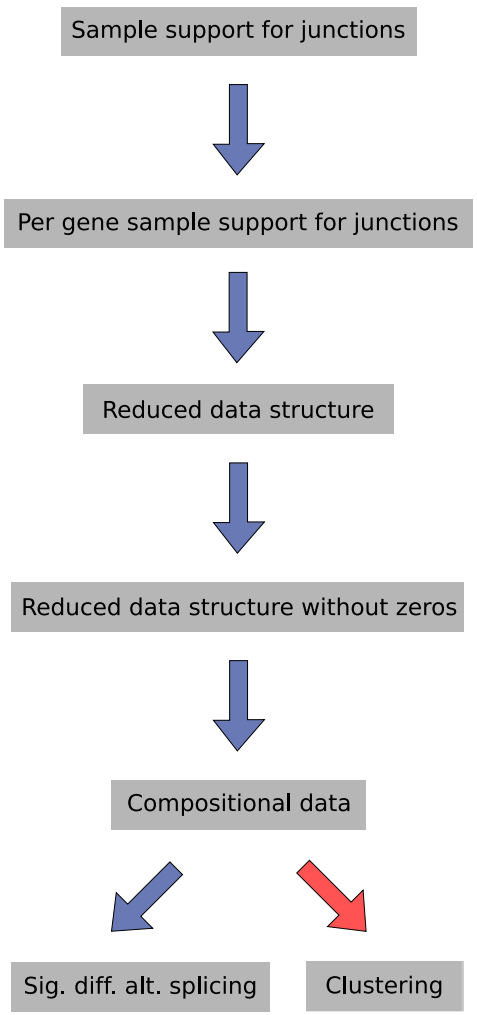
Clustering and outlier detection



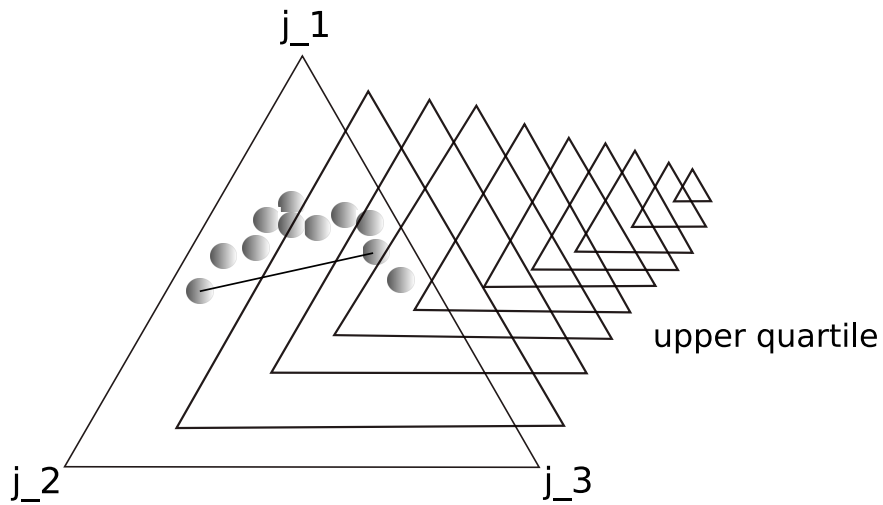
- Finding upper quartil of all genes in regard to average distance between the centre and each of the n compositional vectors



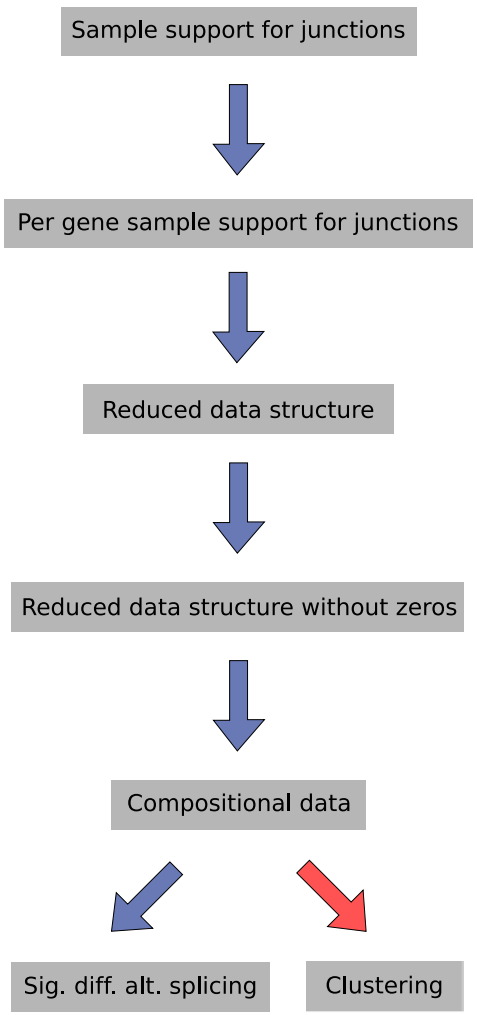
Clustering and outlier detection



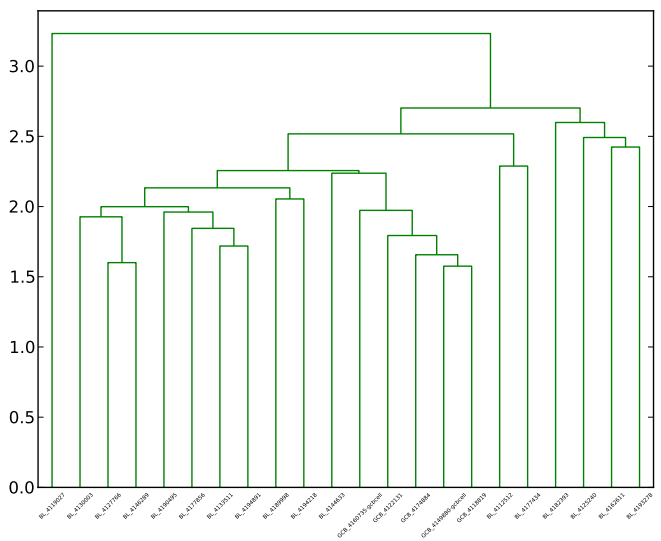
- Finding upper quartil of all genes in regard to average distance between the centre and each of the n compositional vectors
- Calculating all $\binom{n}{2}$ pairwise sample combinations averaged over this set of genes



Clustering and outlier detection



- Finding upper quartil of all genes in regard to average distance between the centre and each of the n compositional vectors
- Calculating all $\binom{n}{2}$ pairwise sample combinations averaged over this set of genes
- Hierarchical agglomerative clustering



