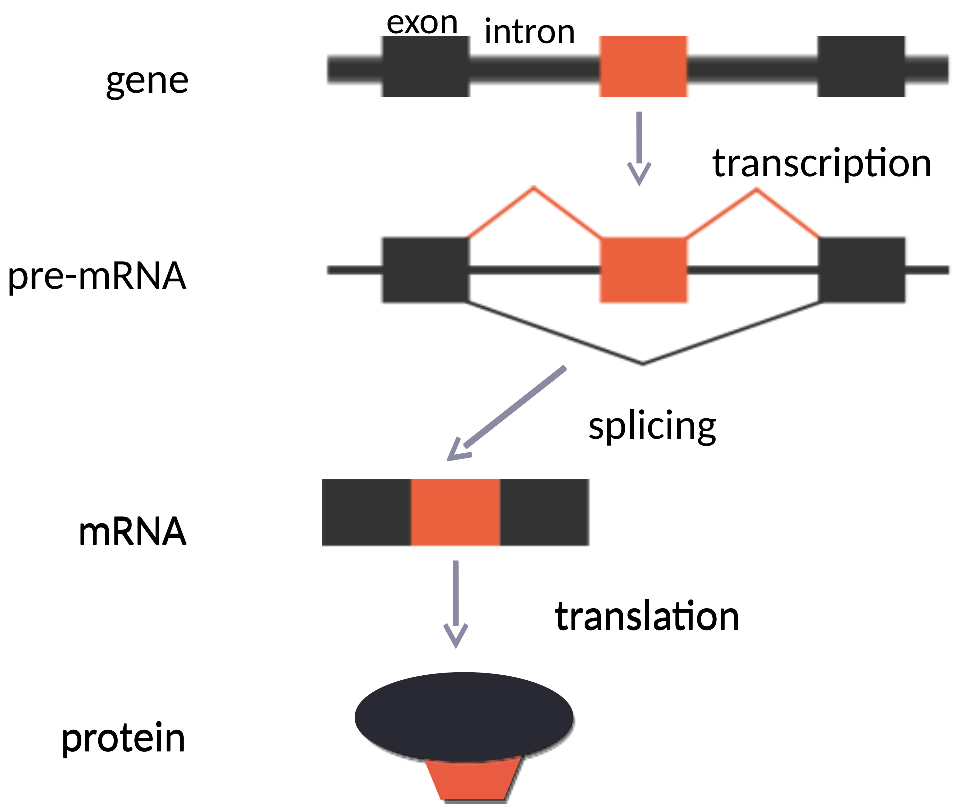
Splicing



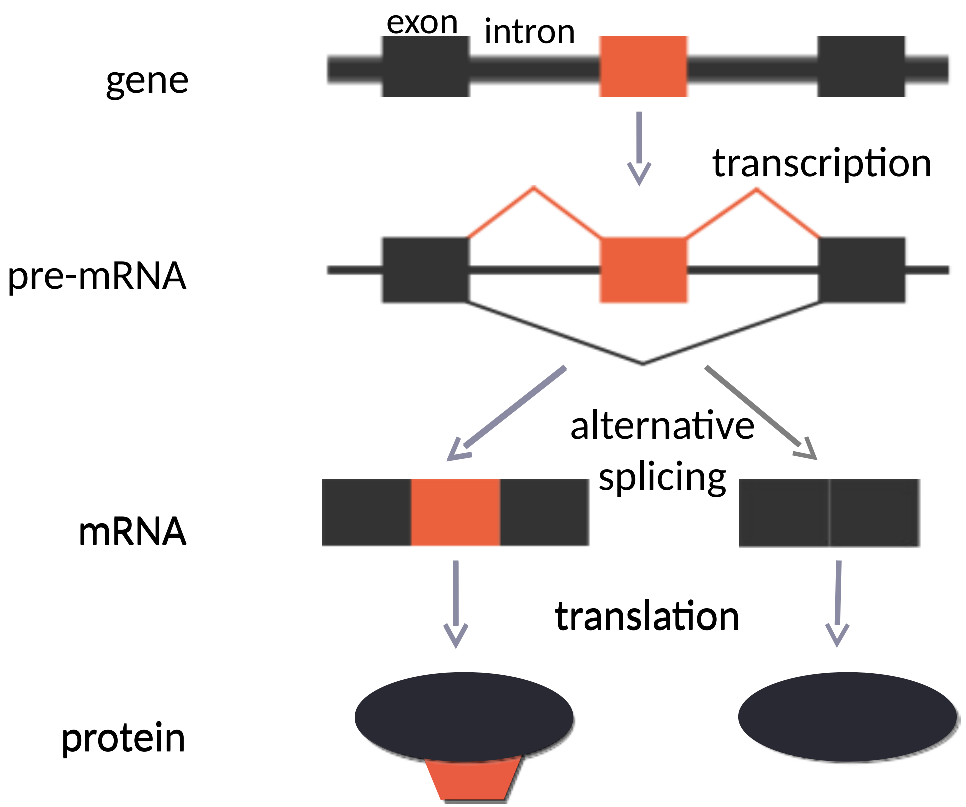
gene

pre-mRNA

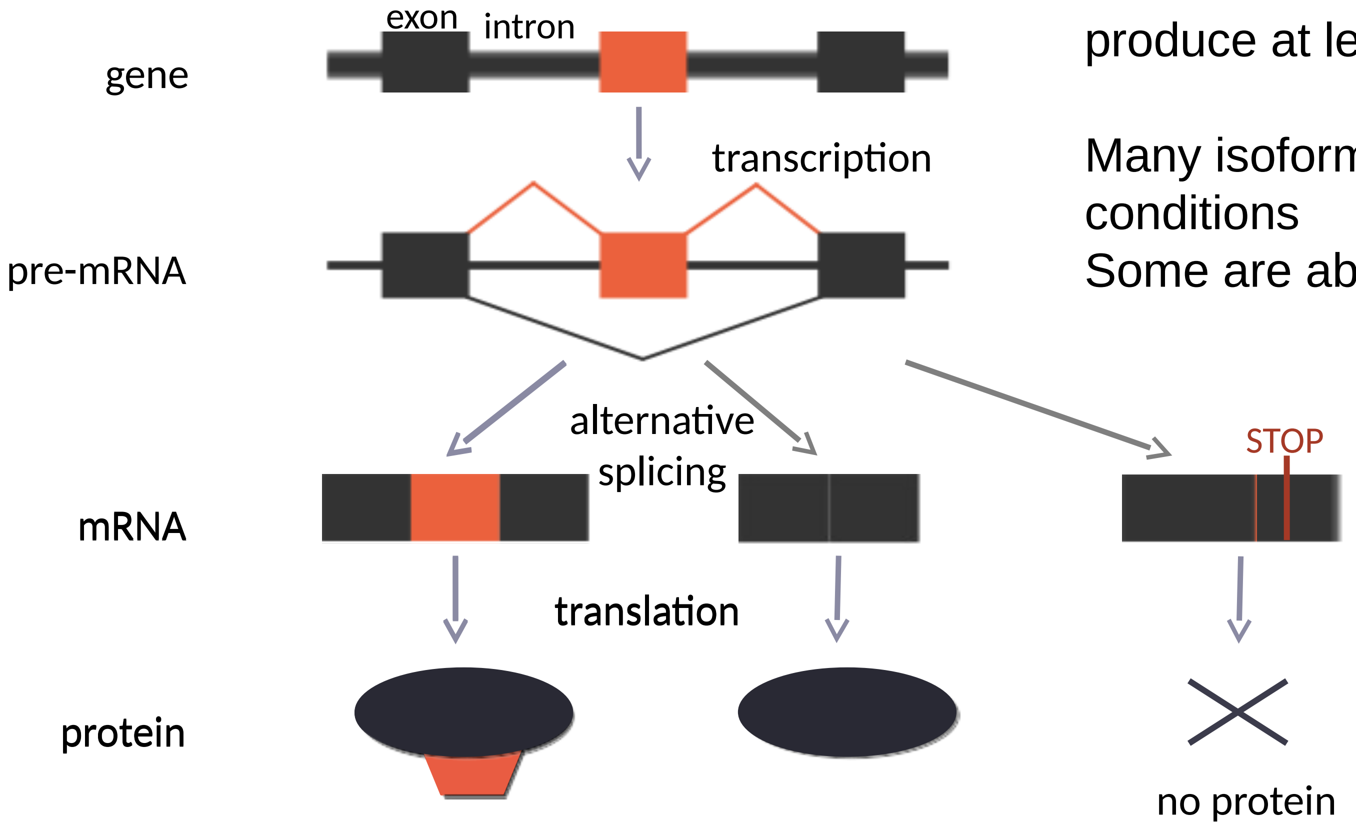
mRNA

protein

Alternative Splicing



Alternative Splicing

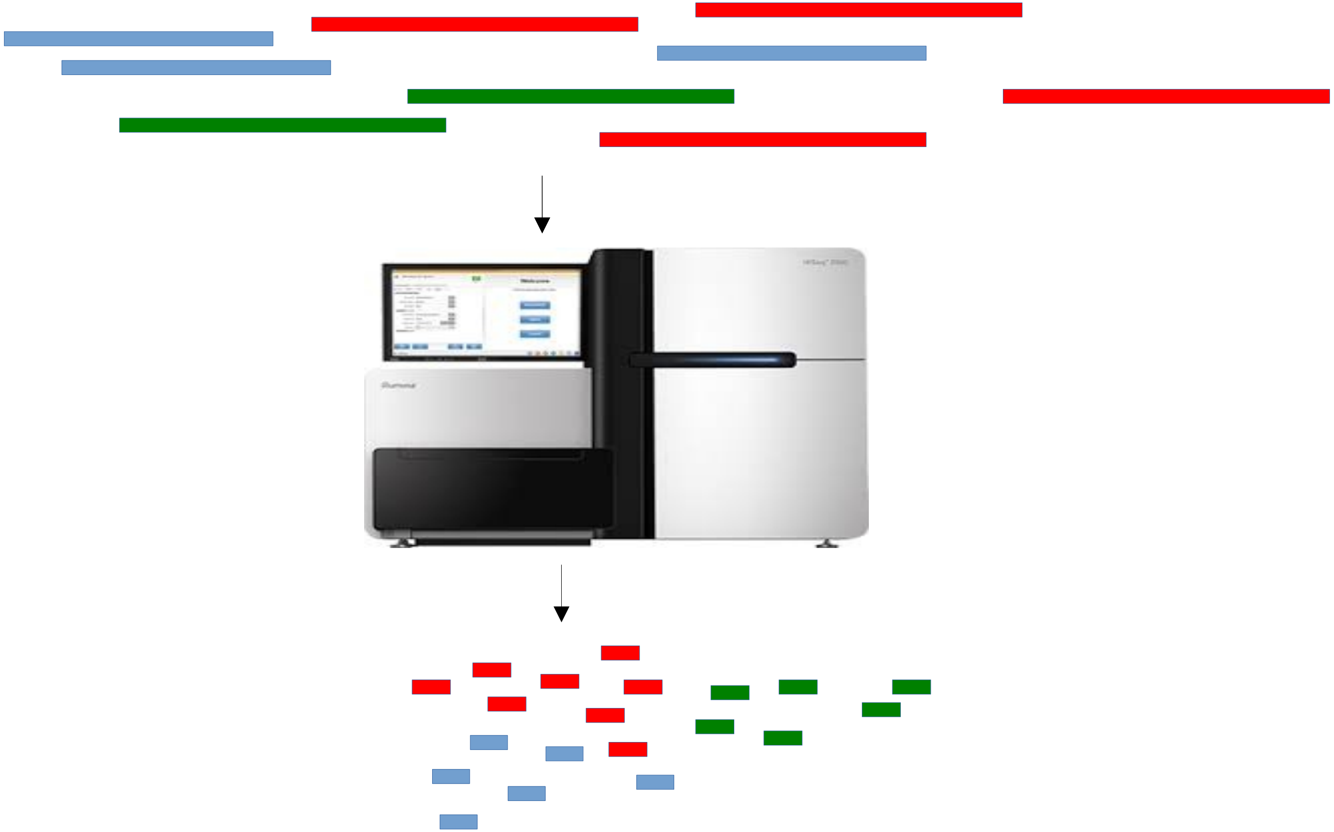


More than 90 % of multi-exon genes produce at least 2 isoforms

Many isoforms are rare in physiological conditions

Some are abundant in a specific condition

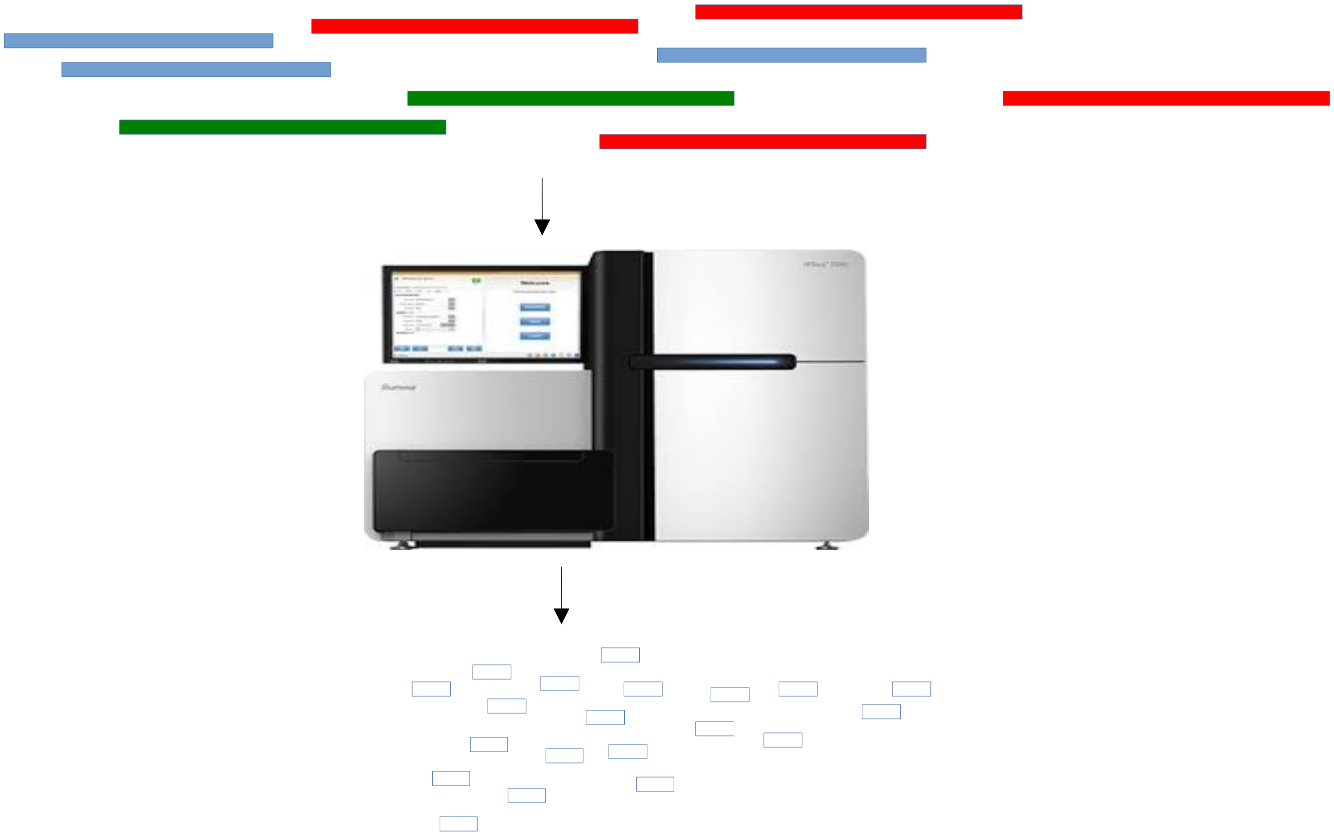
