Analyse de données de séquençage haut débit

Vincent Lacroix
Laboratoire de Biométrie et Biologie Évolutive
INRIA ERABLE

- Sequencing is cheap
- Applications include
  - De novo genome assembly
  - Resequencing (SNPs, indels, rearrangements)
  - Transcriptome analysis (RNA-seq)
  - Protein-DNA interactions (Chip-seq)
  - Metagenomics
- The difficulty is not to generate data, but to analyse it
  - Many datasets are superficially analysed, or not analysed at all
  - Bioinformatics bottleneck

- Bioinformatics analysis of sequencing data starts by either
  - Mapping your reads to a reference genome
  - De novo assembling your reads
- Downstream analysis depends on the application (RNAseq, ChipSeq, etc)
- For each application, there is not yet a unique reference pipeline
- Choosing which pipeline to use requires to understand what it captures/misses

- Bioinformatics analysis of sequencing data starts by either
  - Mapping your reads to a reference genome
  - De novo assembling your reads
- Downstream analysis depends on the application (RNAseq, ChipSeq, etc)
- For each application, there is not yet a unique reference pipeline
- Choosing which pipeline to use requires to understand what are its limitation

aviesan
alliance nationale
pour les sciences de la vie et de la santé

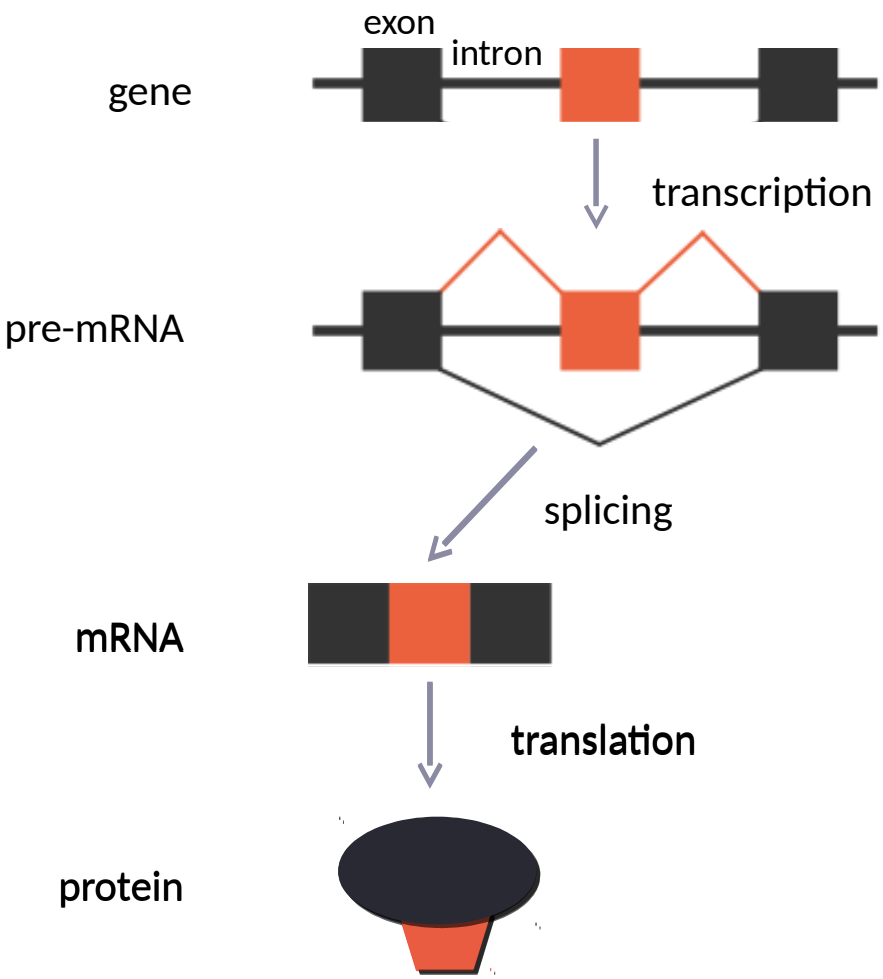
Institut Thématique Multi-Organismes
Technologies pour la santé

Complementarity of assembly-first and mapping-first approaches for alternative splicing annotation and differential analysis from RNAseq data

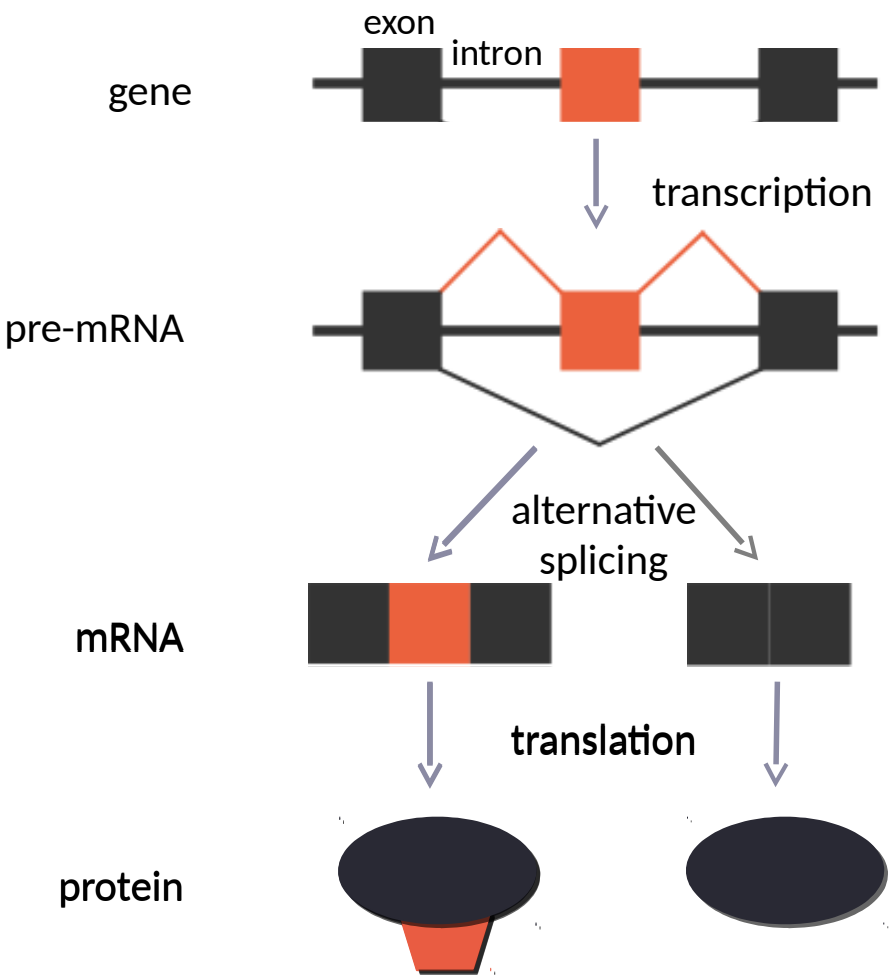
Clara Benoit-Pilven, Camille Marchet, Emilie Chautard, Leandro Lima, Marie-Pierre Lambert, Gustavo Sacomoto, Amandine Rey, Audric Cologne, Sophie Terrone, Louis Dulaurier, Jean-Baptiste Claude, Cyril Bourgeois, Didier Auboeuf, Vincent Lacroix