Section 3

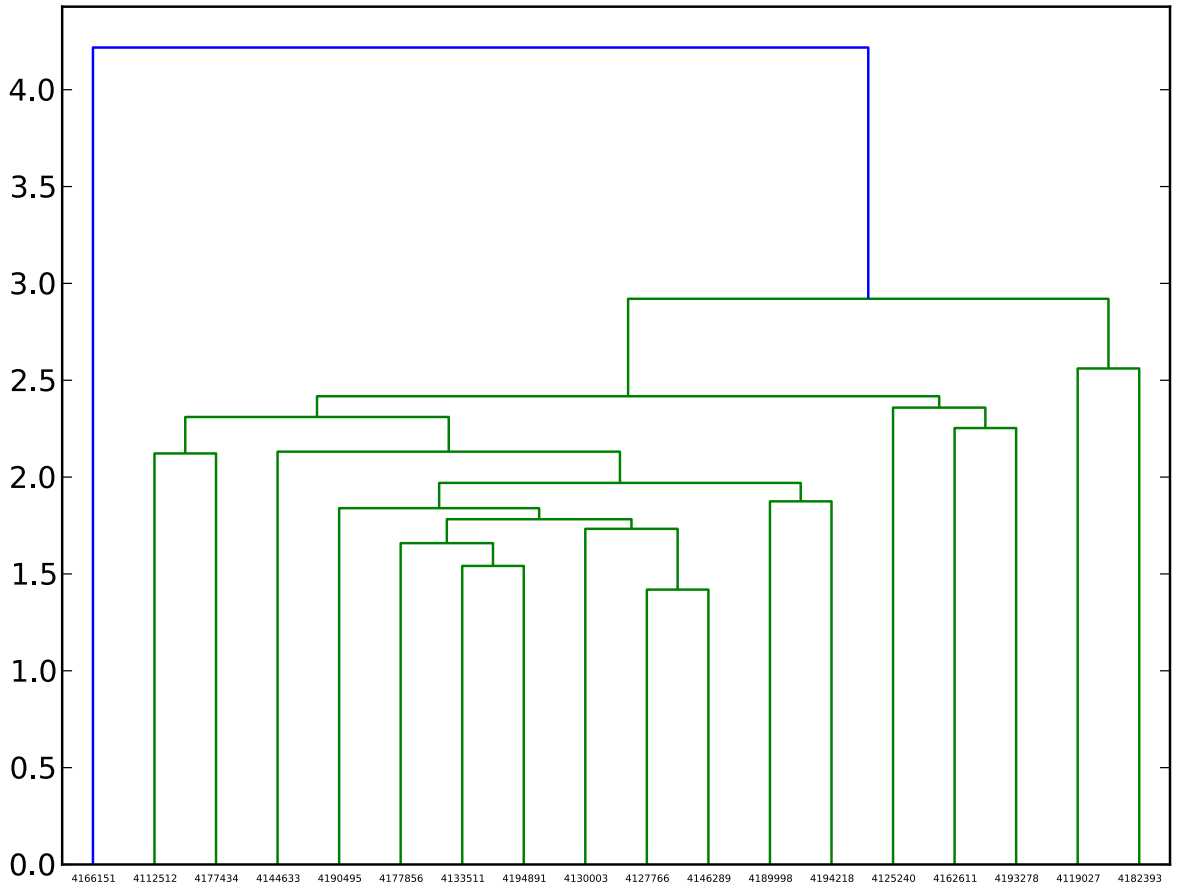
Results

First glimpse at the hidden treasures within the ICGC RNA-seq data

ICGC data set: Germinal Center B-cell Derived Lymphomas

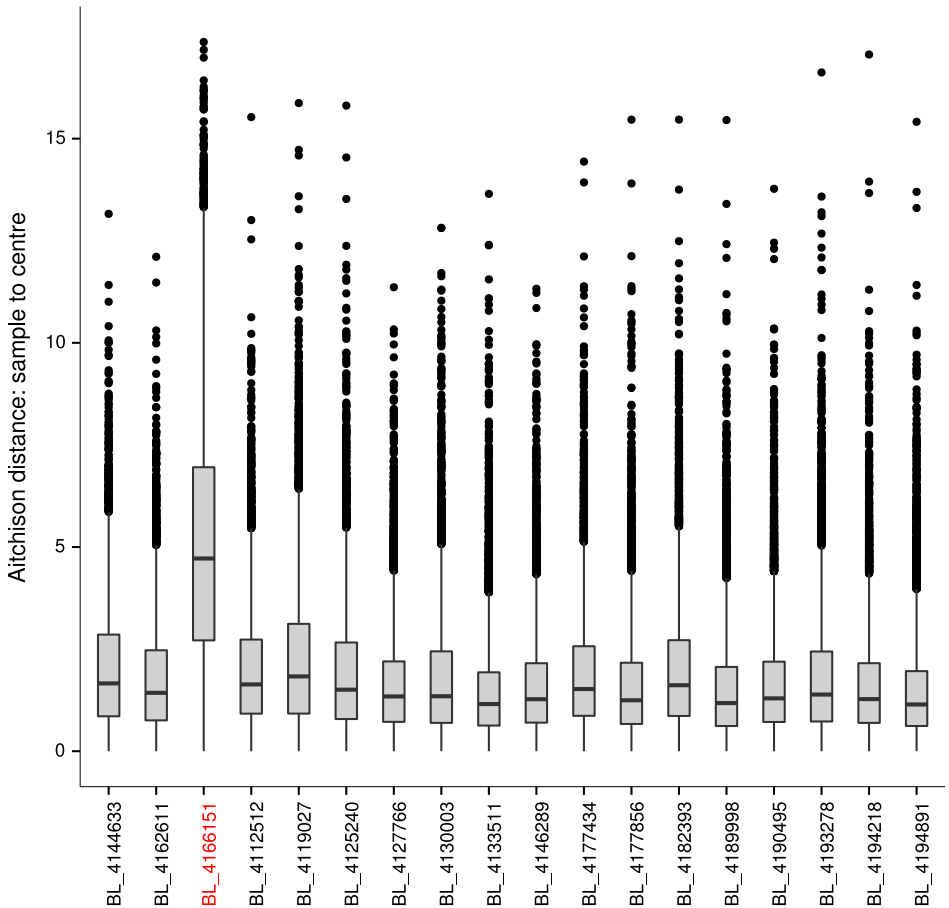
- 6 comparisons (4 conditions: GCB, BL, FL, DLBCL)
- 126 samples (5 GCB, 18 BL, 46 FL, 47 DLBCL)