RNAseq data



mRNAs
(~1000nt)

Reads
(100nt)

RNAseq data



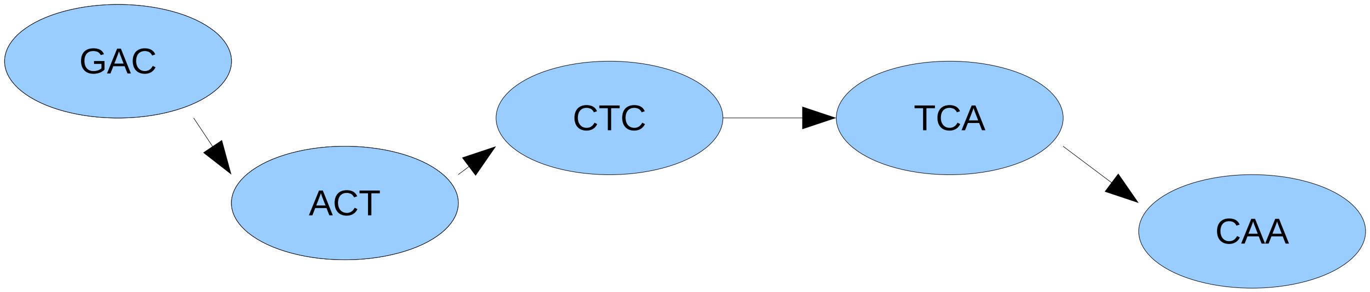
mRNAs
(~1000nt)

Reads
(100nt)

De Bruijn graph

- De Bruijn graphs (DBG) are used as a first step in many short reads assemblers.
- Node = k-mer, Edge = overlap of k-1 bases
- Example:

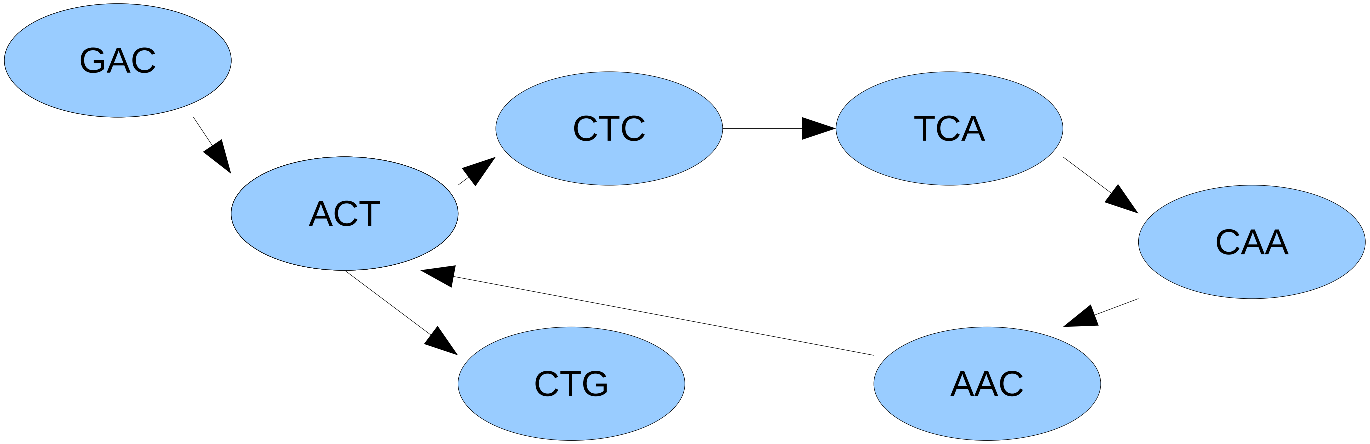
GACTCAA, k=3



De Bruijn graph

- More complicated example
- reference GACTCAACTG (unknown)

