Identification and quantification of isoforms in RNAseq data : deep short reads Vs shallow long reads

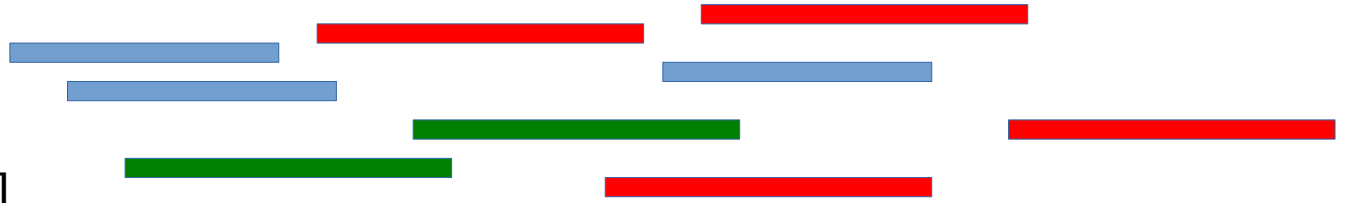
Vincent Lacroix
Laboratoire de Biométrie et Biologie Évolutive
INRIA ERABLE

What do we do in Lyon

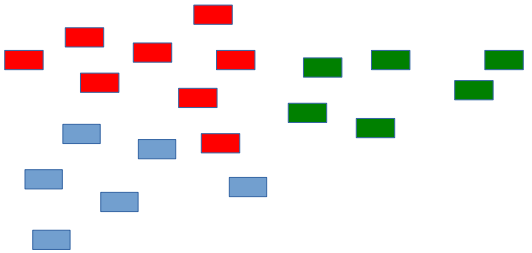
- We are interested in **developing** bioinformatics methods to study alternative splicing
- KisSplice assembles AS events from short RNAseq reads efficiently. It is based on principled models and efficient data structures.
- It is available, maintained and used : www.kissplice.prabi.fr
- Question : when/how to move to long reads ?

RNAseq with Illumina

mRNAs
[500-5000nt]

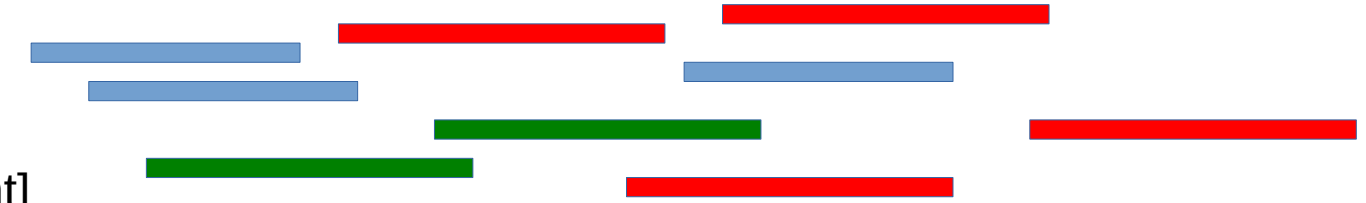


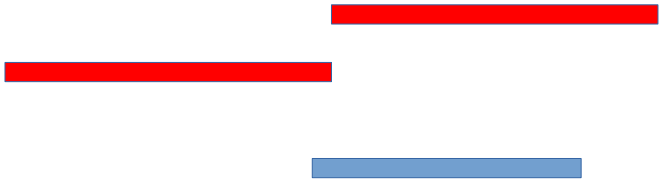
Reads
Length : 100nt
Number : 100M
Error : 0.5 %



RNAseq with Nanopore

mRNAs
[500-5000nt]