Outlier detection



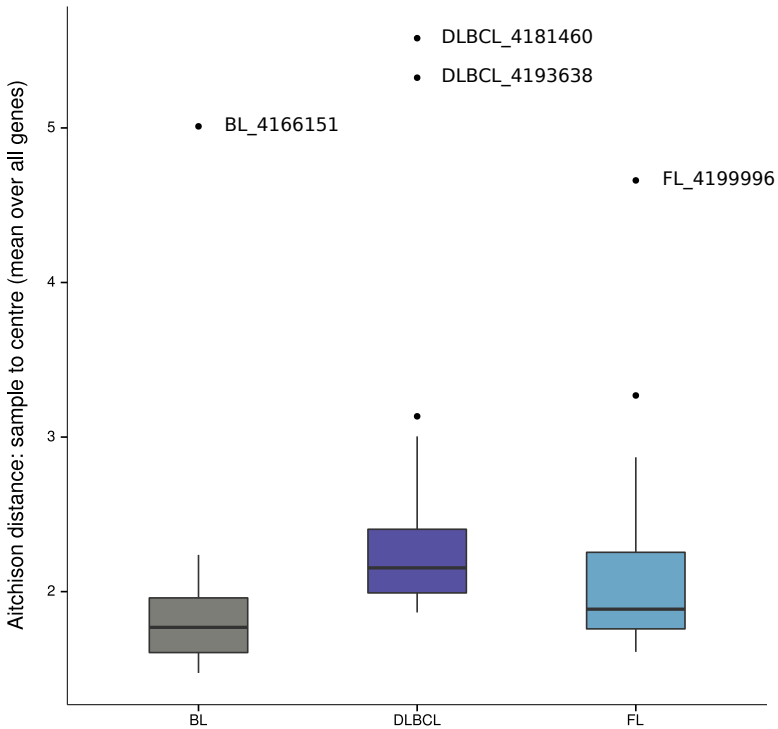
Dendrogram for BL samples

Outlier detection



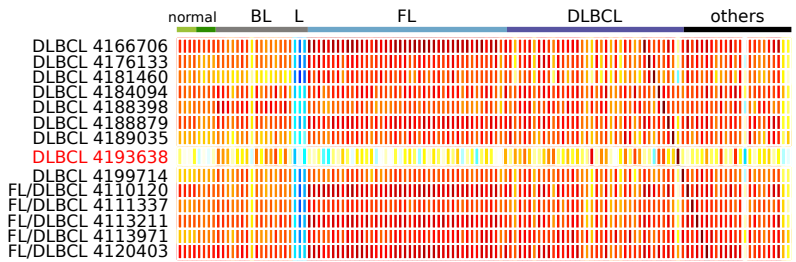
Distance distribution for BL samples

Outlier detection



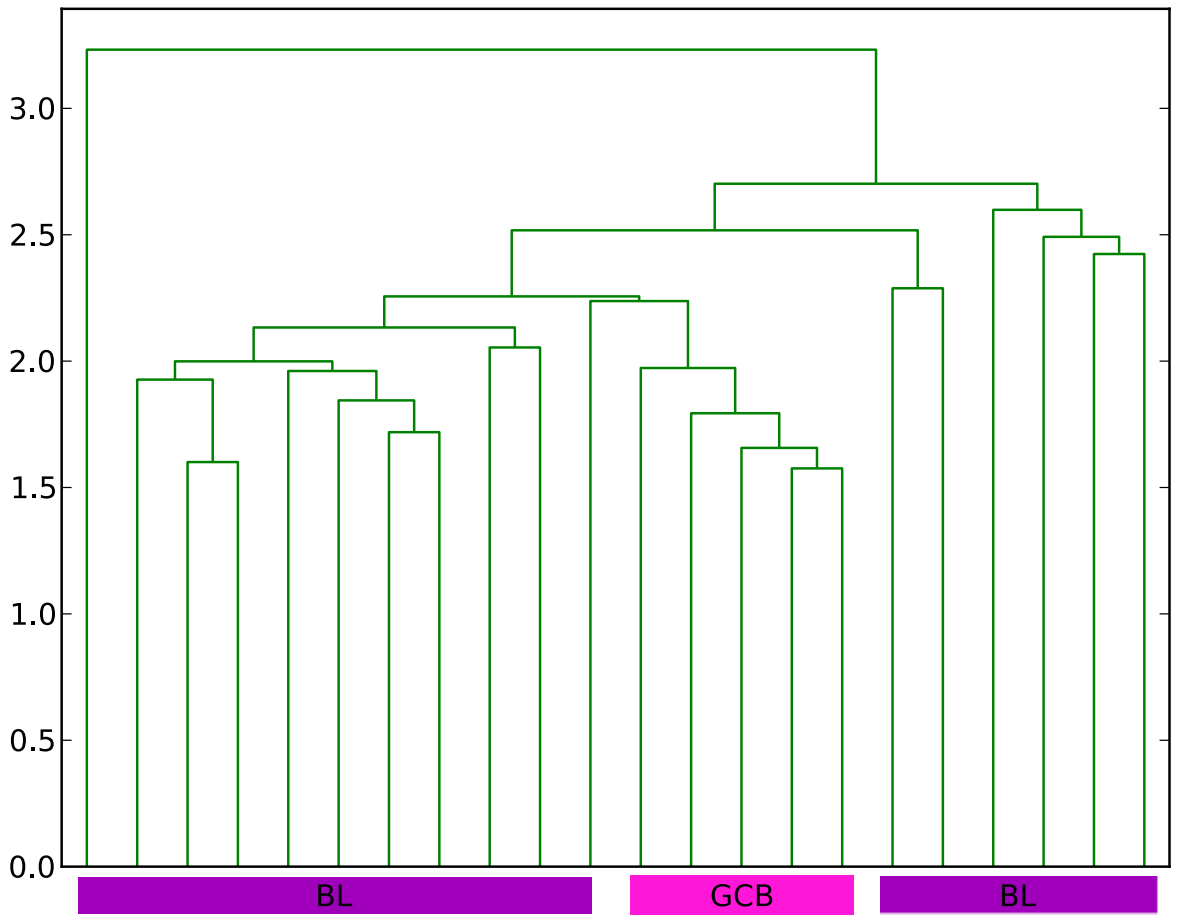
Outlier

- BL 4166151
  - Age 74 (median $\approx$ 10)
- DLBCL 4193638
  - Also an outlier in terms of expression
- FL 4199996 and DLBCL 4181460
  - not conspicuous in terms of methylation or expression



Expression correlation

Subgroup detection



Dendrogram for GCB and BL samples

Differential Alternative Splicing - a statistical overview

- 6 comparisons (4 conditions: GCB, BL, FL, DLBCL)
- 126 samples (5 GCB, 18 BL, 46 FL, 47 DLBCL)
- 442,738 supported splice junctions
- 18,700 supported genes (16,208 protein coding, 2,492 lincRNA)
- Significant if $|\text{abundance change}| \geq 1$ and q-value < 0.01

Differential Alternative Splicing - a statistical overview

- 6 comparisons (4 conditions: GCB, BL, FL, DLBCL)
- 126 samples (5 GCB, 18 BL, 46 FL, 47 DLBCL)
- 442,738 supported splice junctions
- 18,700 supported genes (16,208 protein coding, 2,492 lincRNA)
- Significant if $|\text{abundance change}| \geq 1$ and q-value < 0.01

	GCB-BL	GCB-FL	GCB-DLBCL	BL-FL	BL-DLBCL	FL-DLBCL
	splice junctions					
tested	100,315	106,890	107,498	126,667	127,919	138,212
sig.	1,012	1,629	1,499	1,089	610	132
percent	1.01	1.52	1.39	0.86	0.48	0.10
	genes					
tested	8,229	8,419	8,450	10,304	10,426	11,418
sig.	756	1,077	1,016	738	424	108
percent	9.19	12.79	12.02	7.16	4.07	0.95

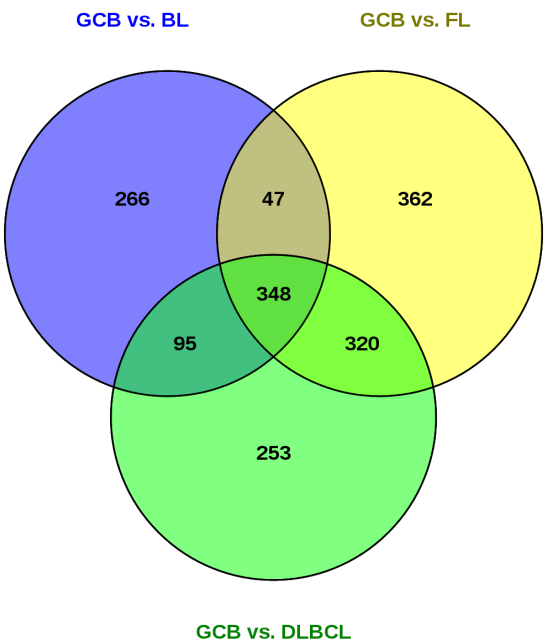
Differential Alternative Splicing - a statistical overview

- 6 comparisons (4 conditions: GCB, BL, FL, DLBCL)
- 126 samples (5 GCB, 18 BL, 46 FL, 47 DLBCL)
- 442,738 supported splice junctions
- 18,700 supported genes (16,208 protein coding, 2,492 lincRNA)
- Significant if $|\text{abundance change}| \geq 1$ and q-value < 0.01