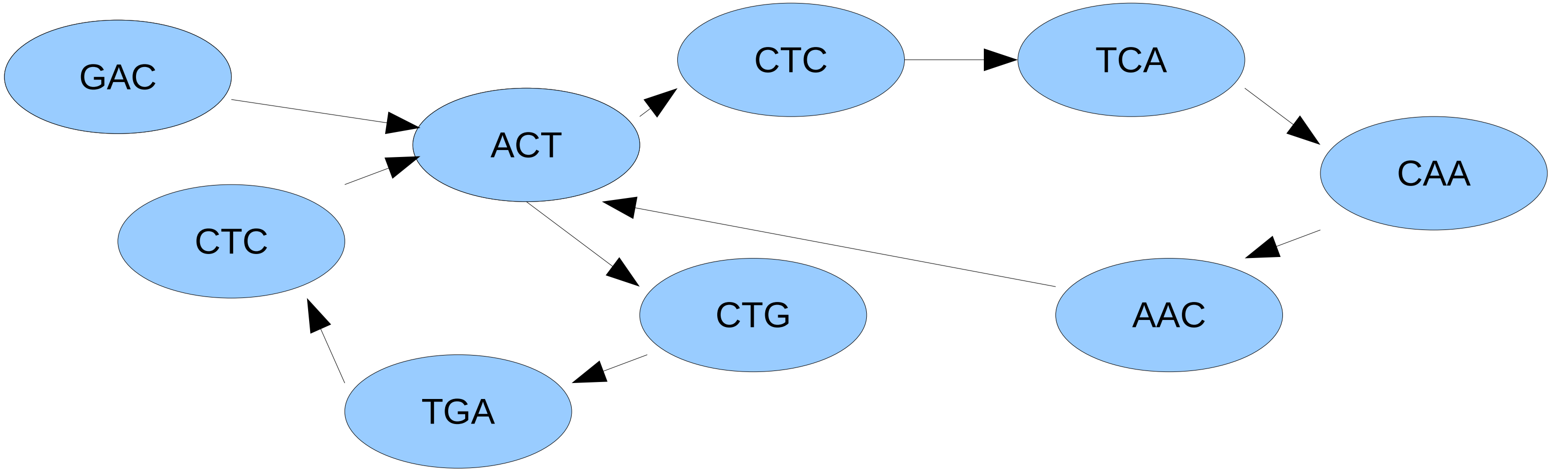
read1 GACTCA
read2 CAACTG



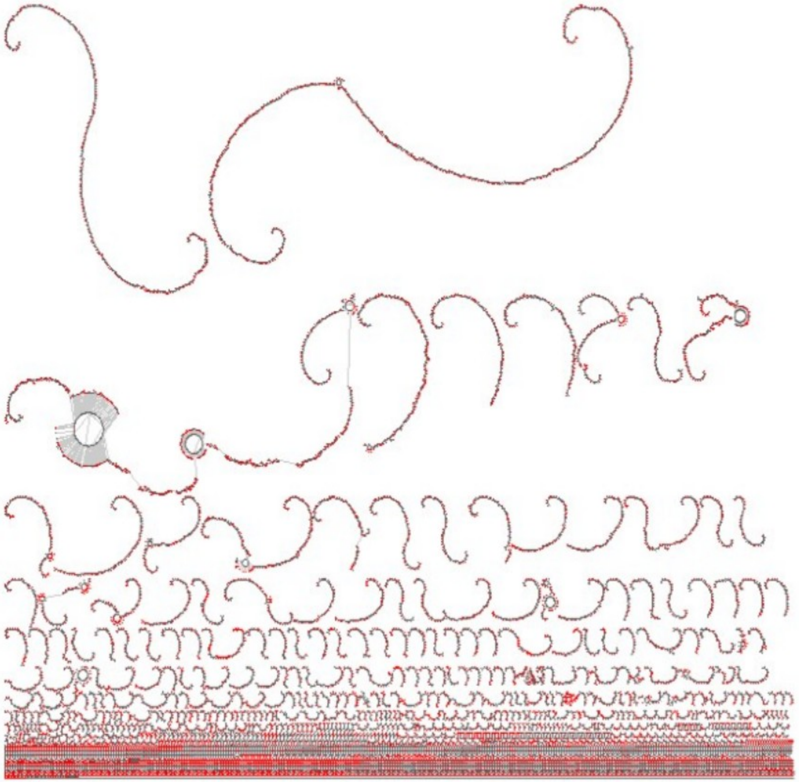
De Bruijn graph

- Even more complicated example
- reference GACTCAACTGACT (unknown)

  read1 GACTCA
  read2 CAACTG
  read3 CTGACT

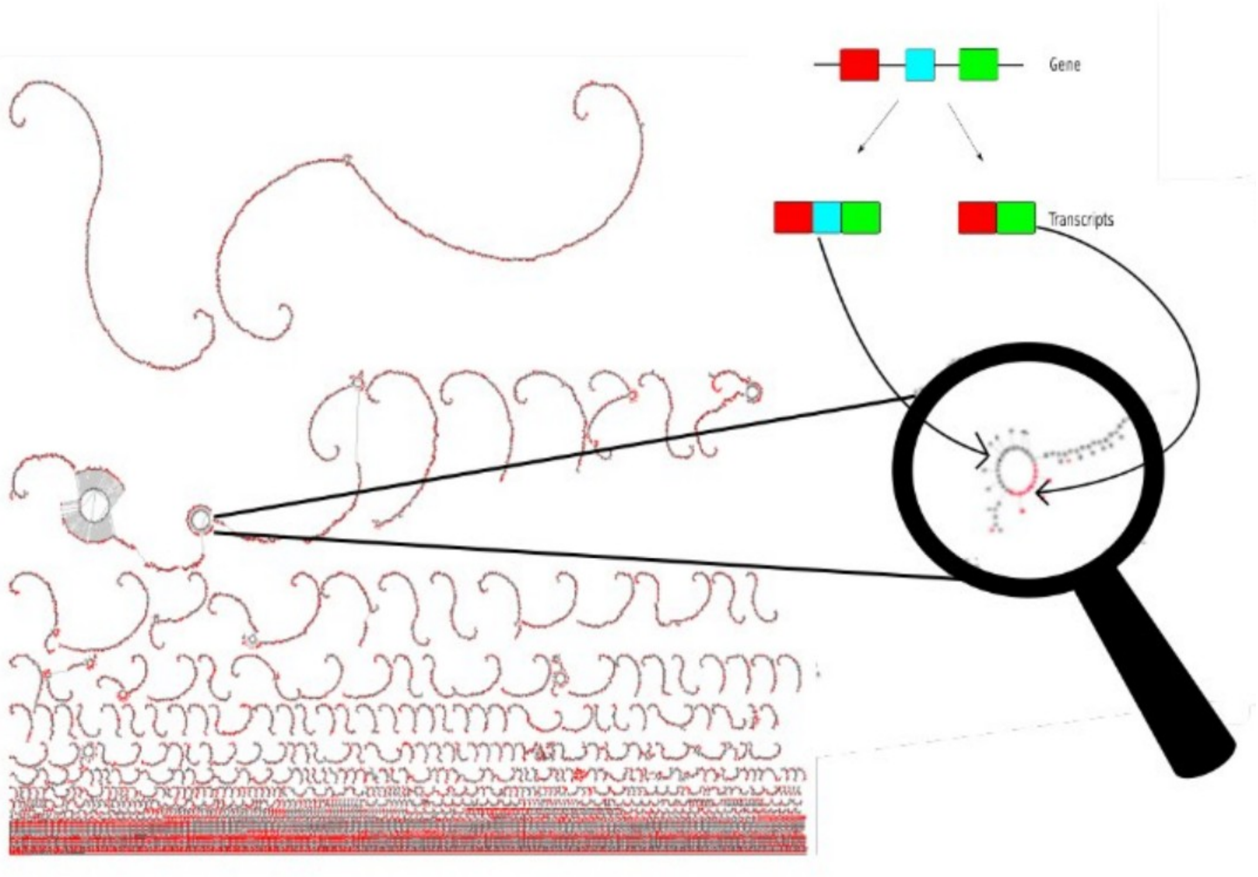


DBG from RNAseq data



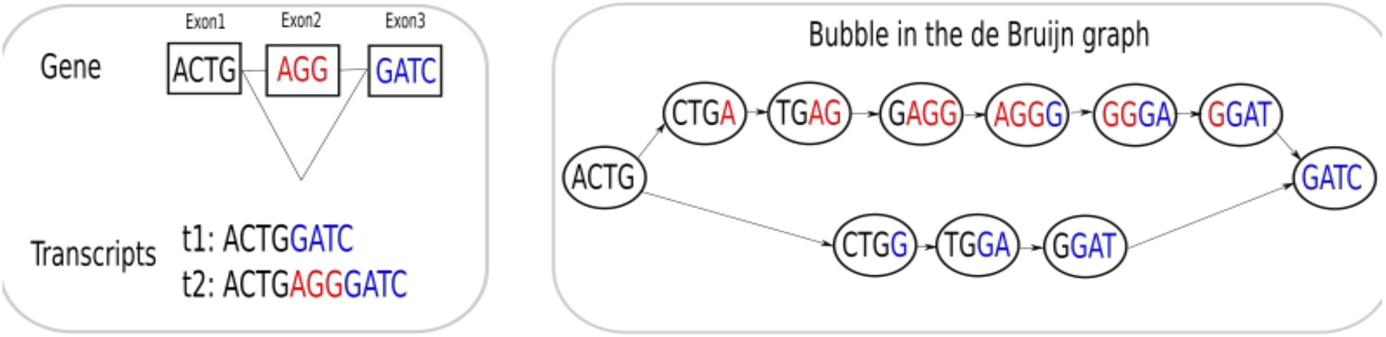
Drosophila transcriptome, shallow coverage (100k reads)