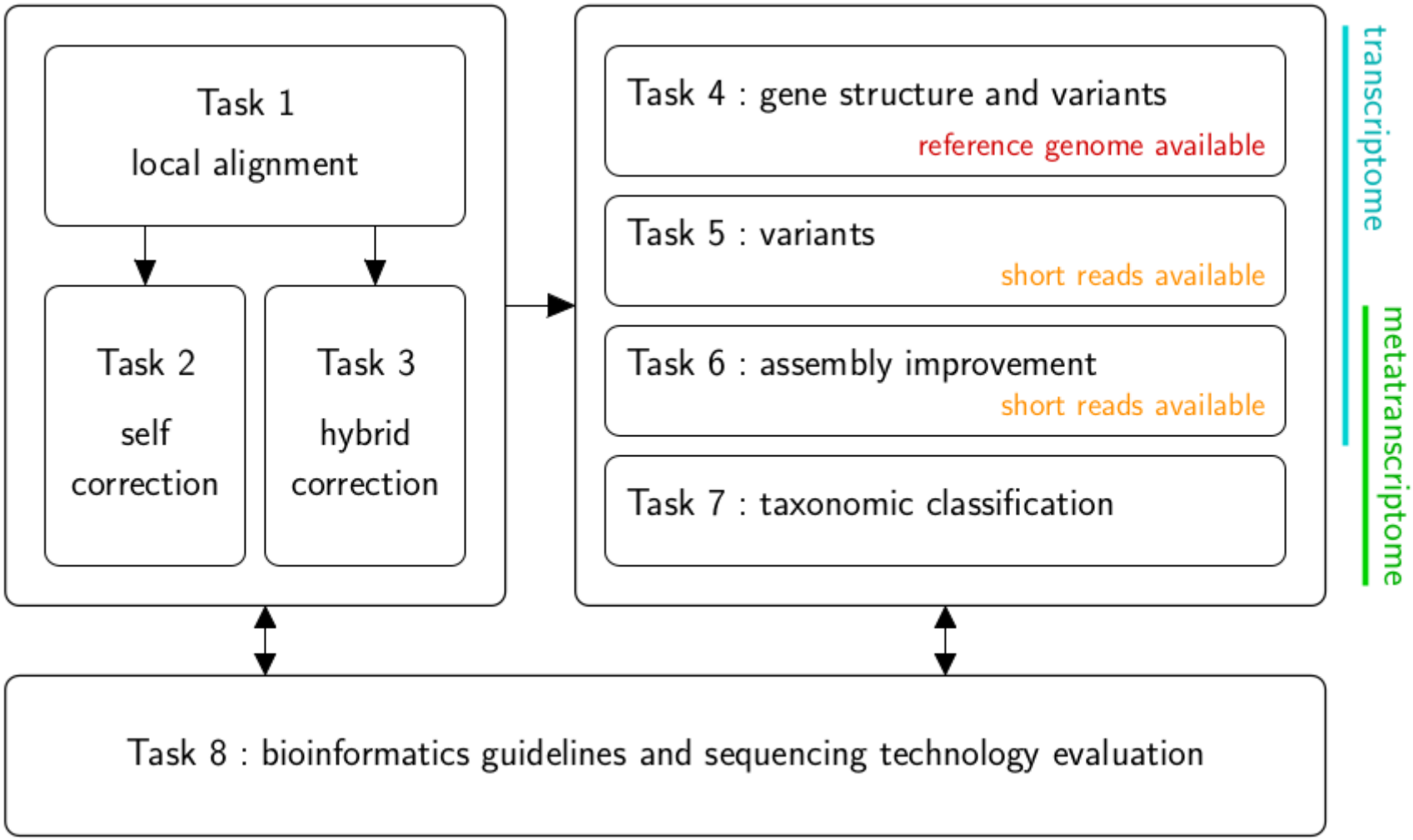
Reads
Length : 1000nt
Number : 1M
Error : 10 %

Purpose of RNAseq

- Annotation
  - Identify and quantify all transcripts present in a given condition
- Differential analysis
  - Identify genes whose expression significantly changed across conditions
  - Identify exons whose inclusion levels significantly changed across conditions