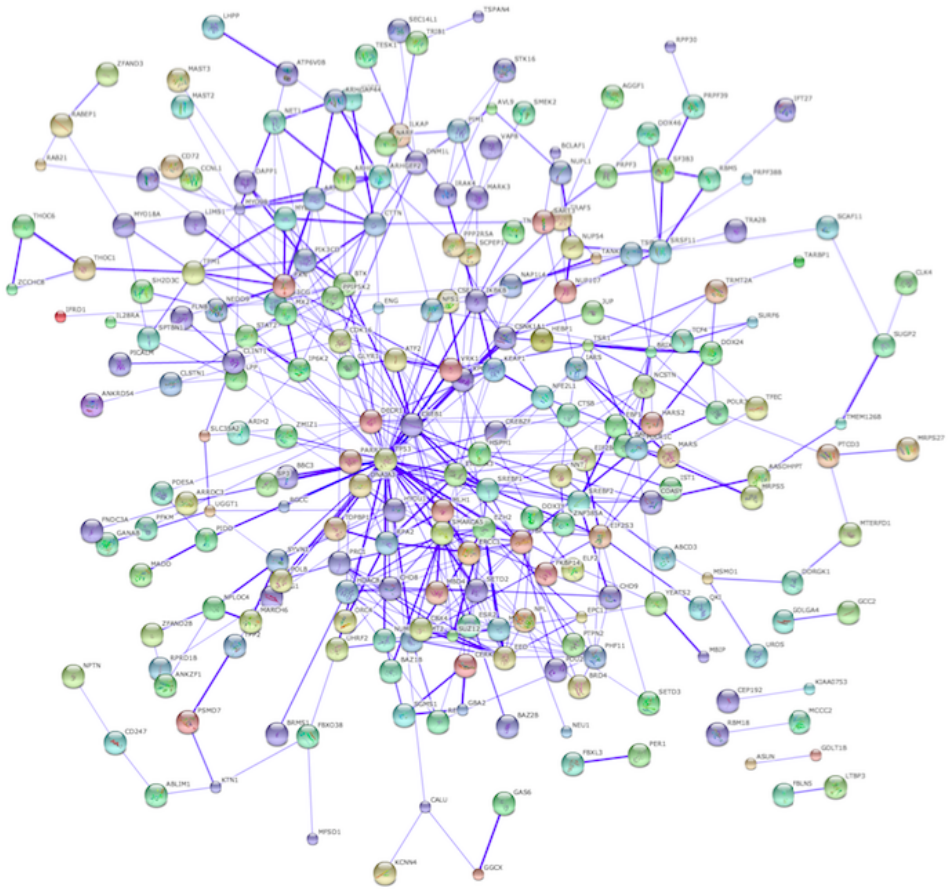
	GCB-BL	GCB-FL	GCB-DLBCL	BL-FL	BL-DLBCL	FL-DLBCL
	splice junctions					
tested	100,315	106,890	107,498	126,667	127,919	138,212
sig.	1,012	1,629	1,499	1,089	610	132
percent	1.01	1.52	1.39	0.86	0.48	0.10
	genes					
tested	8,229	8,419	8,450	10,304	10,426	11,418
sig.	756	1,077	1,016	738	424	108
percent	9.19	12.79	12.02	7.16	4.07	0.95

- Randomizing group info $\Rightarrow$ no significant results

Lymphoma common genes



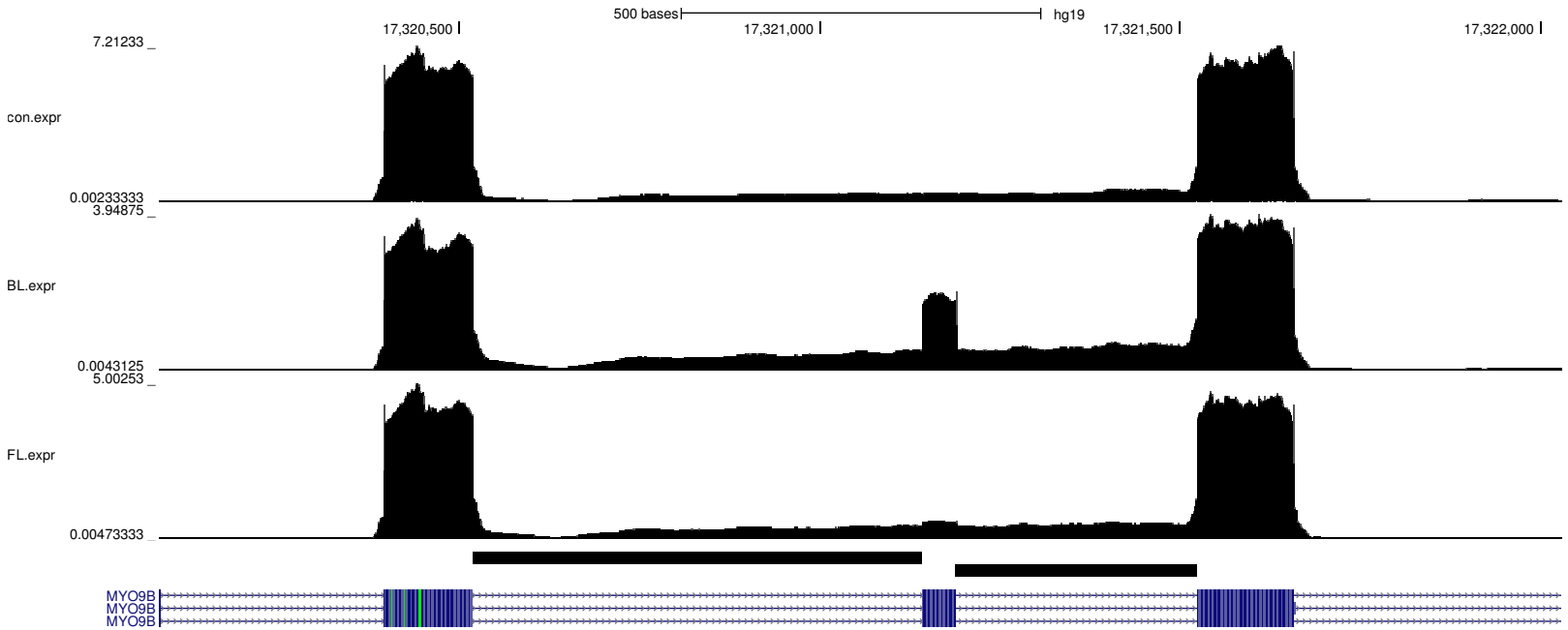
Overlap of sig. genes



Network of overlap

Lymphoma common genes

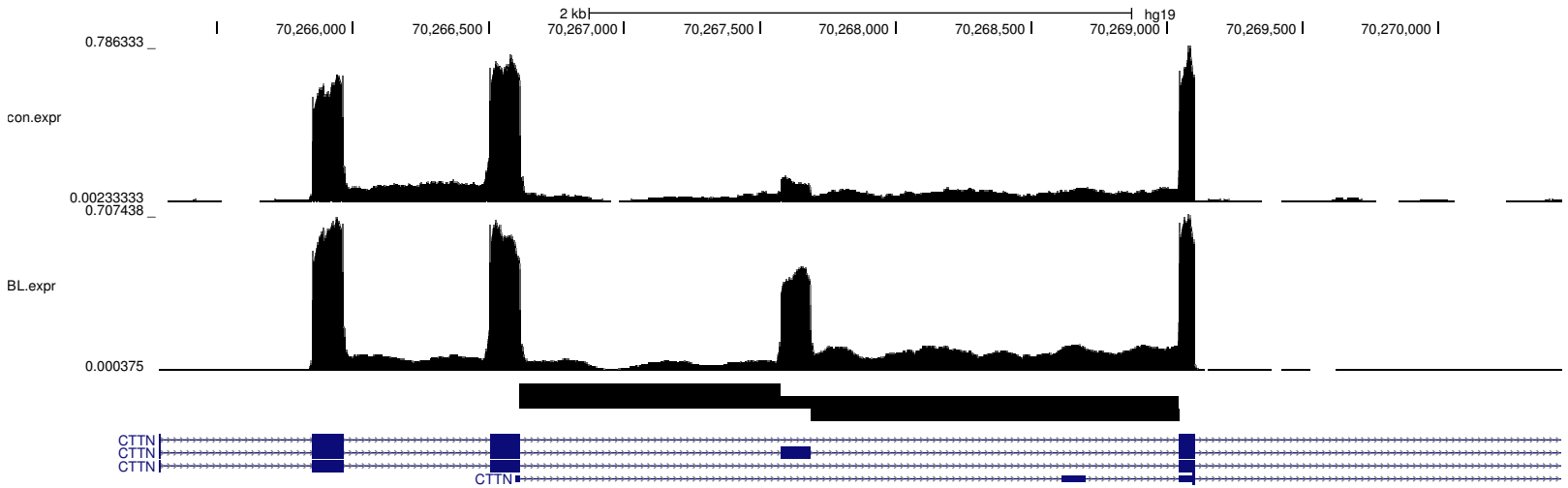
- High protein expression in lymph nodes and other immune related tissues