DBG from RNAseq data



Drosophila transcriptome, shallow coverage (100k reads)

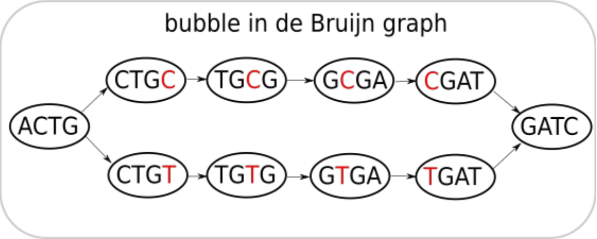
An alternative splicing event corresponds to a bubble in the DBG



SNPs and indels also generate bubbles in the DBG

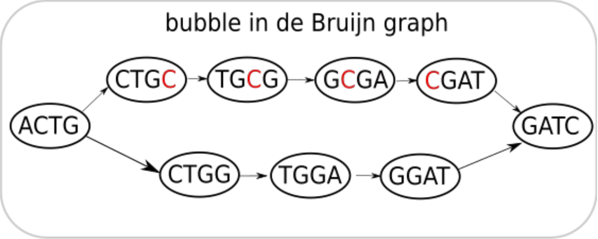
SNP

ACTGCGATC
ACTGTGATC

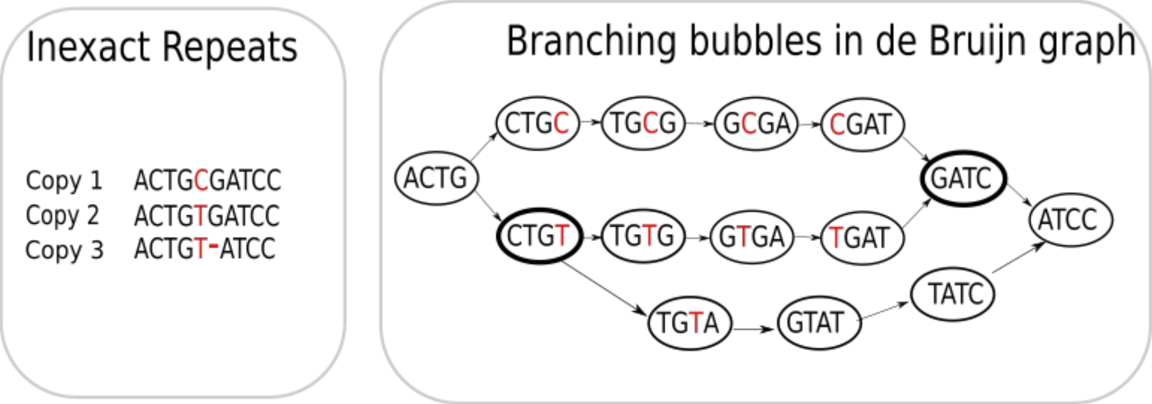

bubble in de Bruijn graph

Indel

ACTGCGATC
ACTG-GATC


bubble in de Bruijn graph

Inexact repeats generate branching bubbles in the DBG



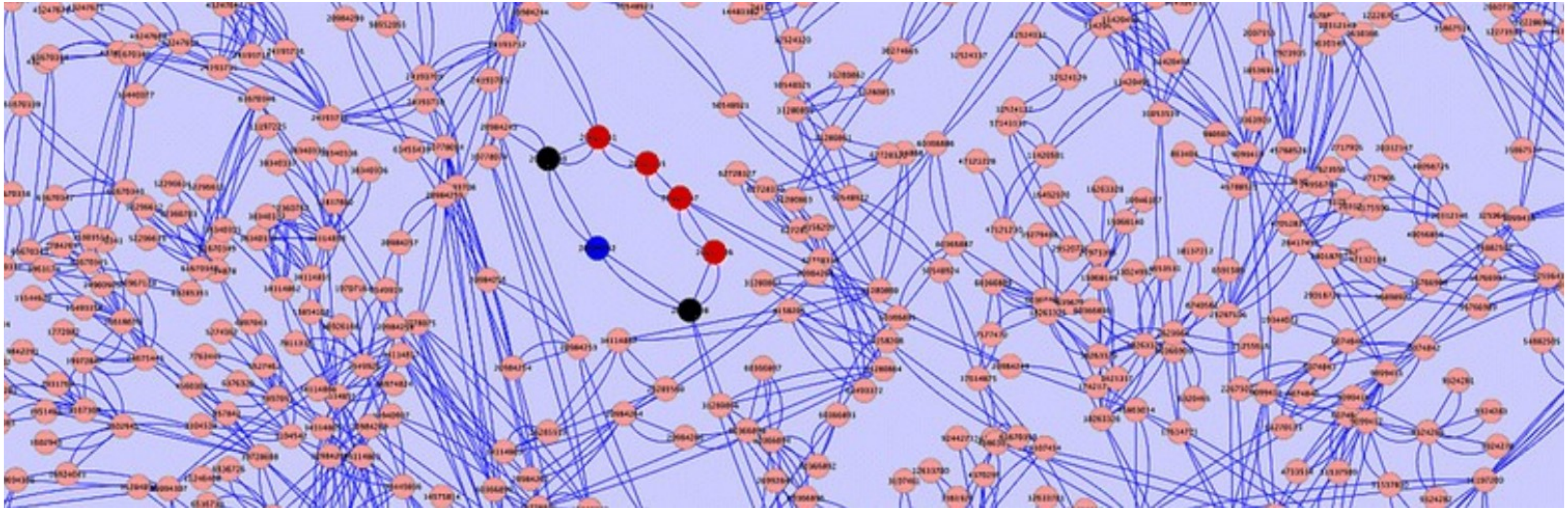
Inexact Repeats

Copy 1	ACTGCGATCC
Copy 2	ACTGTGATCC
Copy 3	ACTGT-ATCC

Branching bubbles in de Bruijn graph

Issue: Some repeats are present in very high copy number (even in transcriptomes) and generate very dense subgraphs, which is the main cause for the combinatorial explosion

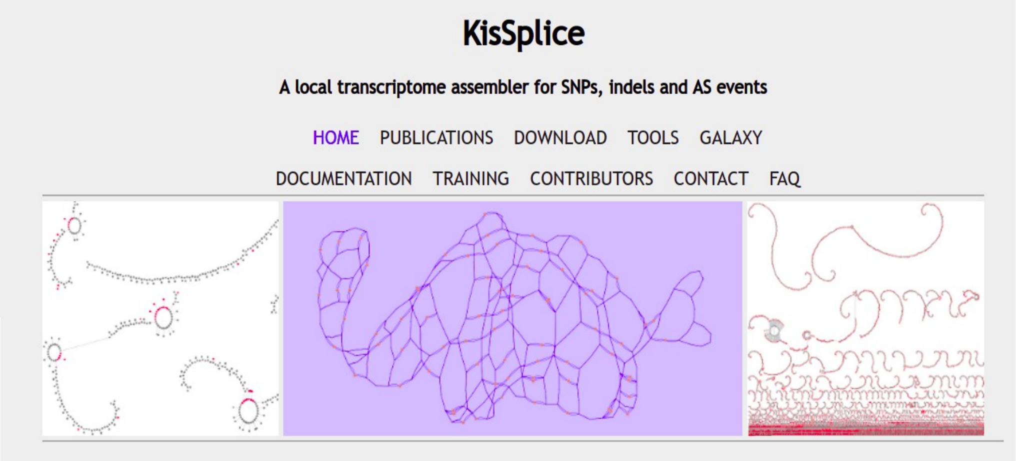
AS event flanked by repeats



SCN5A gene in patients with myotonic dystrophy
Trick: this bubble has less than 5 branching nodes

KisSplice pipeline

- Input: RNAseq data (.fastq)
- KisSplice :
  - Build DBG from RNAseq data
  - Enumerate all bubbles
  - Quantify bubbles



KisSplice

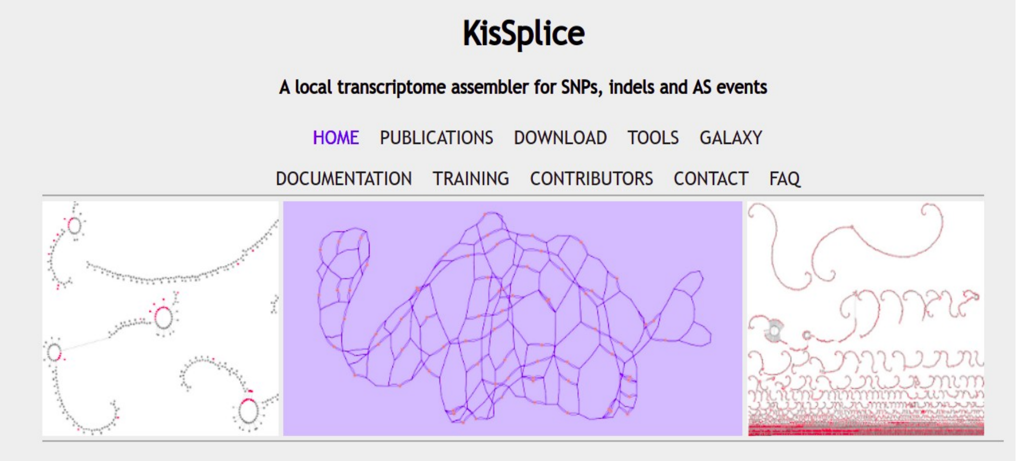
A local transcriptome assembler for SNPs, indels and AS events