ASTER

Algorithms & software for 3rd generation RNA sequencing



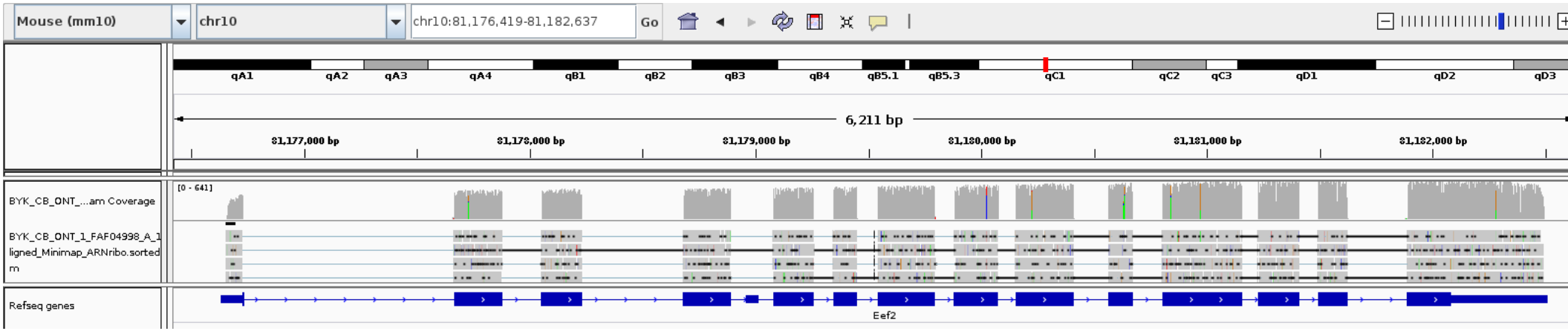
Data generated by Genoscope

- Mouse brain / liver transcriptome
  - Nanopore cDNA : 1.2M reads
  - Illumina : 60M reads
- Using existing software, how can we analyse this dataset ?
- What are the open questions ?