RhôneAlpes Région
arc COMMUNAUTÉS DE RECHERCHE ACADÉMIQUE RhôneAlpes SANTÉ
Plan cancer 2009 2013
Instituts thématiques Inserm Institut national de la santé et de la recherche médicale
cnrs
Inria INVENTEURS DU MONDE NUMÉRIQUE
UCBL Lyon 1
LBBE BIOMETRIE ET BIOLOGIE EVOLUTIVE
LBMC
ENS DE LYON

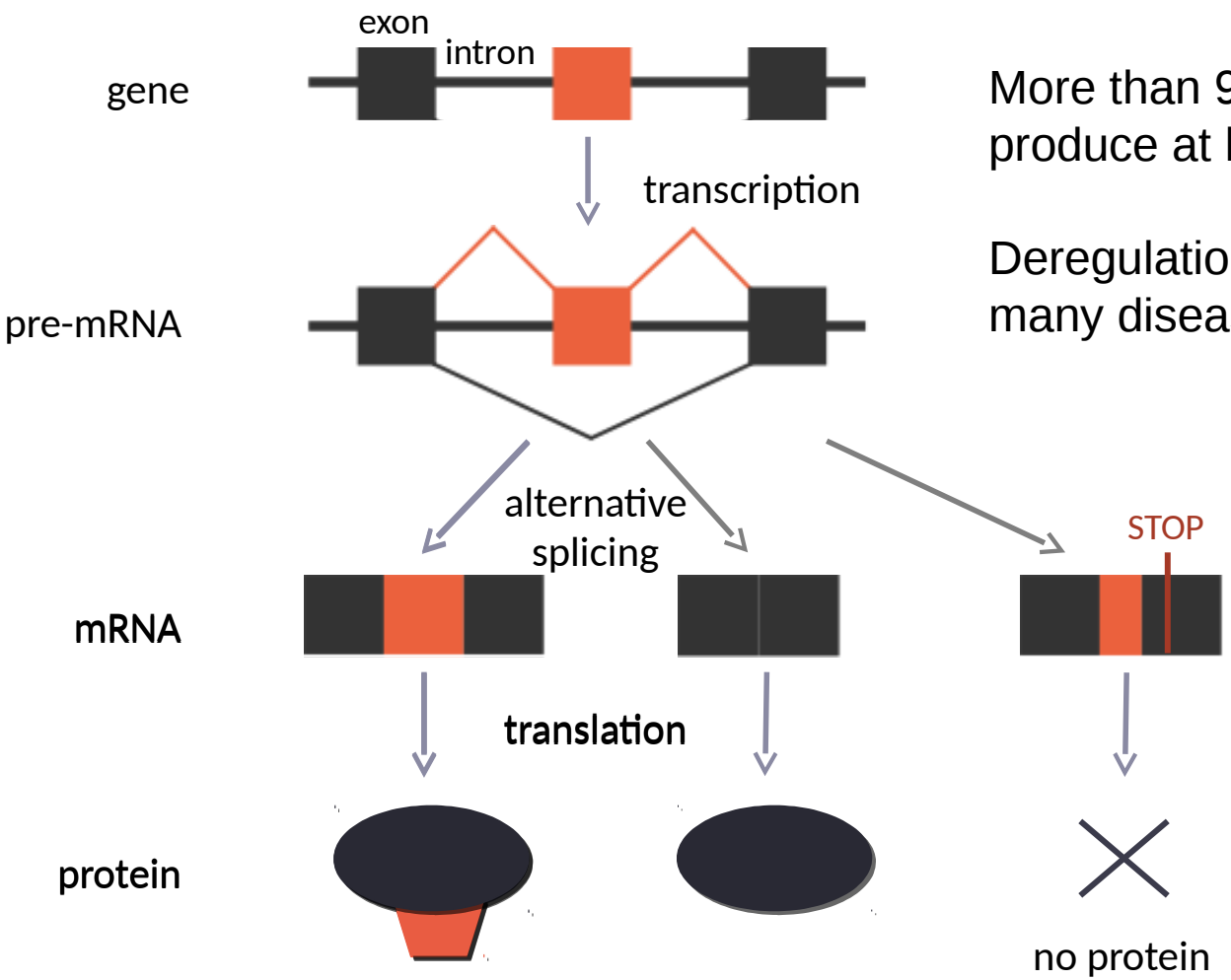
Alternative Splicing



Alternative Splicing



Alternative Splicing

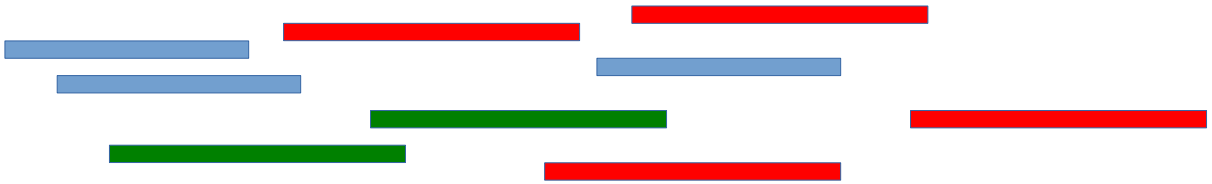


More than 90 % of multi-exon genes produce at least 2 isoforms

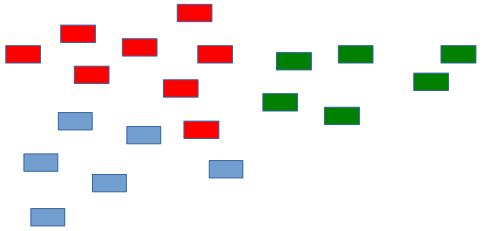
Deregulation of alternative splicing in many diseases (rare diseases & cancer)

RNAseq data

mRNAs
(~1000nt)

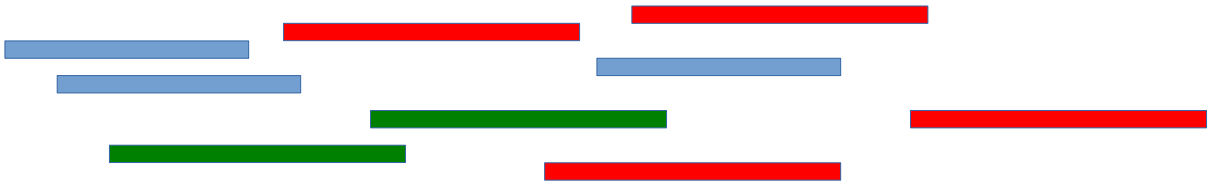


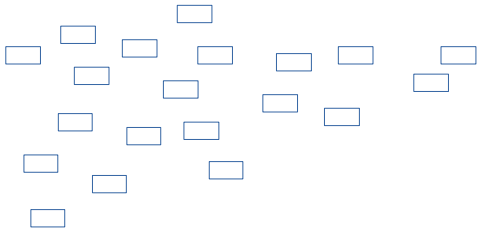
Reads
(100nt)



RNAseq data

mRNAs
(~1000nt)





Reads
(100nt)