HOME PUBLICATIONS DOWNLOAD TOOLS GALAXY
DOCUMENTATION TRAINING CONTRIBUTORS CONTACT FAQ

Sacomoto et al. BMC Bioinformatics 2012

KisSplice pipeline

- Input: RNAseq data (.fastq)
- KisSplice :
  - Build DBG from RNAseq data
  - Enumerate all bubbles
  - Quantify bubbles
- KissDE :
  - Differential analysis



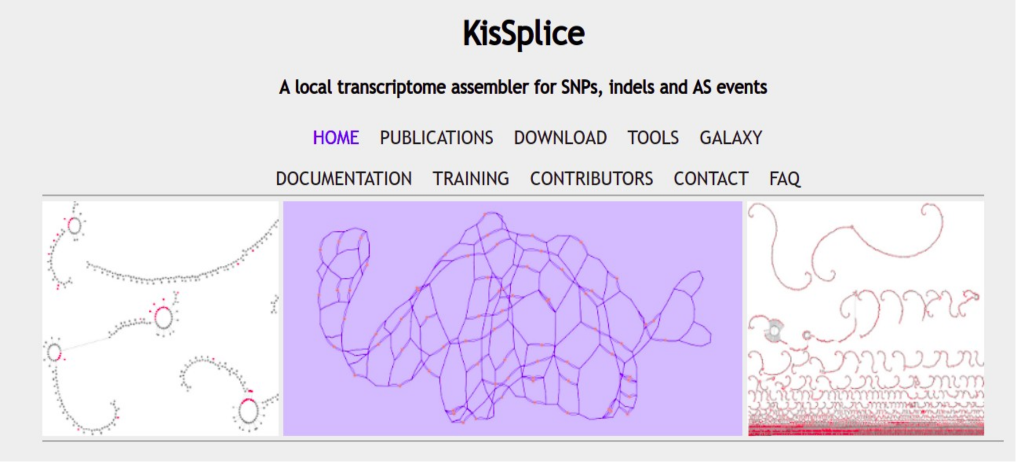
KisSplice

A local transcriptome assembler for SNPs, indels and AS events

Sacomoto et al. BMC Bioinformatics 2012
Lopez-Maestre et al. NAR 2016

KisSplice pipeline

- Input: RNAseq data (.fastq)
- KisSplice :
  - Build DBG from RNAseq data
  - Enumerate all bubbles
  - Quantify bubbles
- KissDE :
  - Differential analysis
- KisSplice2RefGenome
  - Annotate bubbles (if reference genome is available)
- Output: List of differentially spliced genes

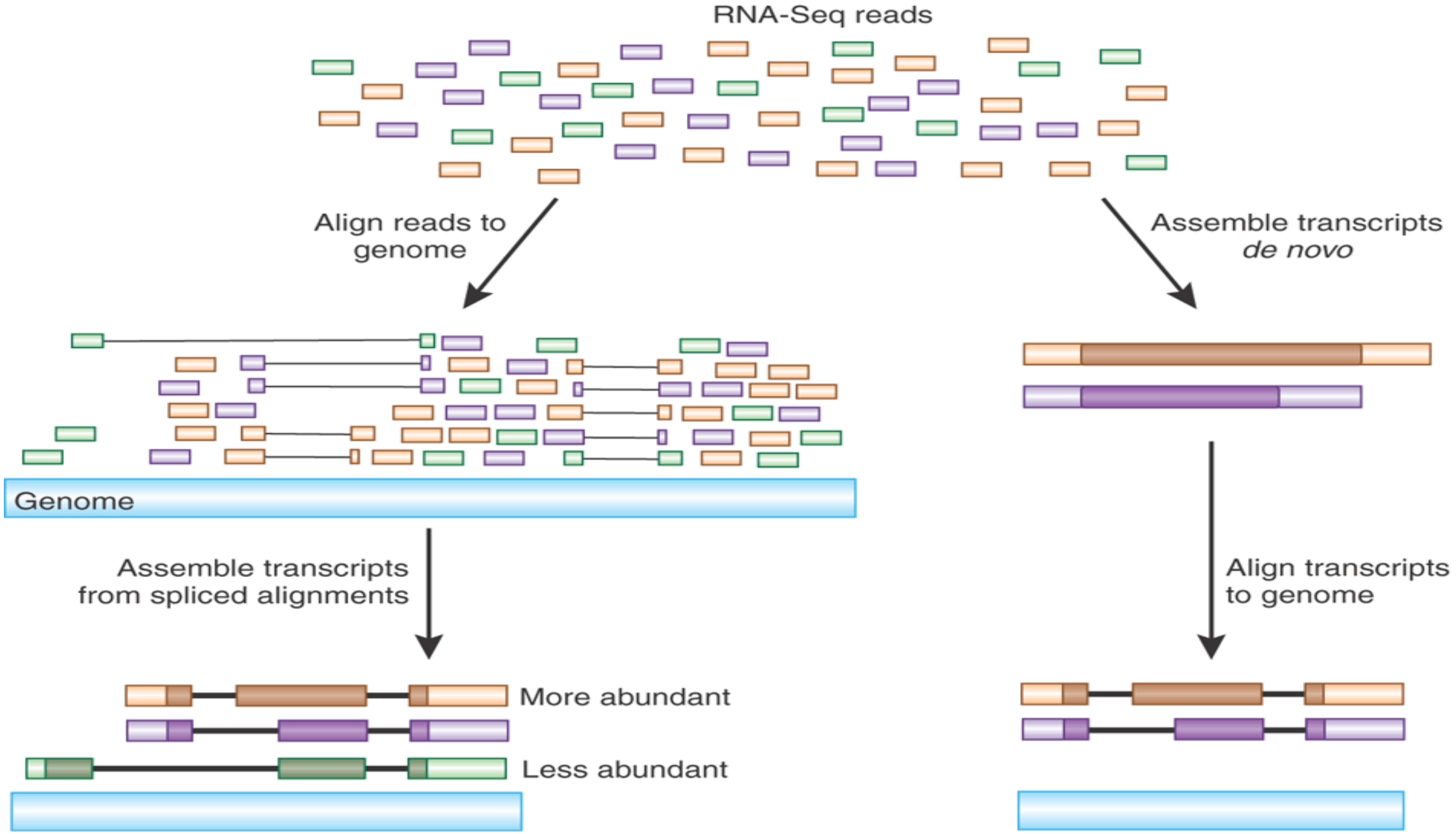


KisSplice

A local transcriptome assembler for SNPs, indels and AS events

Sacomoto et al. BMC Bioinformatics 2012
Lopez-Maestre et al. NAR 2016
Benoit-Pilven et al. Scientific Reports 2018

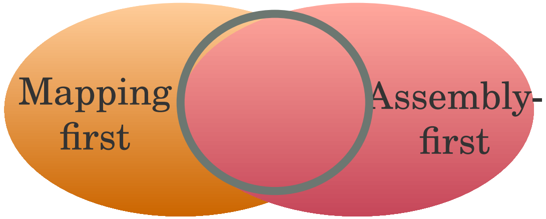
Two approaches to assemble transcripts


RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

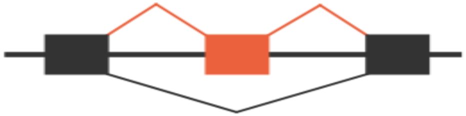
Haas and Zody, Nature Biotech 2010

What is the overlap between the predictions of the two approaches ?

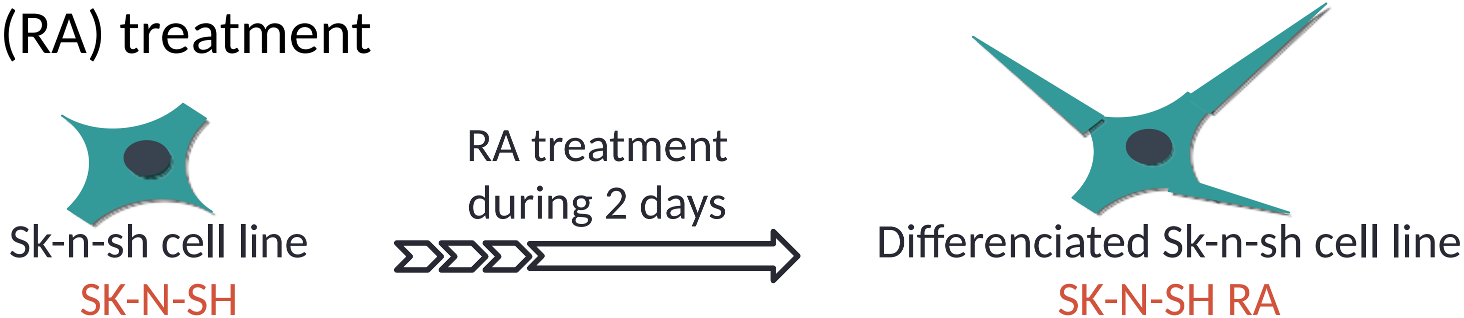
Identify pros and cons of assembly-first and mapping-first methods