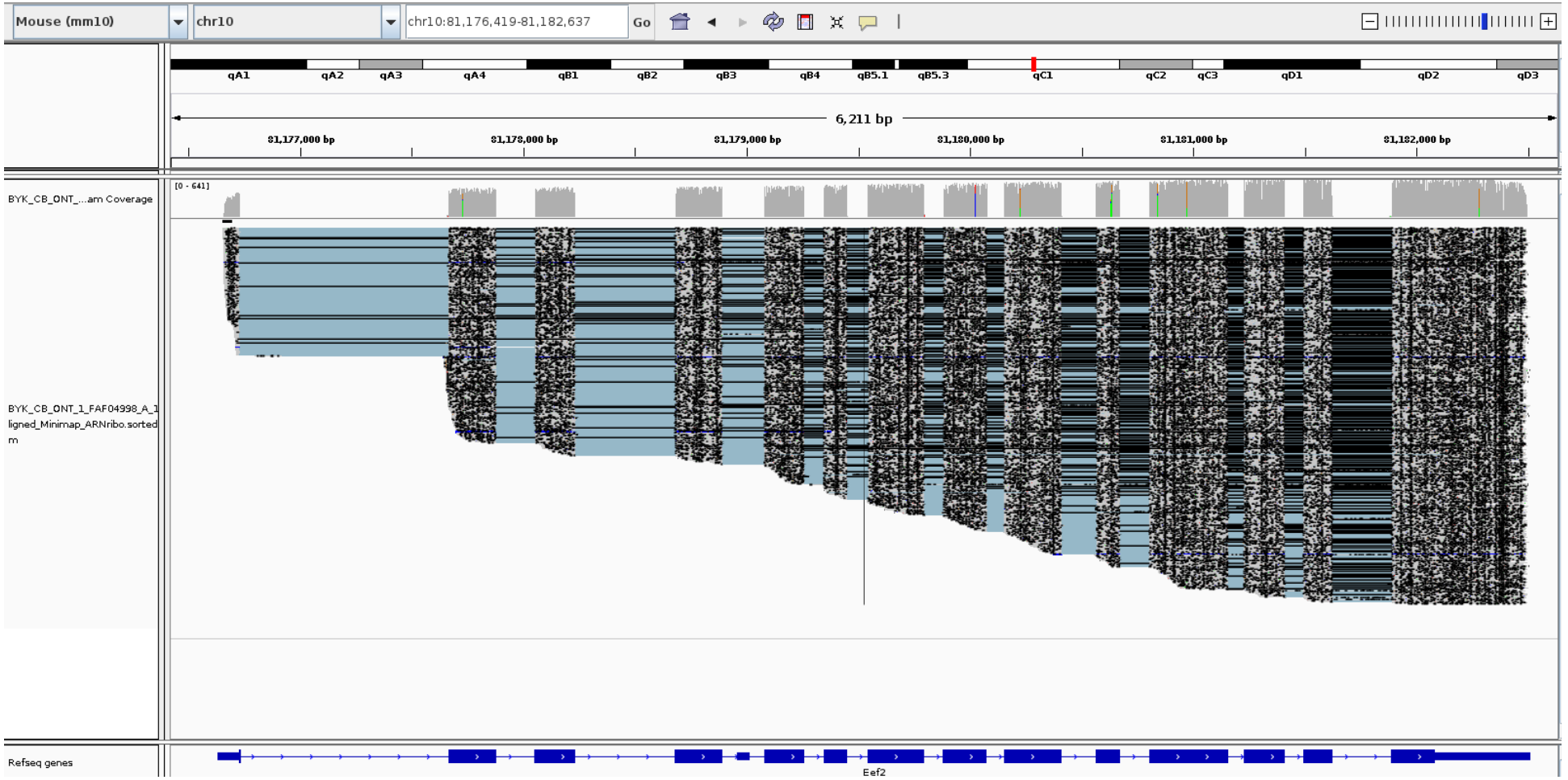
Two mapping strategies

- Map to genome with minimap2 splice
  - 85 % of reads are mapped with 80 % query coverage
- Map to transcriptome with bwa-mem -x ont2d
  - 85 % of reads are mapped with 80 % query coverage

Example of EEF2 gene
Reads are indeed quite long !