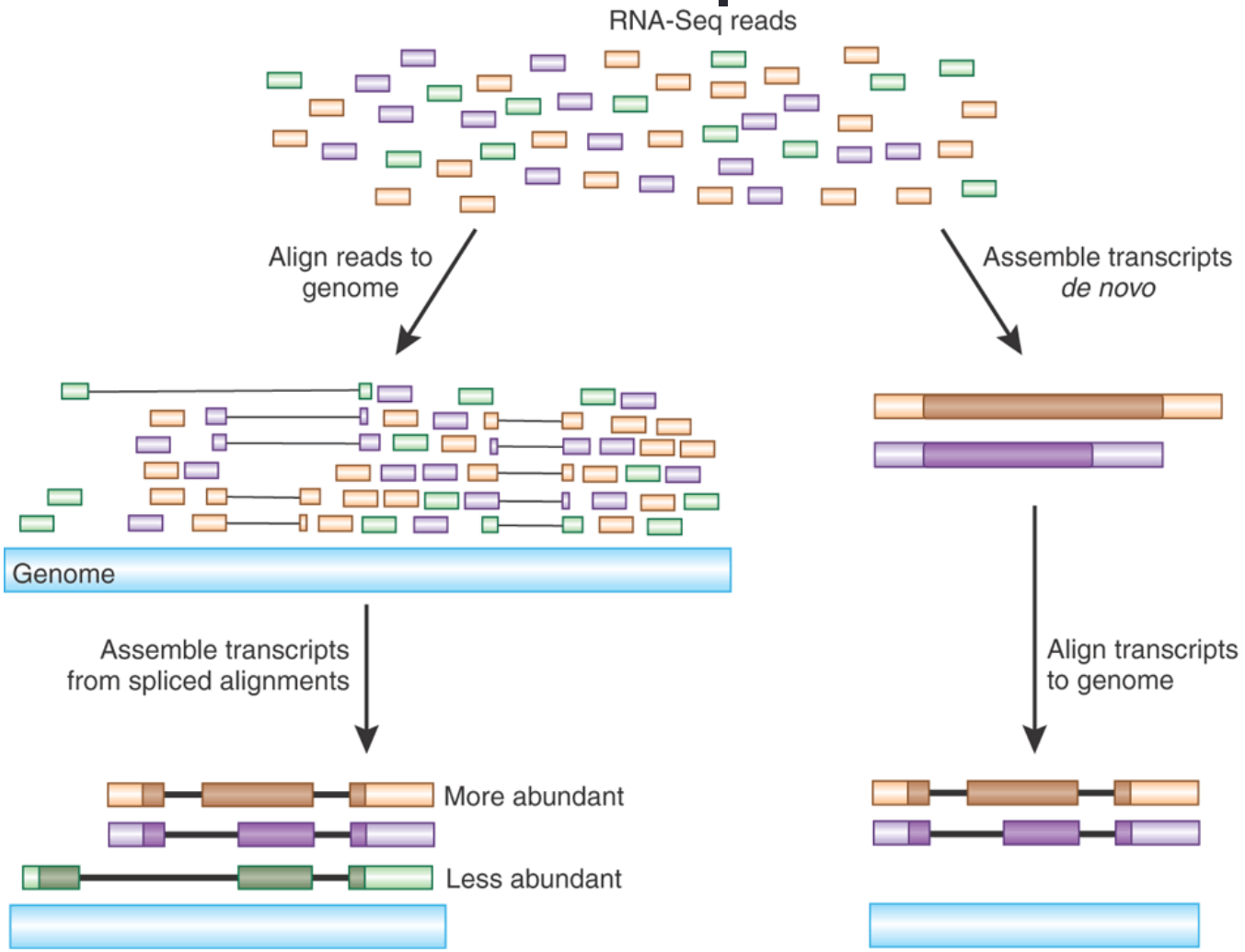
Alternative Splicing and RNA-seq data

- Gencode Annotations : 60 000 genes, 3 transcripts per gene.
- Assessing which gene/transcript is expressed in which tissue/condition can in principle be done through RNAseq
- The main challenges are:
  - Reads are short (100nt) and can be assigned to multiple transcripts (1000nt)
  - Some transcripts are annotated, some are novel
  - Some transcripts are highly expressed, many are poorly expressed

Annotation and Differential Analysis

- Annotation: Identify and quantify all transcripts present in a given condition
- Differential Analysis : Assess which genes are differentially spliced across conditions (treatment / control, population 1 / population 2, disease / control)