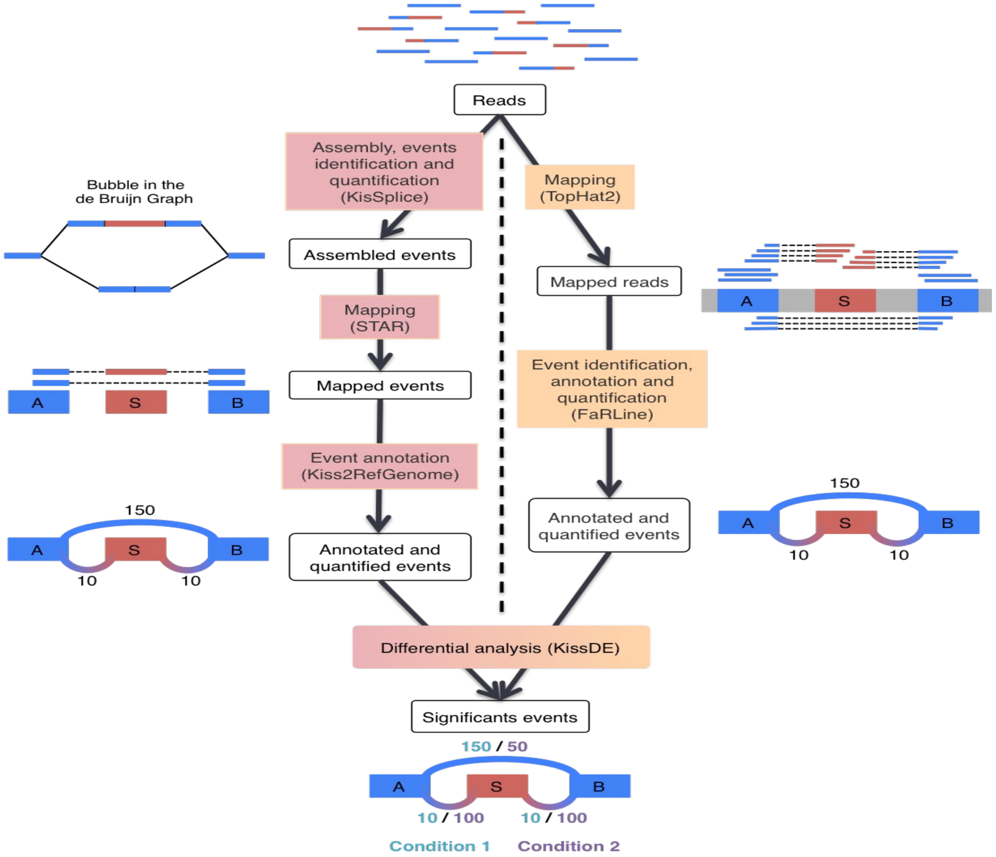
→ Comparison done on alternative skipped exon (ASE) events only



→ Public dataset (ENCODE) from neuroblastoma SK-N-SH cell line with or without retinoic acid (RA) treatment

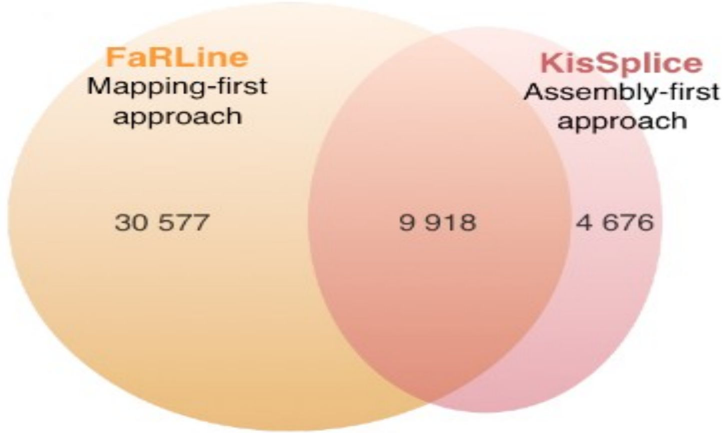


Compared pipelines

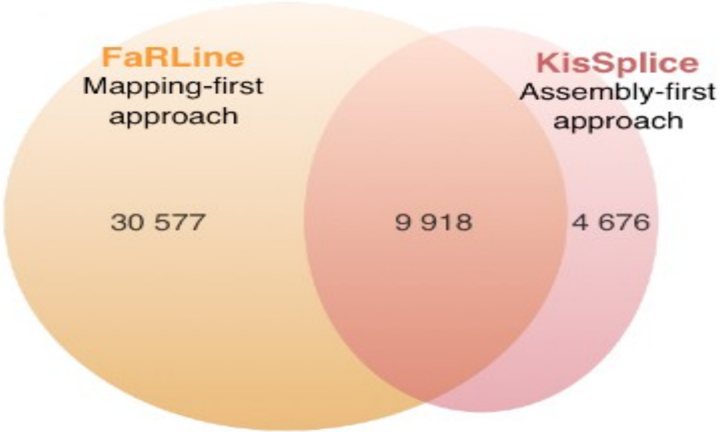

Reads
Assembly, events identification and quantification (KisSplice)
Mapping (TopHat2)
Bubble in the de Bruijn Graph
Assembled events
Mapped reads
A
S
B
Mapping (STAR)
Mapped events
A
S
B
Event identification, annotation and quantification (FaRLine)
Event annotation (Kiss2RefGenome)
150
A
S
B
10
10
Annotated and quantified events
Annotated and quantified events
150
A
S
B
10
10
Differential analysis (KissDE)
Significants events
150 / 50
A
S
B
10 / 100
10 / 100
Condition 1
Condition 2

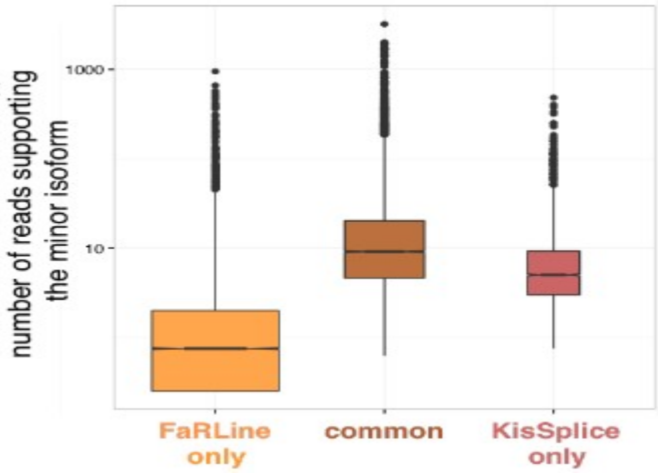
FaRLine developed in the group of Didier Auboeuf

Mapping-first approach finds many unfrequent variants

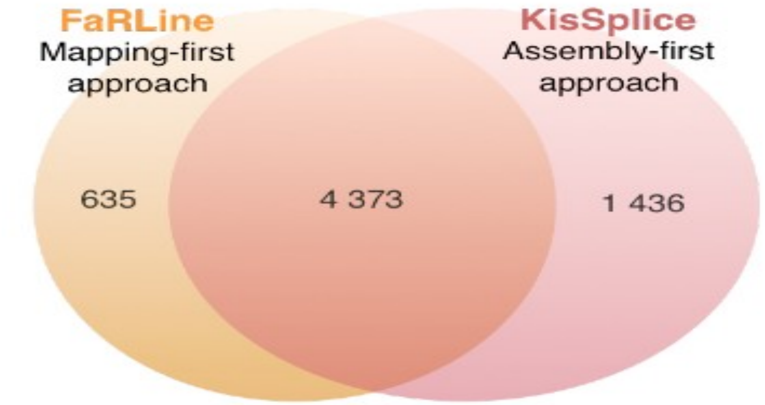


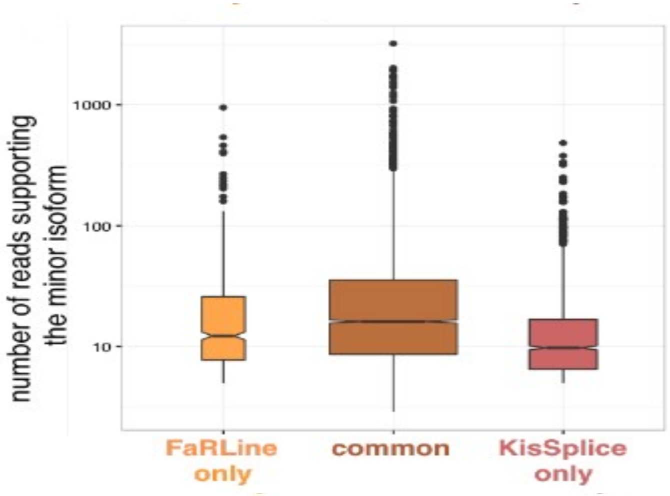
Mapping-first approach finds many unfrequent variants





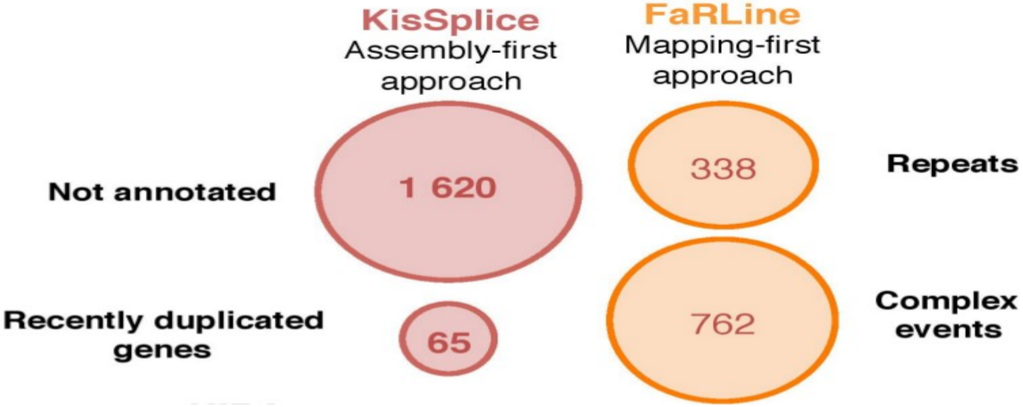
The overlap between methods increases when unfrequent variants are filtered out