MYO9B

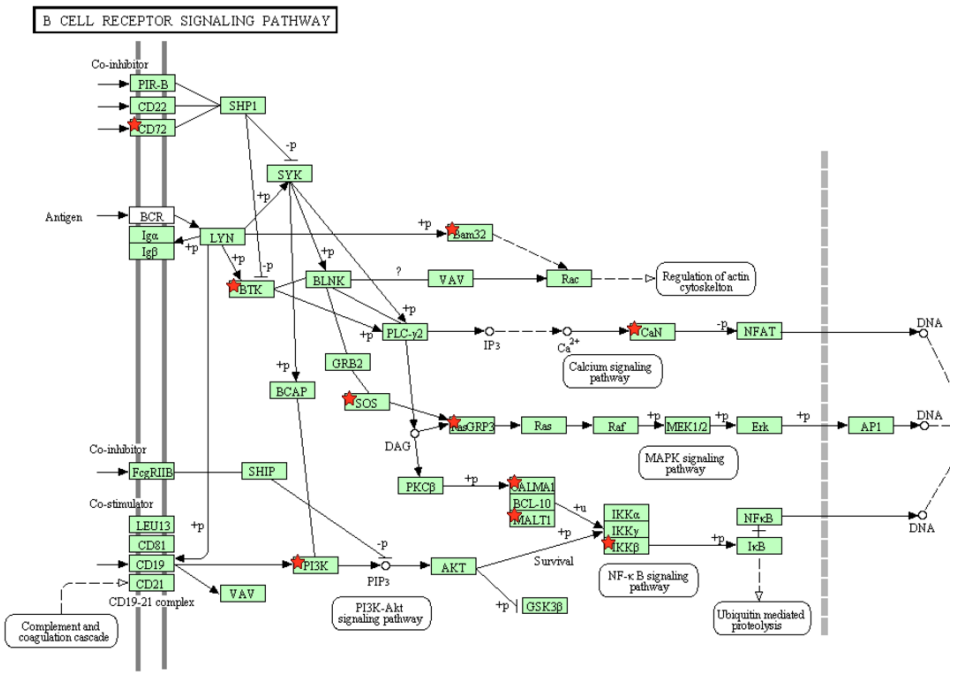
Lymphoma common genes

- Plays a role in the invasiveness of cancer cells, and the formation of metastases

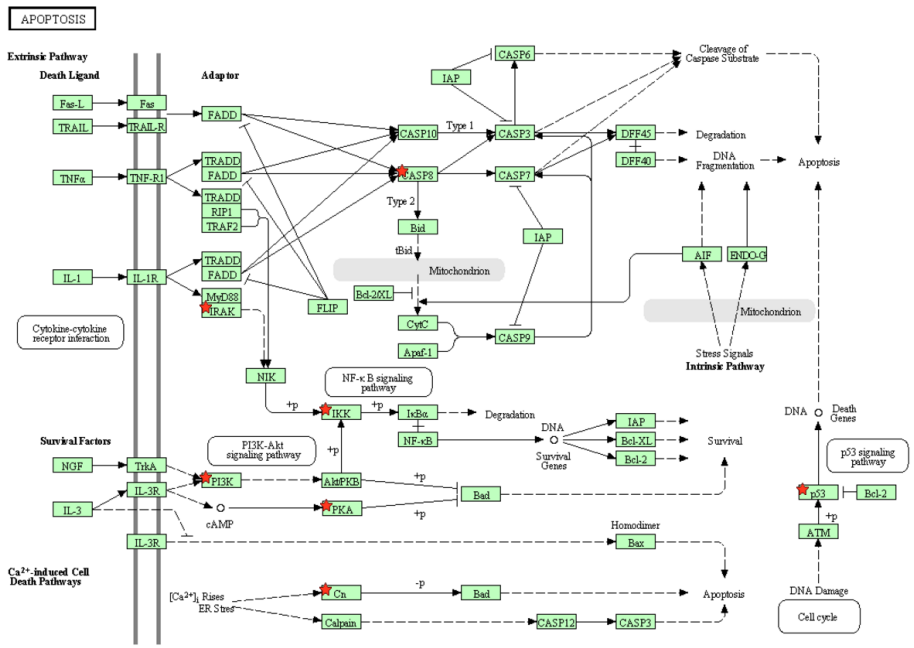


CTTN

Enriched pathways (GCB vs BL)

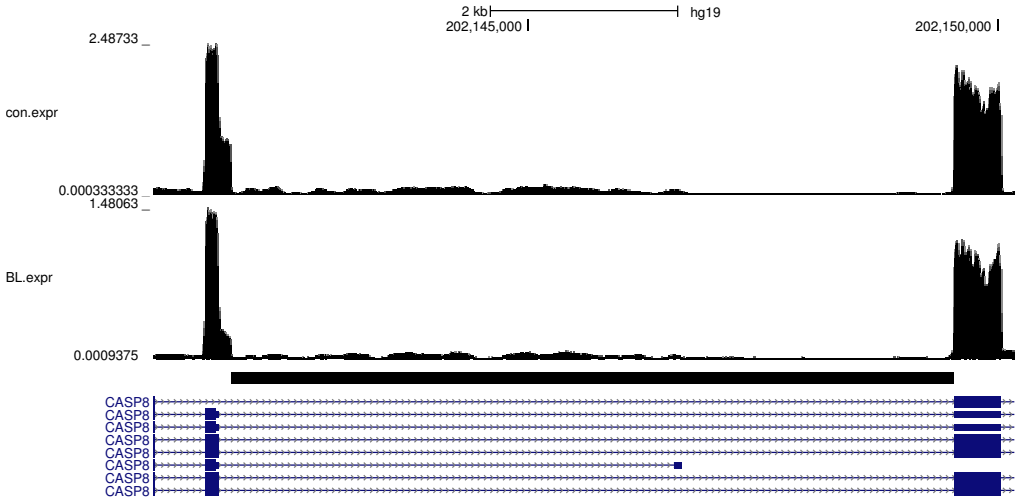


B-cell receptor

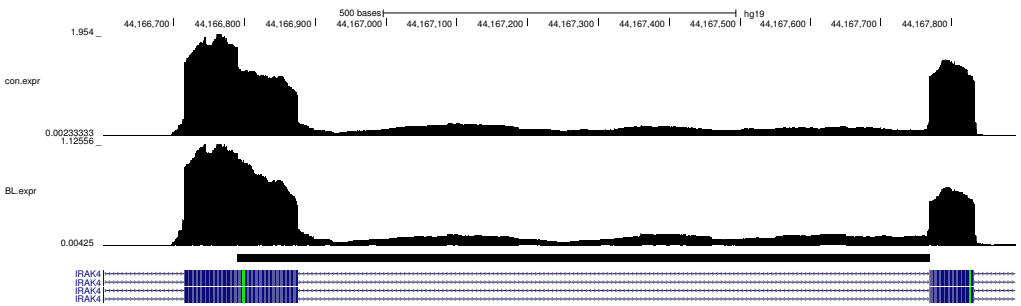


Apoptosis

Apoptosis pathway (GCB vs. BL)