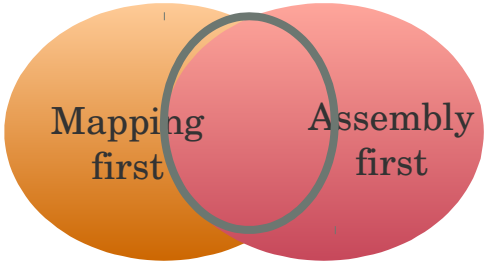
Two approaches to assemble transcripts



Haas and Zody, Nature Biotech

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

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

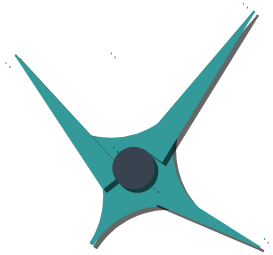


→ Comparison done on alternative skipped exon (ASE) events only

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

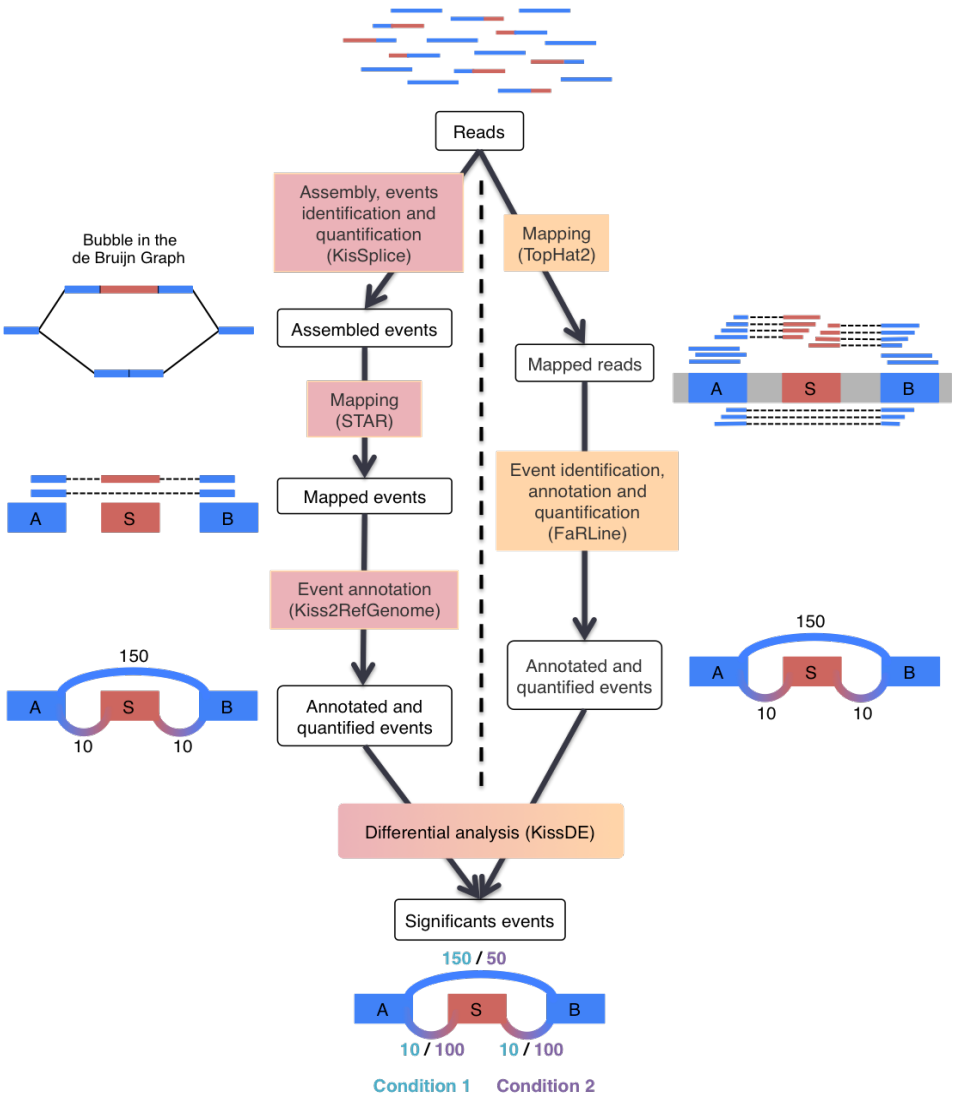