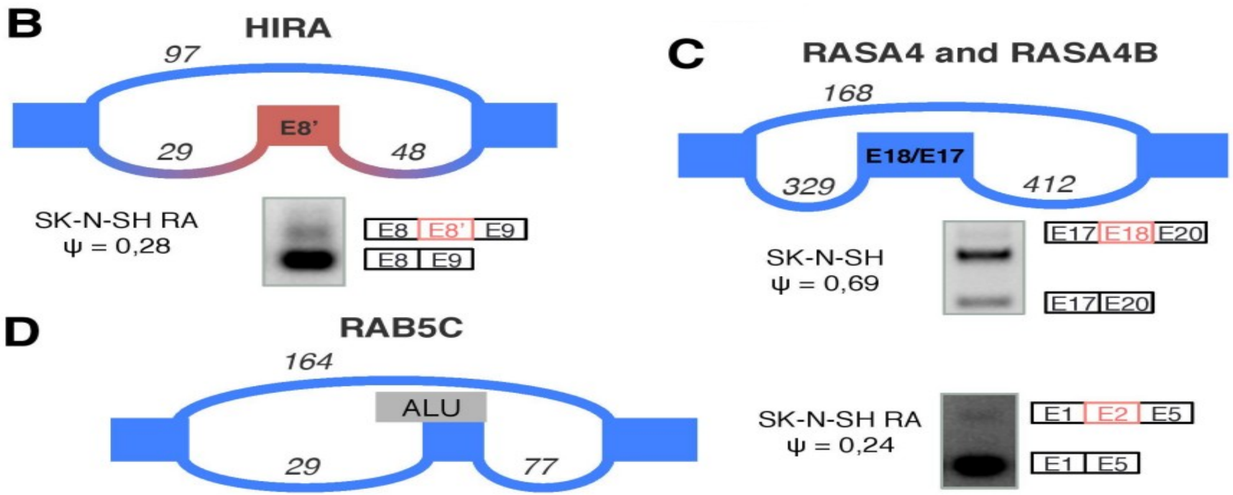


Unfrequent variant = less than 5 reads, or relative abundance < 10 %

Some abundant transcripts are systematically missed by one approach



Experimental Validations

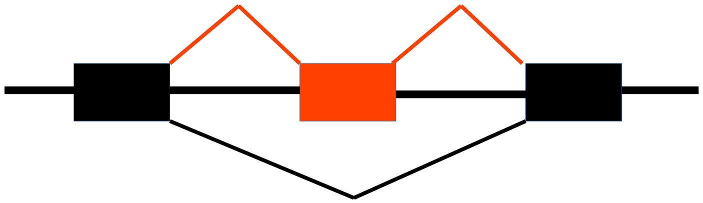


Annotation summary

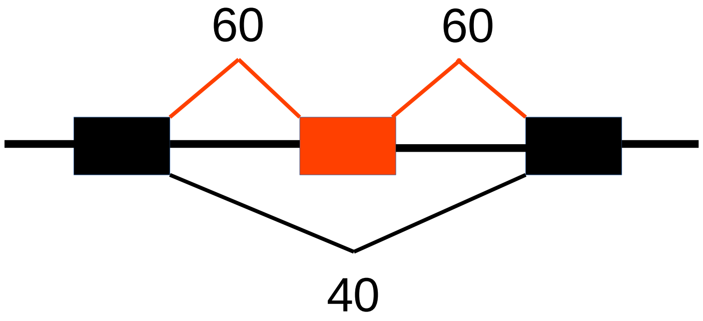
Mapping-first is stronger for rare variants and exonised Alus
Assembly-first is stronger for novel variants and recent paralogs

Should I care about these differences ?
Does it have an impact on my differential analysis ?

Magnitude of the effect

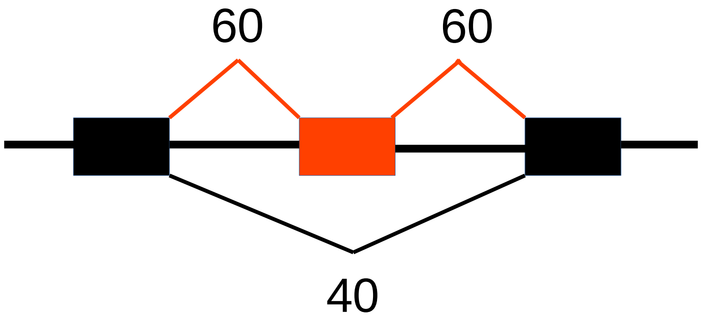


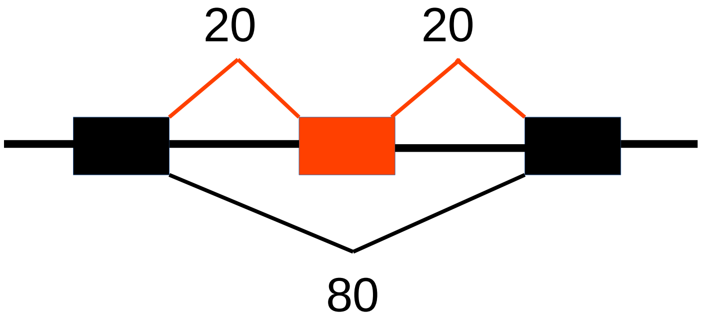
Magnitude of the effect



Percent Spliced In (PSI) = 60 / (60 + 40) = 60%
The major isoform is the inclusion isoform, the exon is included in 60% of cases

Magnitude of the effect





Condition 1: PSI1 = 60%

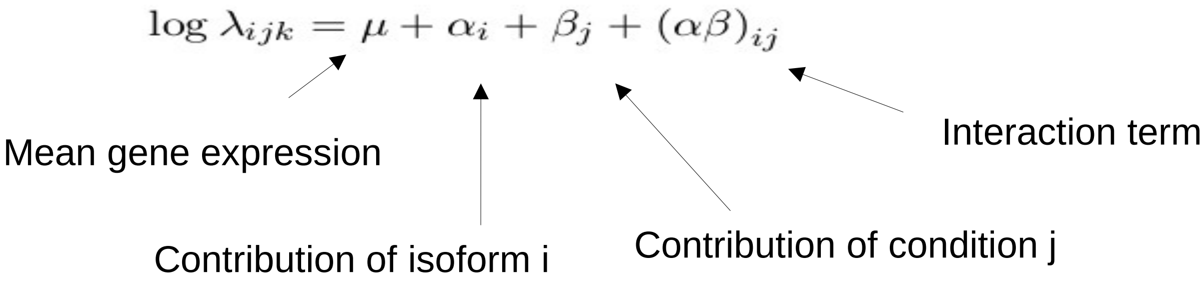
Condition 2: PSI2=20%

DeltaPSI = PSI1 – PSI2 = 60-20 = 40%

The inclusion level of the exon decreased by 40%

Statistical Analysis

- Count regression with negative binomial distribution
- Generalised linear model, 2 way design, with interaction



$$\log \lambda_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

Mean gene expression

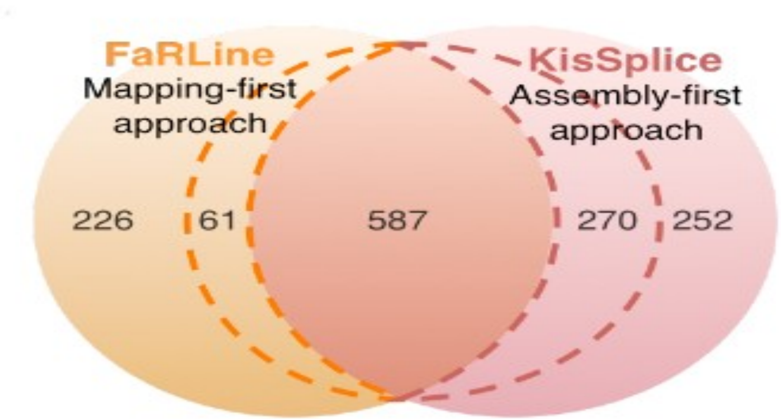
Contribution of isoform i

Contribution of condition j

Interaction term

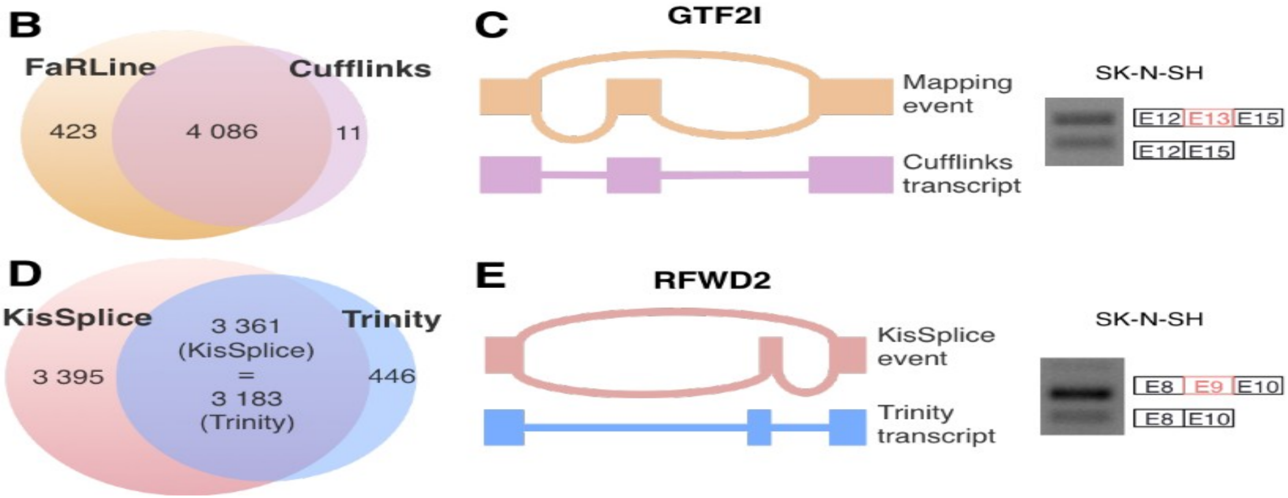
- Target hypothesis: $H_0 : \{(\alpha\beta)_{ij} = 0\}$
- Likelihood ratio test
- P-values adjusted with benjamini-hochberg procedure