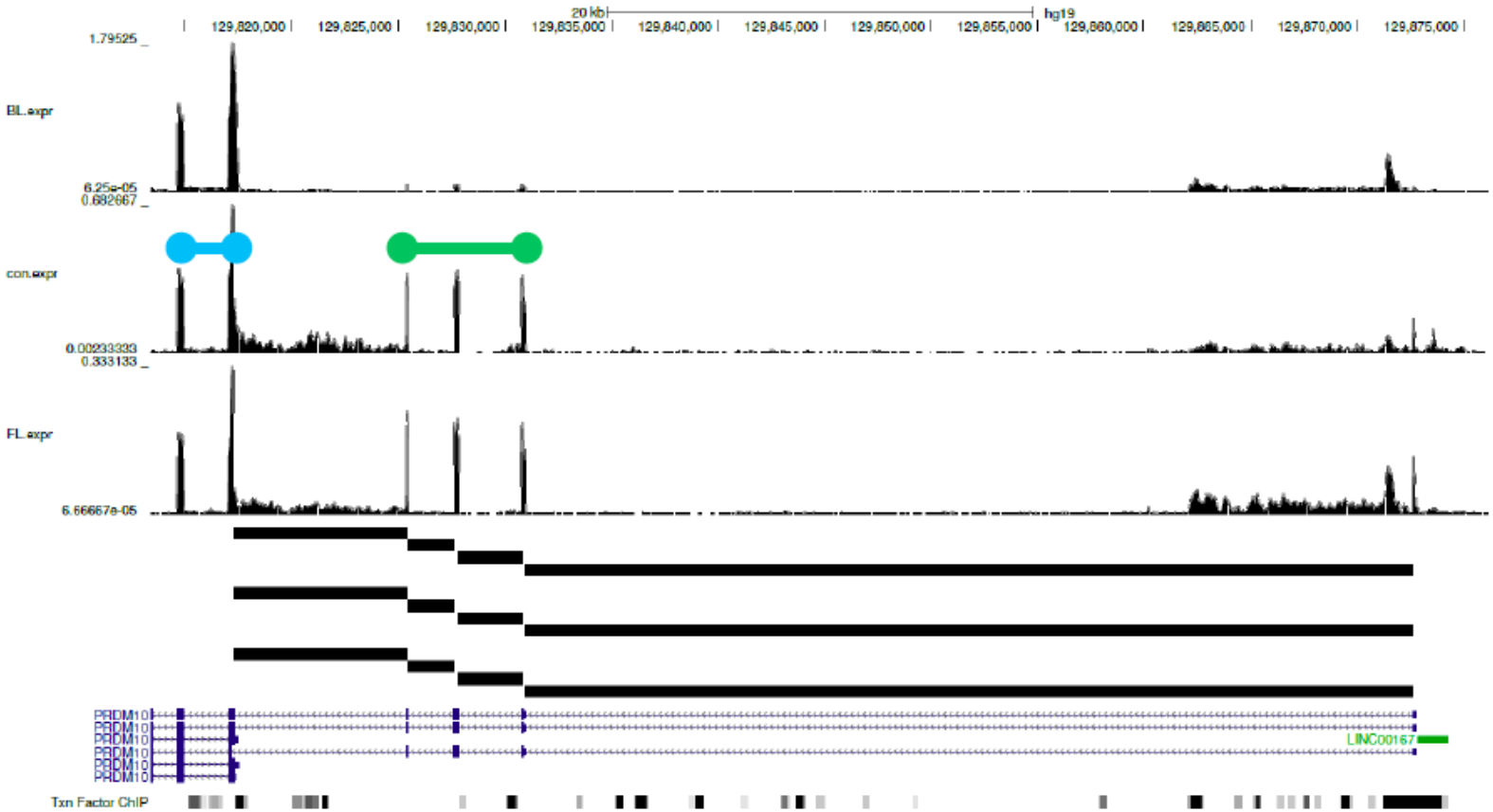


CASP8



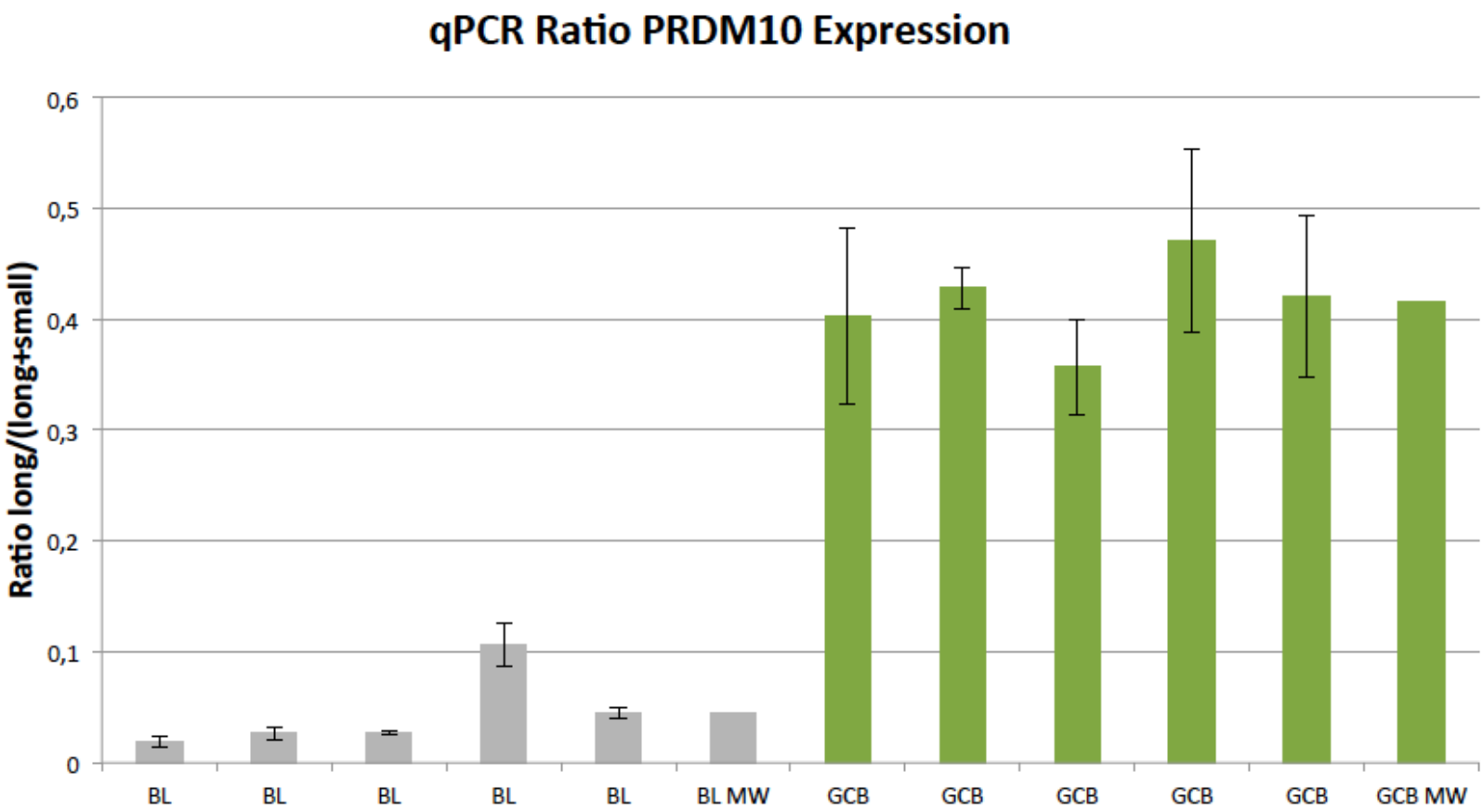
IRAK4

Validation by qPCR



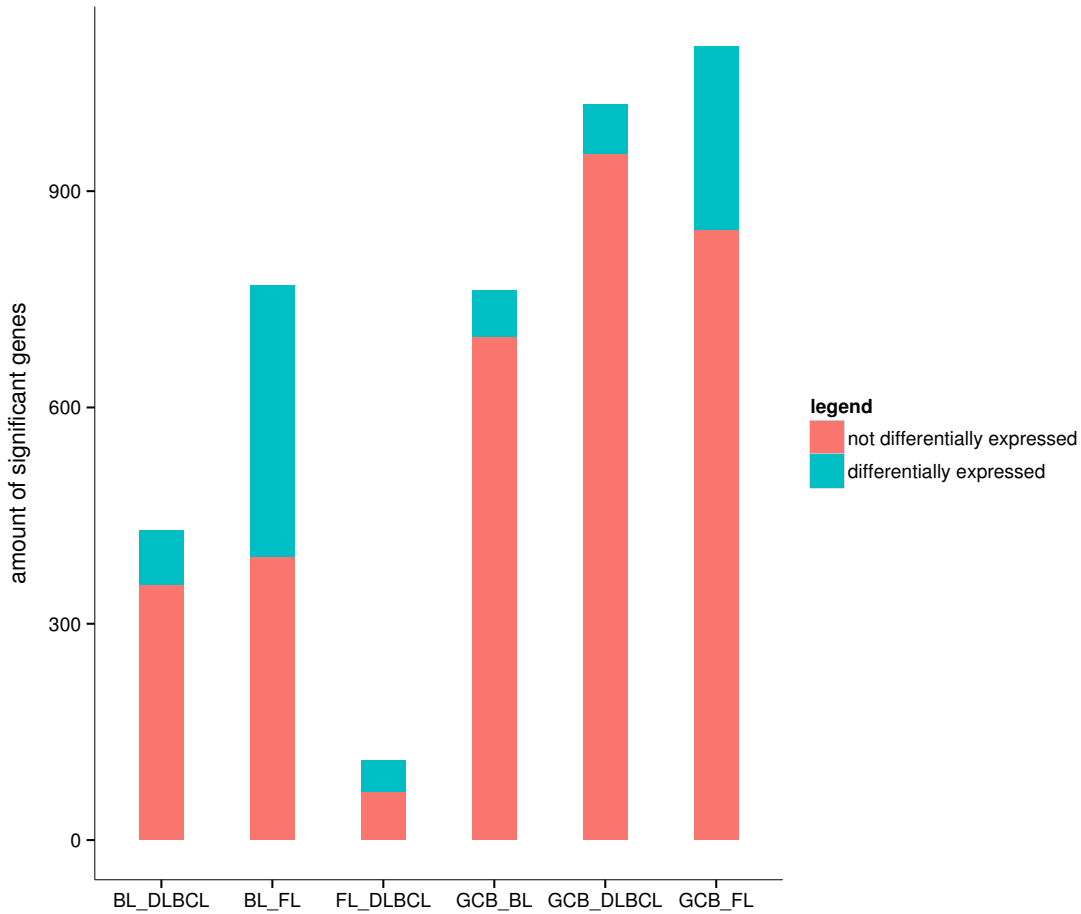
PRDM10

Validation by qPCR



PRDM10

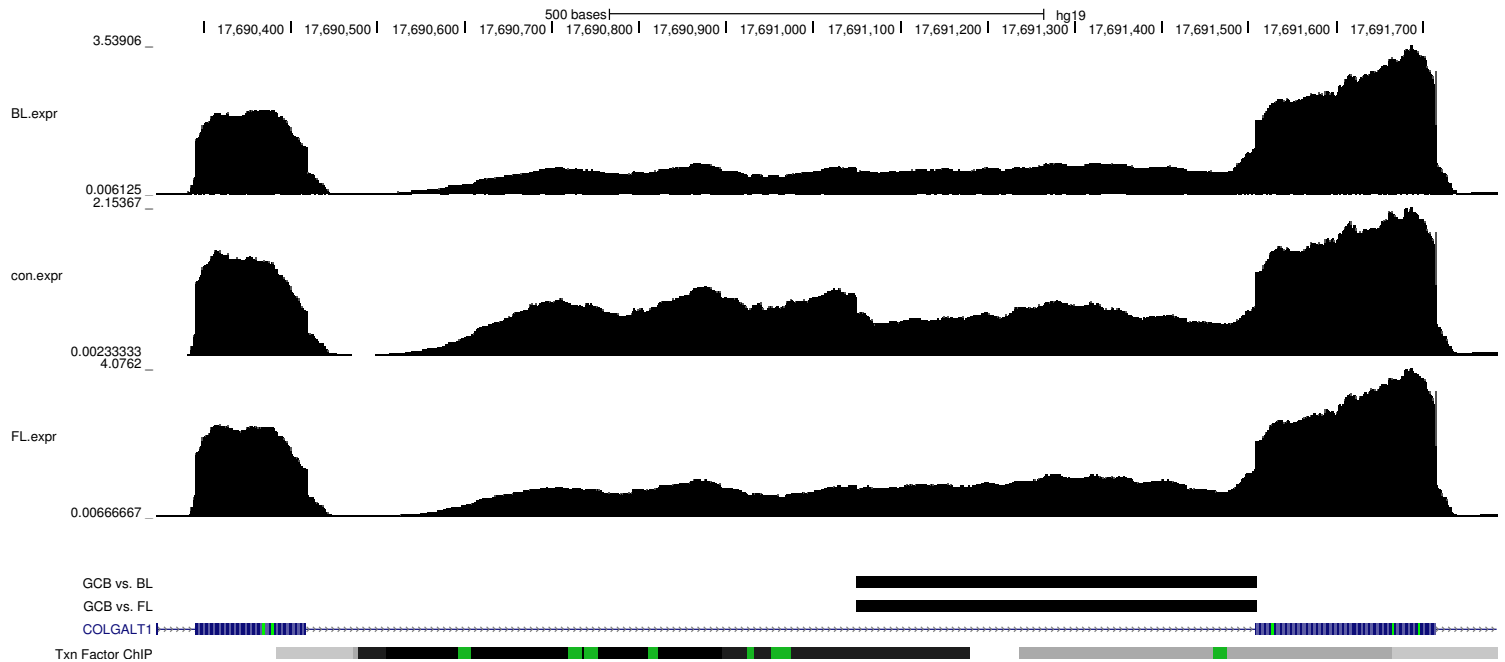
Connection between diff. alt. splicing and diff. expression