Sk-n-sh cell line
SK-N-SH

RA treatment during 2 days

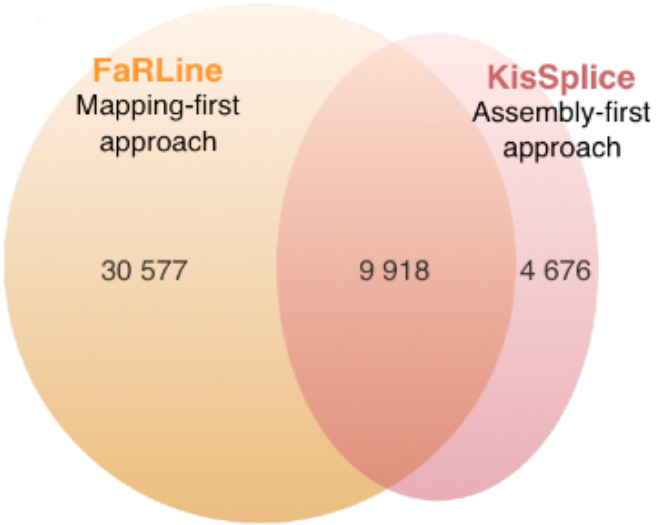


Differenciated Sk-n-sh cell line

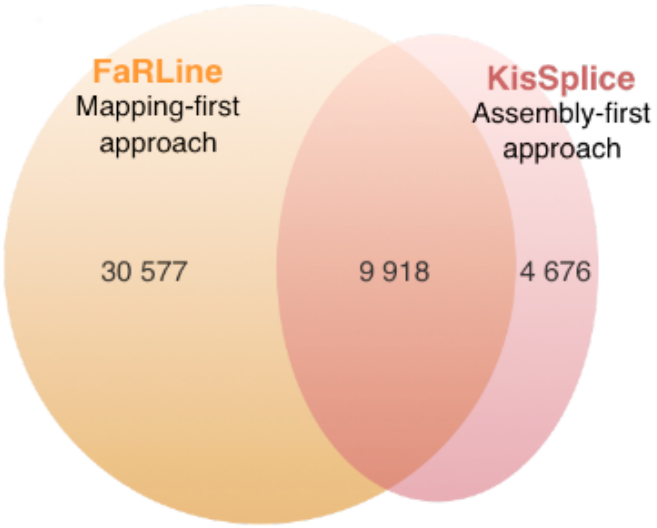
Compared pipelines

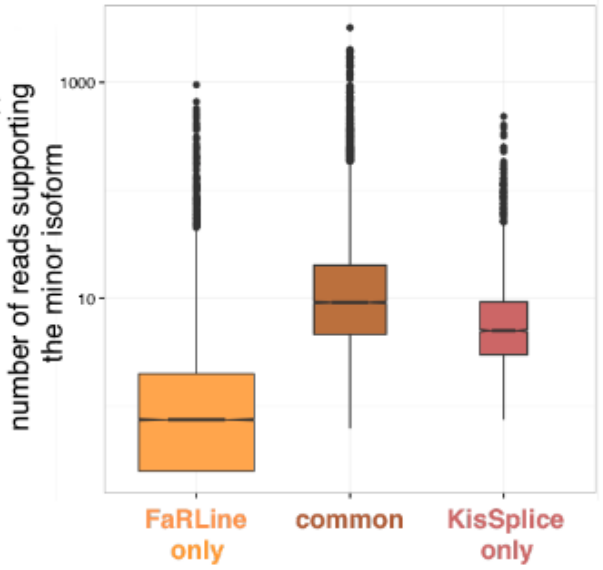


Mapping-first approach finds many unfrequent variants

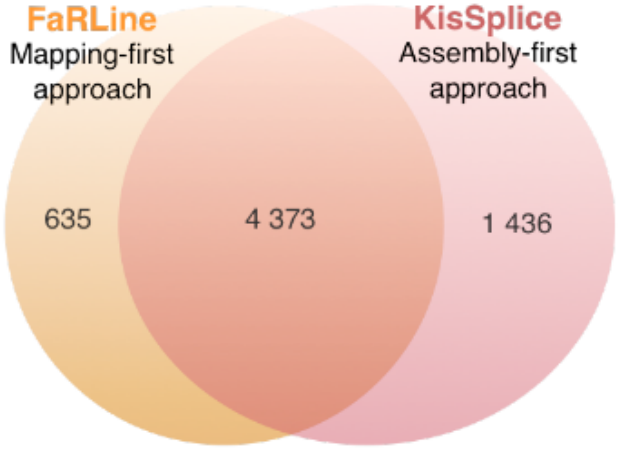


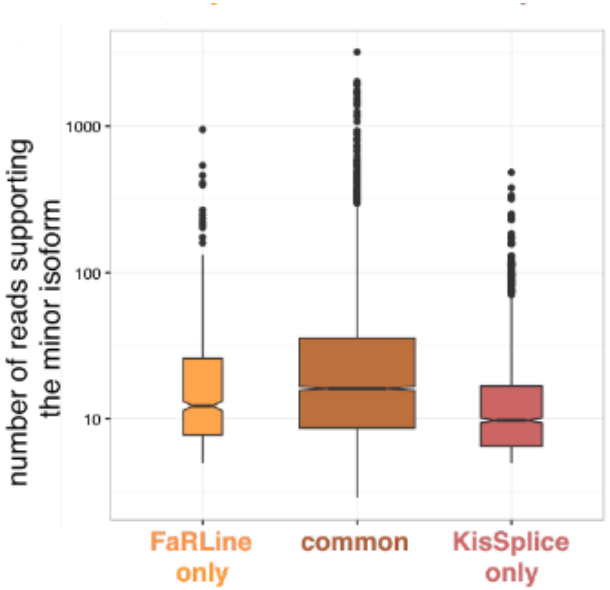
Mapping-first approach finds many unfrequent variants