Comparison after differential analysis



AS events found by one method and not the other can be significant
|DeltaPSI| > 10%, FDR<0.05

Comparison to global methods

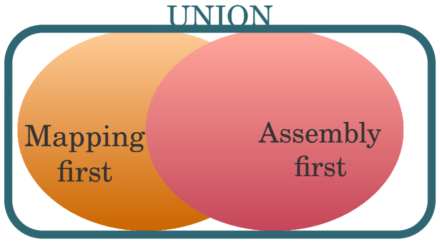


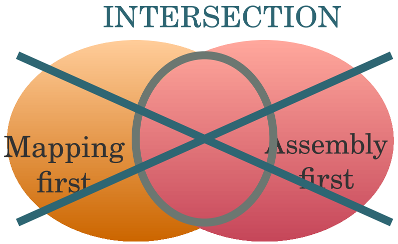
Methods Summary

Annotating alternative splicing with a single approach leads to **missing a large number of candidates**.

These candidates should not be neglected, since many of them are **differentially regulated** across conditions.

We advocate for the use of a combination of both mapping-first and assembly-first approaches for annotation and differential analysis of alternative splicing from RNA-seq data.





Benoit-Pilven et al. Scientific Reports 2018 – kissplice.prabi.fr/pipeline_ks_farline

Two applicative case studies

- Application to a spliceosomopathy
(collaboration with the group of Patrick Edery & Sylvie Mazoyer, HCL)

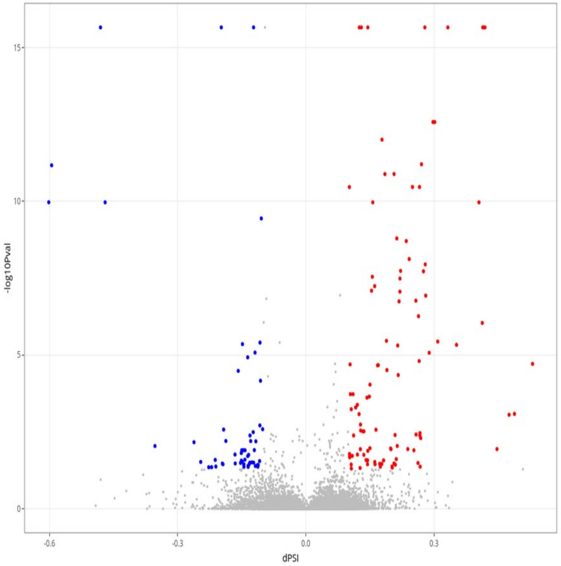
Cologne et al. RNA 2019

Two applicative case studies

- Application to a spliceosomopathy
  (collaboration with the group of Patrick Edery & Sylvie Mazoyer, HCL)
- Application to Influenza A virus infection
  (collaboration with the group of Nadia Naffakh at Institut Pasteur)

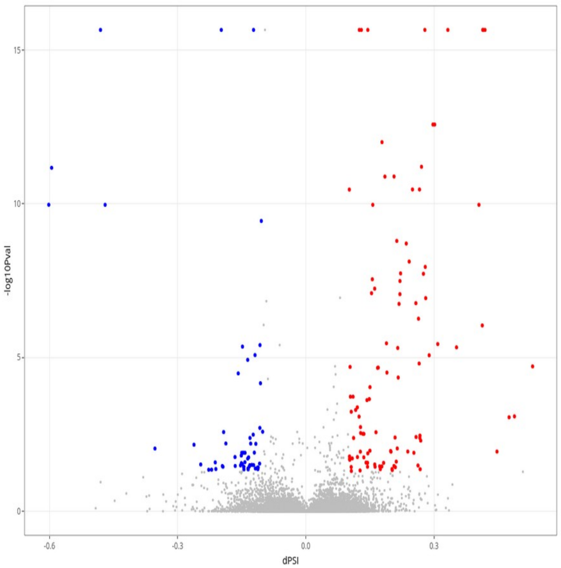
Cologne et al. RNA 2019
Ashraf et al. NAR Genomics & Bionformatics 2020

Volcano Plots

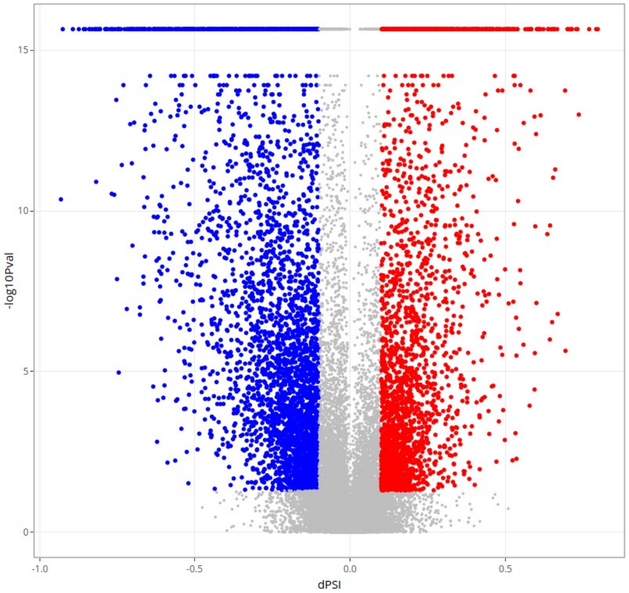


Spliceosomopathy
(TALS patients
fibroblasts)

Volcano Plots

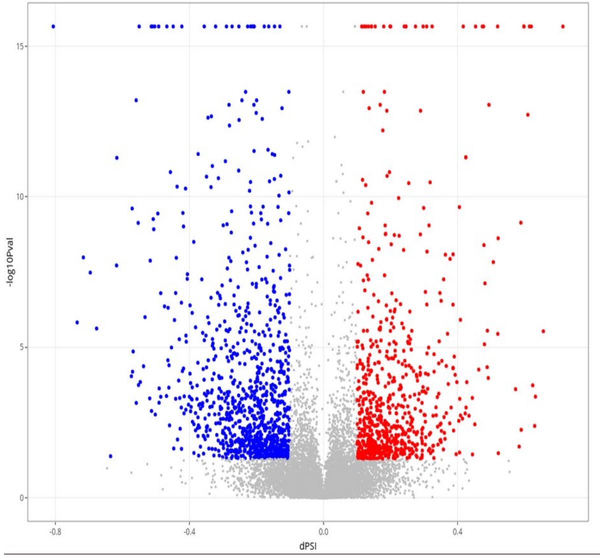


Spliceosomopathy
(TALS patients
fibroblasts)

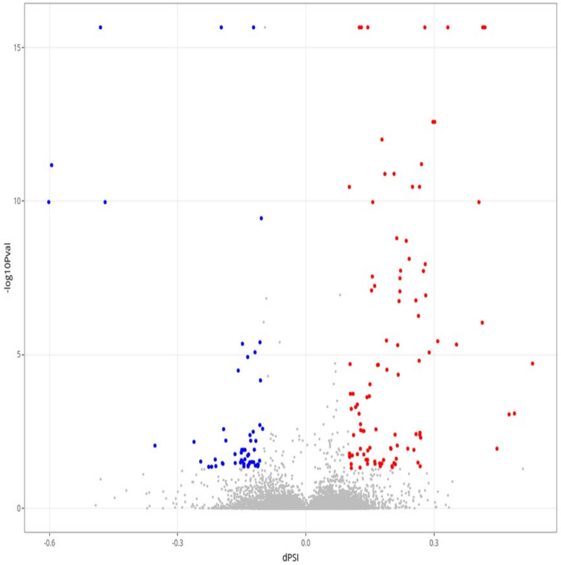


IAV infection
(A549 cells)

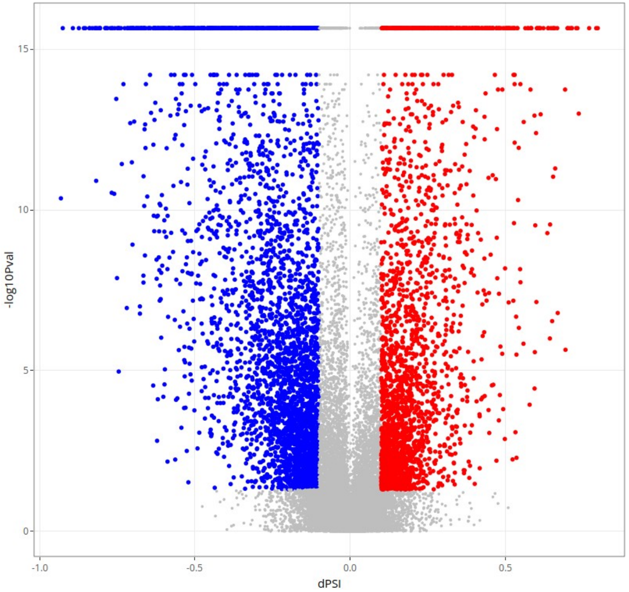
Volcano Plots



Cellular differentiation
(SKNSH cells + RA)



Spliceosomopathy
(TALS patients
fibroblasts)



IAV infection
(A549 cells)