Intersection with sig. genes based on DESeq

Connection between diff. alt. splicing and RNA editing

- Calling sig. differentially RNA editing sites between conditions (with Methylene)
- Intersected with both splice sites of all sig. diff. alt. splice junctions



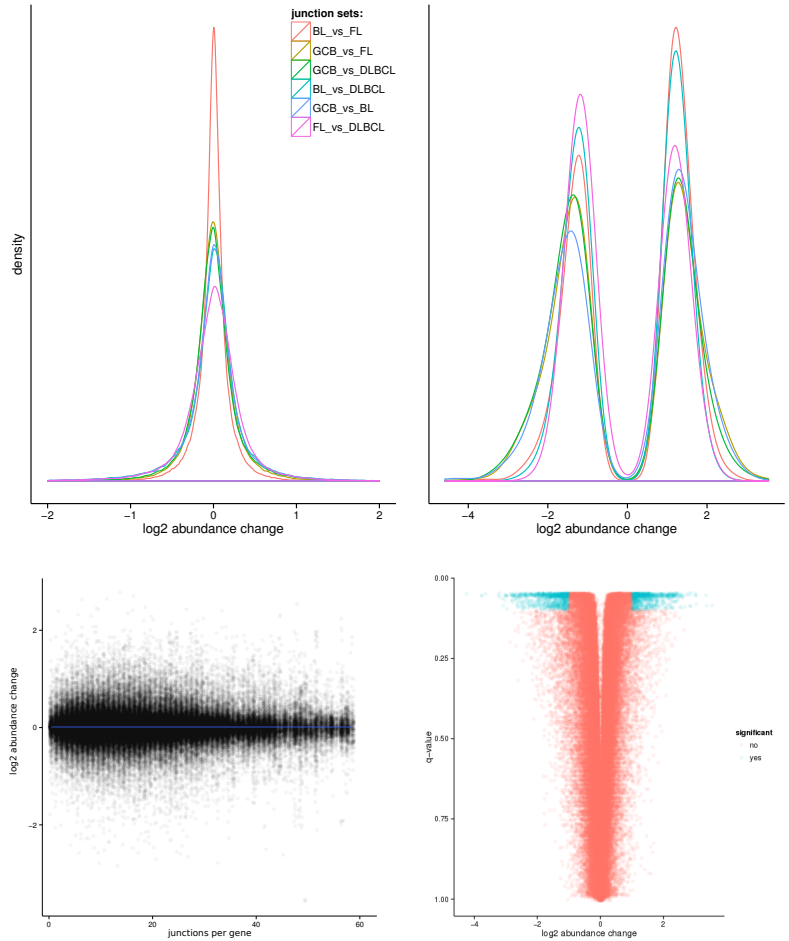
RNA editing at position chr19:17,691,052