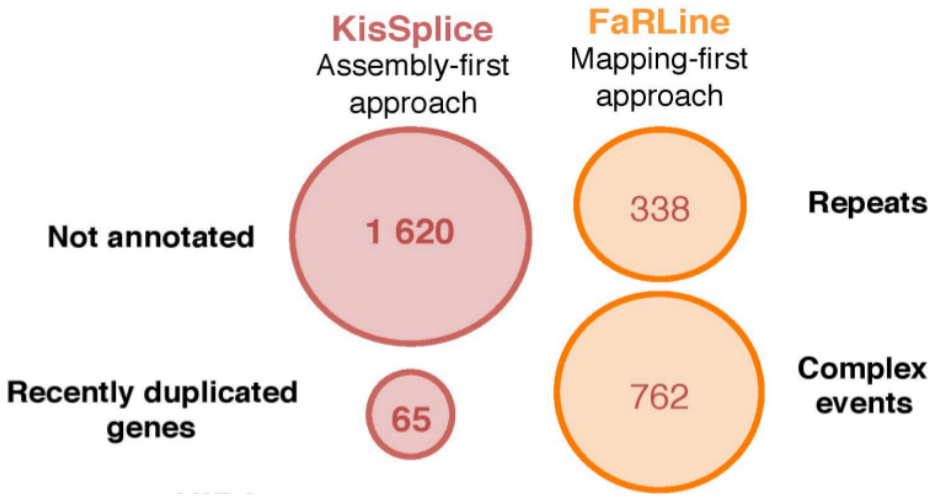
The overlap between methods increases when unfrequent variants are filtered out



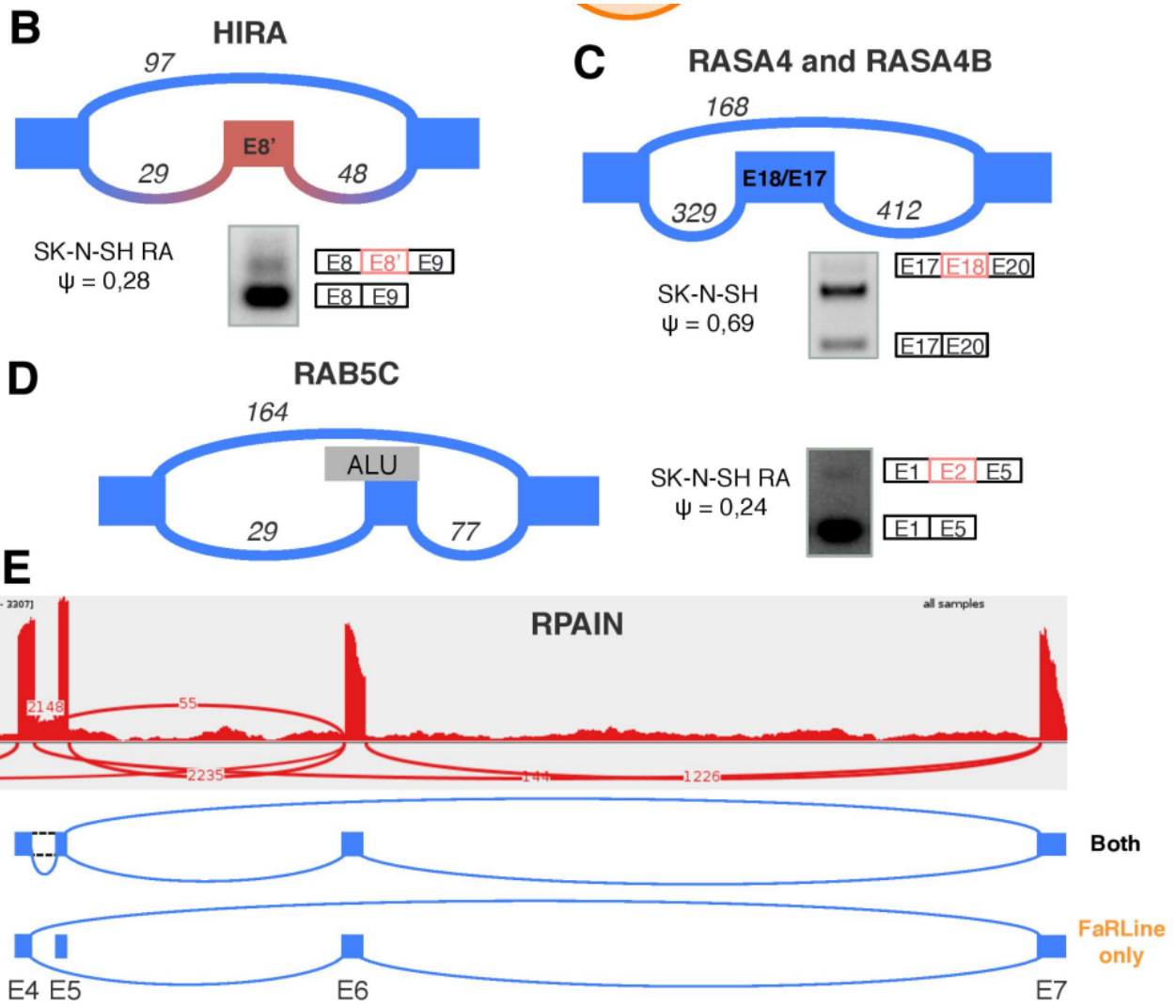


Unfrequent variant = less than 5 reads, or relative abundance < 10 %

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 ?

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

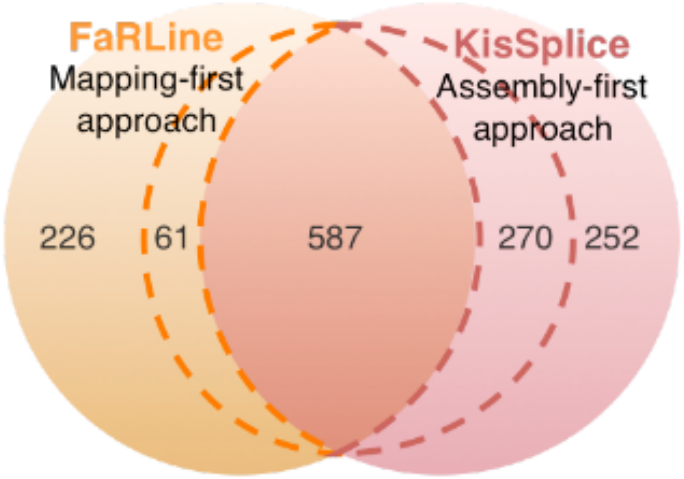
Contribution of isoform i

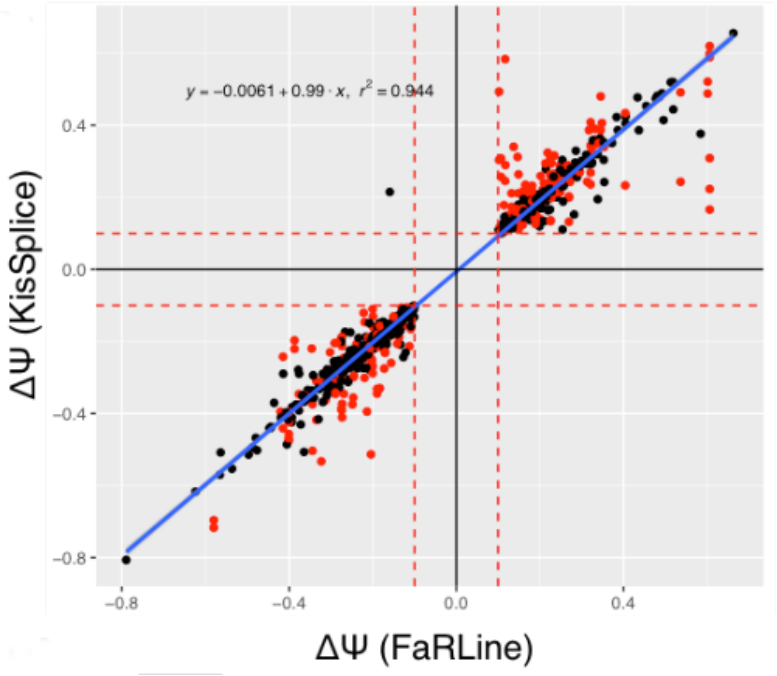
Contribution of condition j

Interaction term

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

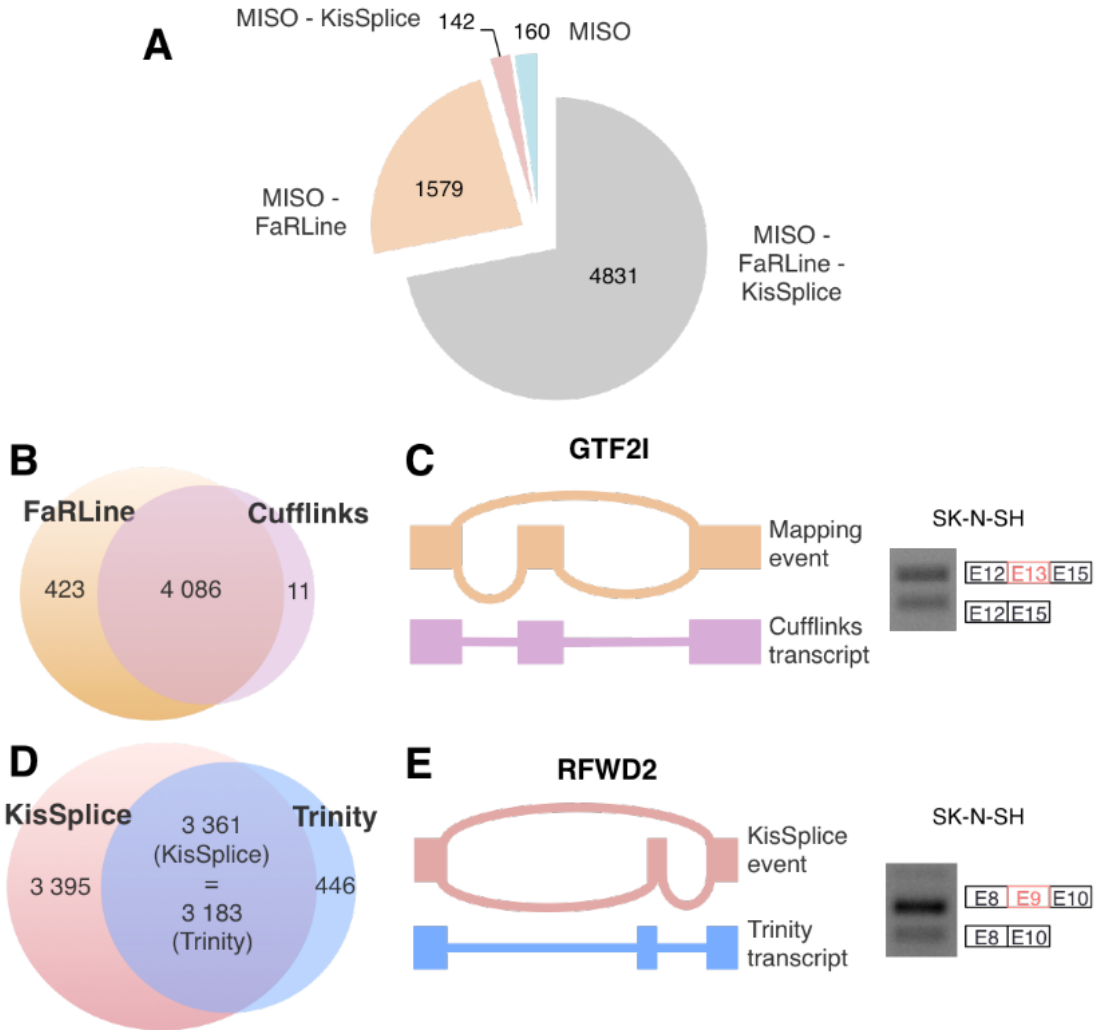
Comparison after differential analysis





AS events predicted by both pipelines have some quantification differences, especially for complex events (red dots)