Connection between diff. alt. splicing and DMRs

- Building 2 x 2 contingency tables of gene counts in regard to diff. alt. splicing and DMR overlap.
- Odds ratios and p-values from Fisher's exact test:

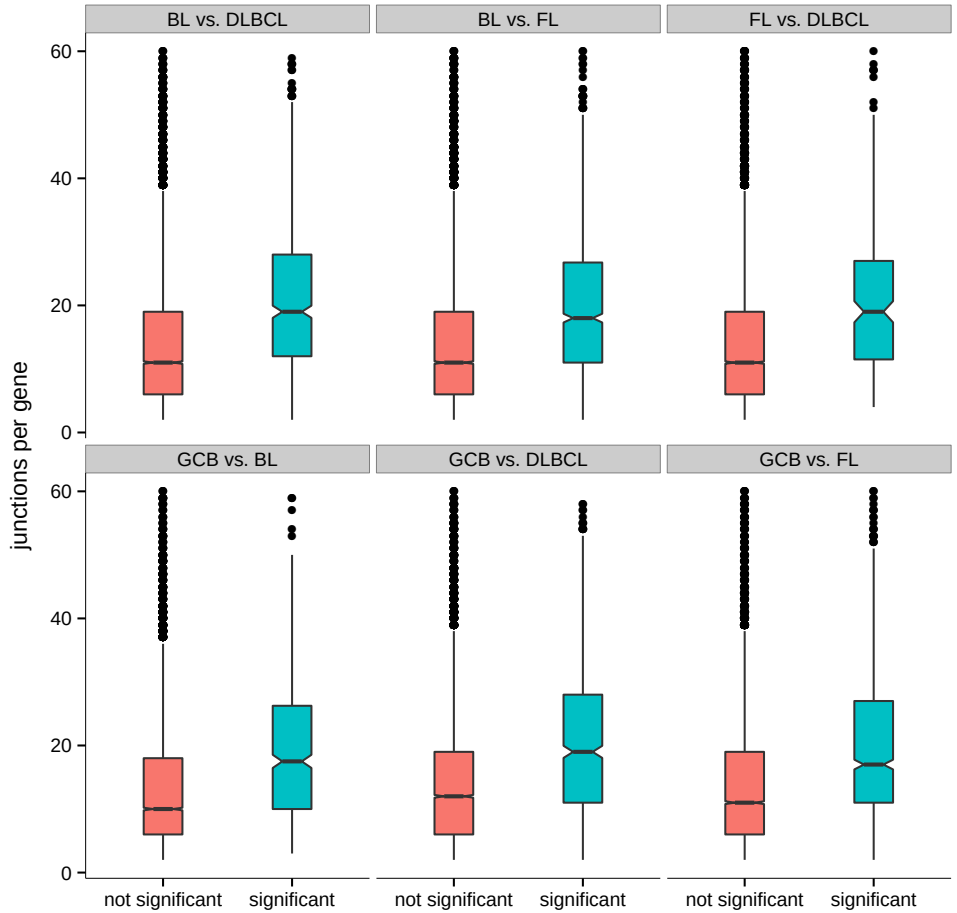
	GCB-BL	GCB-FL	BL-FL
odds ratio	1.3	1.48	1.84
p-value	0.00075	6.273e-09	1.719e-15
genes yes yes	336	449	400

Abundance changes



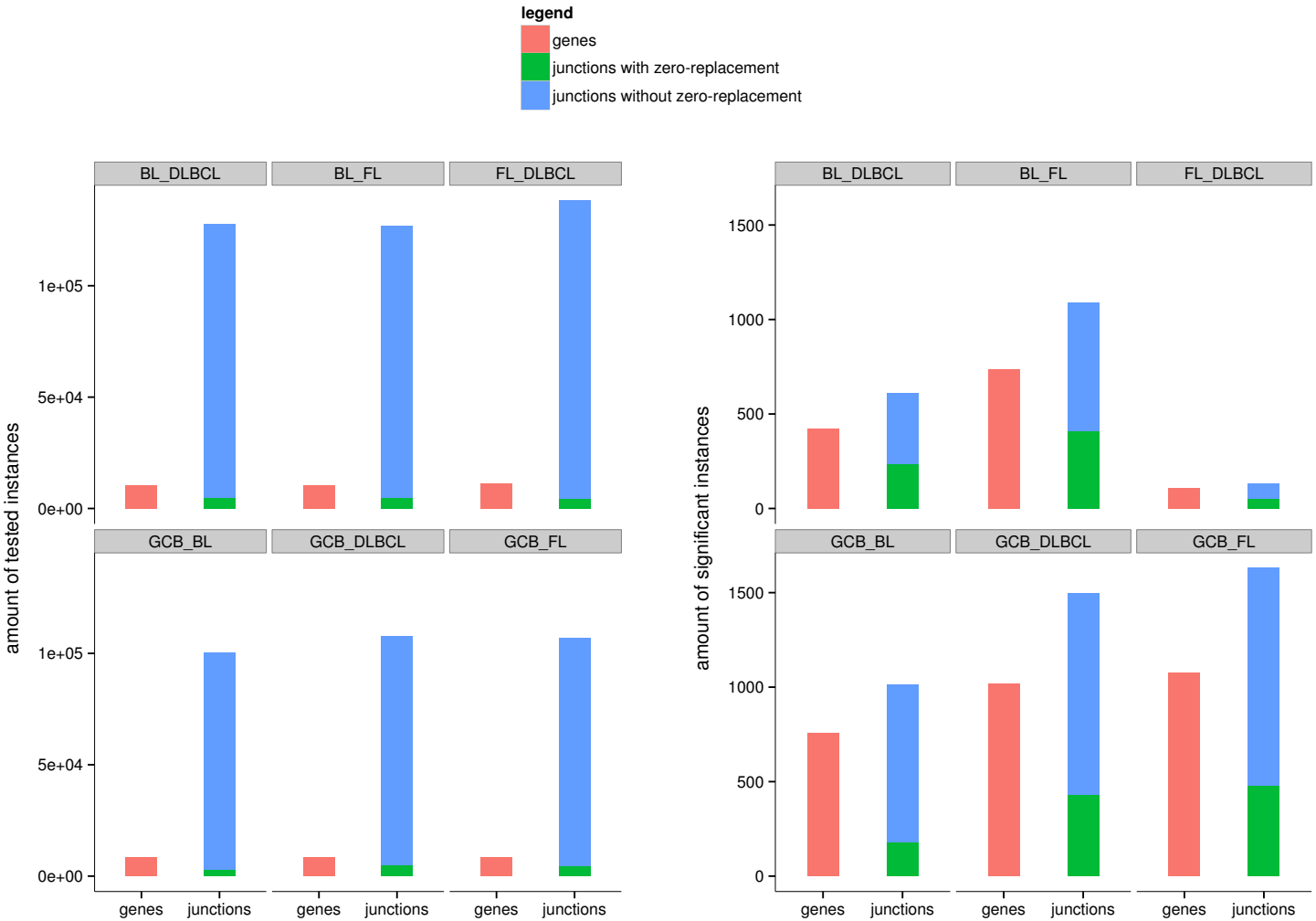
Abundance change as a measure for biological relevance

Junctions per gene



Natural bias towards genes with more junctions

Zero-replacement portion



Tested instances vs. significant instances

Summary

- Simple and robust method
- Based only on direct splice evidence
- Not restricted to current annotations
- Fast core algorithm (3h for 419 samples)
- Plausible results

Thanks to

- Peter F. Stadler
- Steve Hoffmann
- Stephan Bernhart
- Helene Kretzmer
- Reiner Siebert
- Rabea Wagener

Thank you for your attention!

Questions?!

Thank you for your attention!

Questions?!

Thank you for your attention!

Questions?!