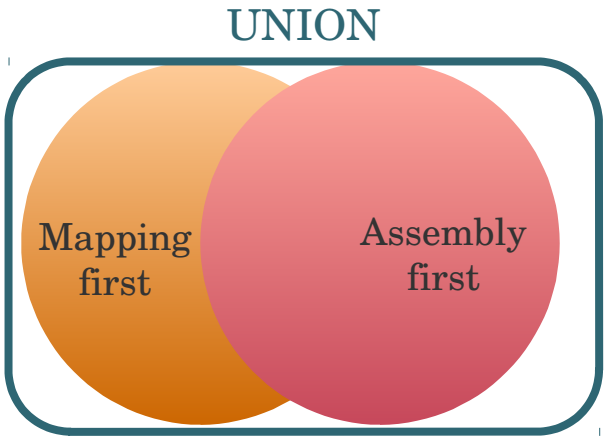
Comparison to other methods

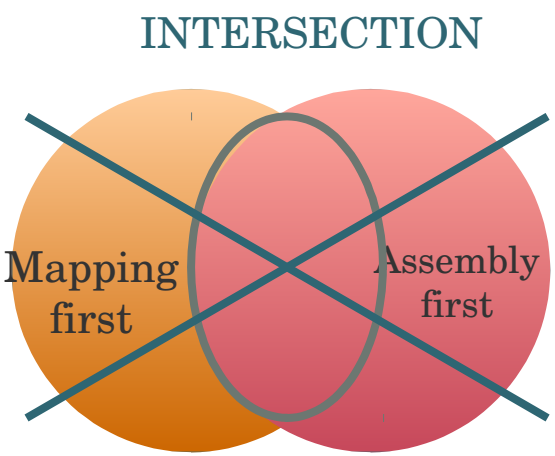


Conclusion & Perspectives

Annotating alternative splicing with a single approach leads to **missing a large number of candidates**.
These candidates cannot 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.





Software availability

- http://kissplice.prabi.fr

KisSplice

A local transcriptome assembler for SNPs, indels and AS events

HOME PUBLICATIONS DOWNLOAD TOOLS GALAXY

DOCUMENTATION TRAINING CONTRIBUTORS CONTACT FAQ

Latest News

- 2017-05-12: kissDE version 1.5.0 Release
- 2017-02-28: Our AMB paper is out.
- 2016-07-31: Our NAR paper is out. Full protocol to reproduce the results is available here

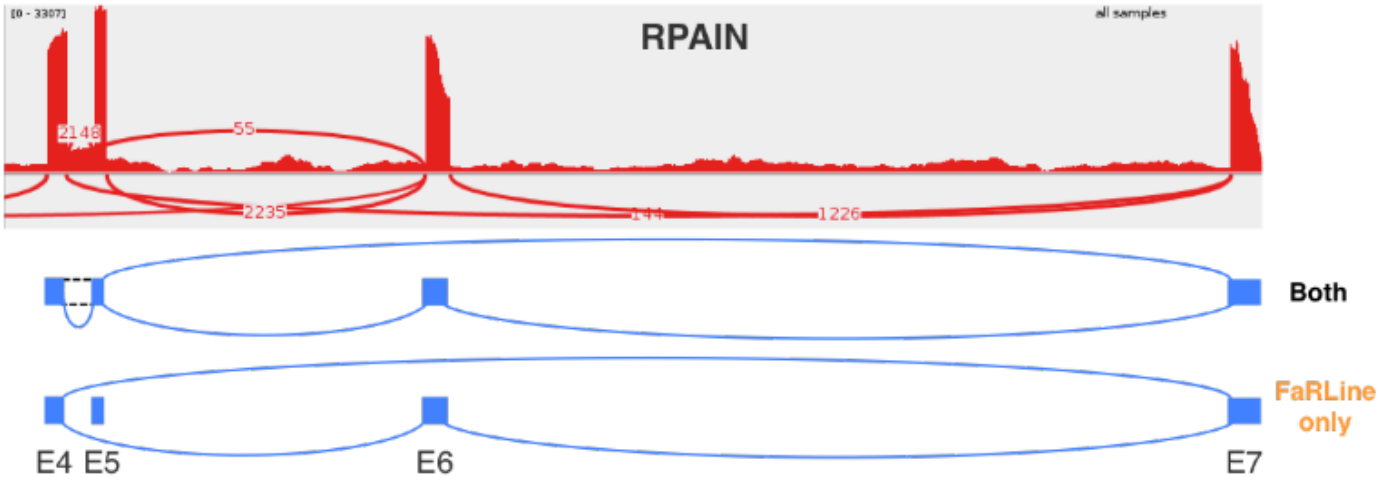
Recent paralogs

Missed by FaRLine
RASA4 and RASA4B are recent paralogs
Multi-mapping reads are discarded by mapping-first approaches
KisSplice co-assembles the two paralogs, and states that they collectively produce two transcripts
Confirmed experimentally by RT-PCR

Exons overlapping repeats

Missed by KisSplice
RAB5C contains an exonised Alu
Since this exon is annotated, FaRLine finds it
KisSplice fails to assemble it, because the bubble has more than 5 branches
(i.e. too many Alu copies in the dataset)
Confirmed experimentally by RT-PCR

Complex events



Missed by KisSplice
The skipping of E6 with E4 and E7 as flanking exons is reported only by FaRLine
KisSplice discards E4-E6 junction because it is supported only by 55 reads, which is less then 2 % of the read flow leaving E4