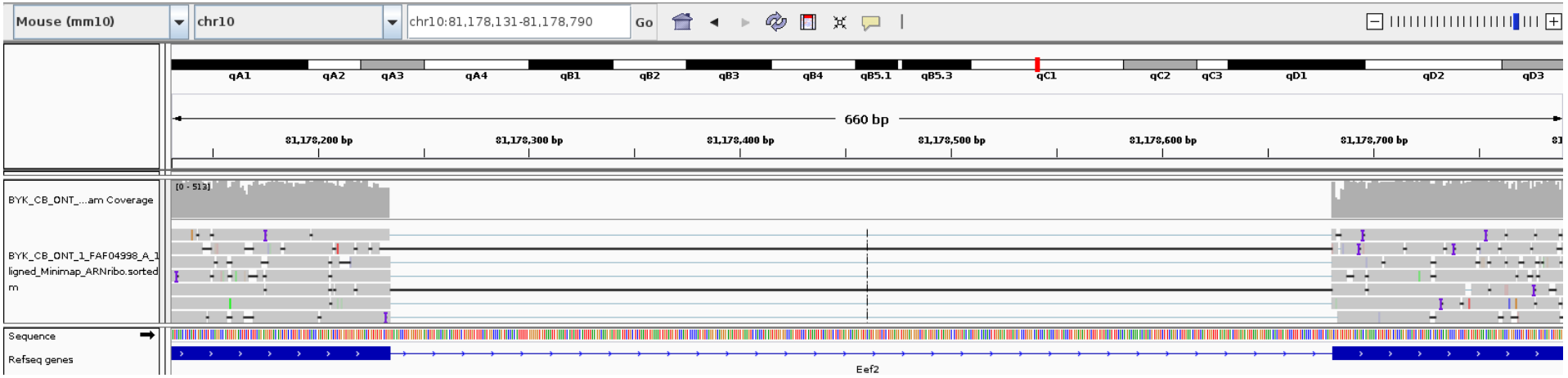


Example of EEF2 gene the staircase effect



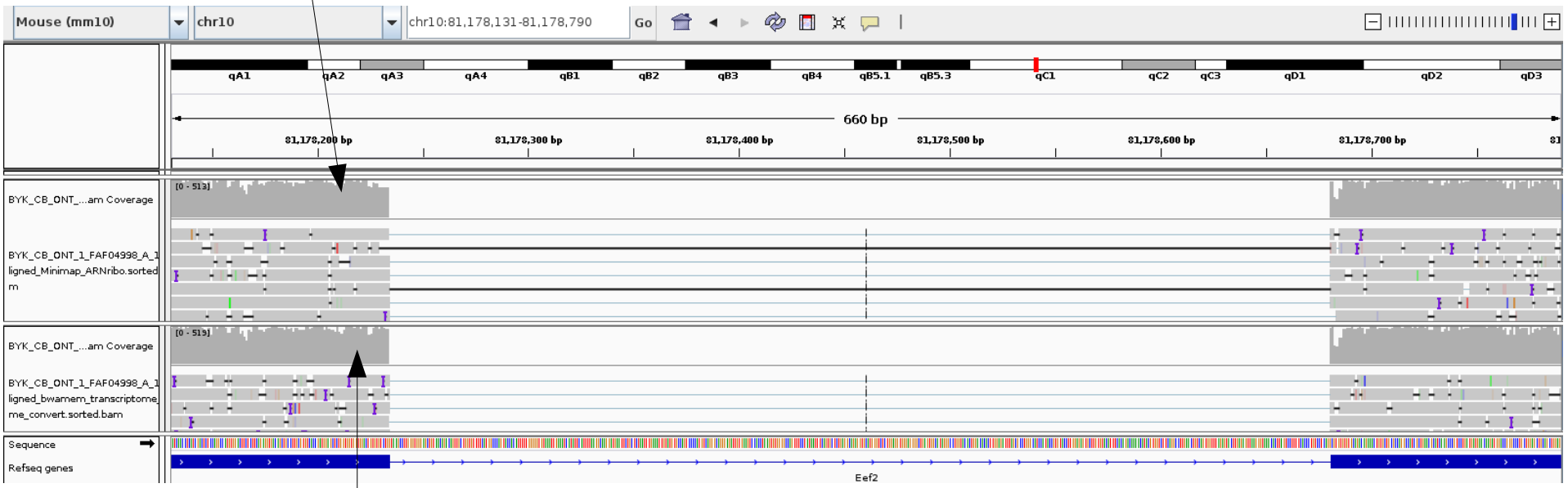
Many reads do not cover the full transcripts
All reads cover the 3'end. This is due to cDNA synthesis which uses polydT primers.

De novo discovery of splice sites is not easy



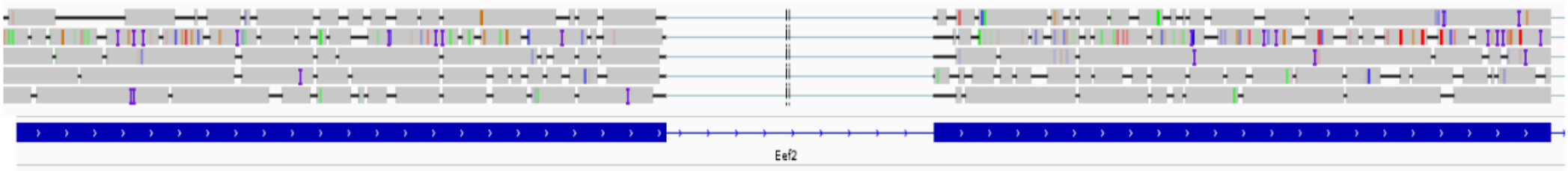
Mapping to annotated splice sites is very easy

Map To Genome



Map To Transcriptome

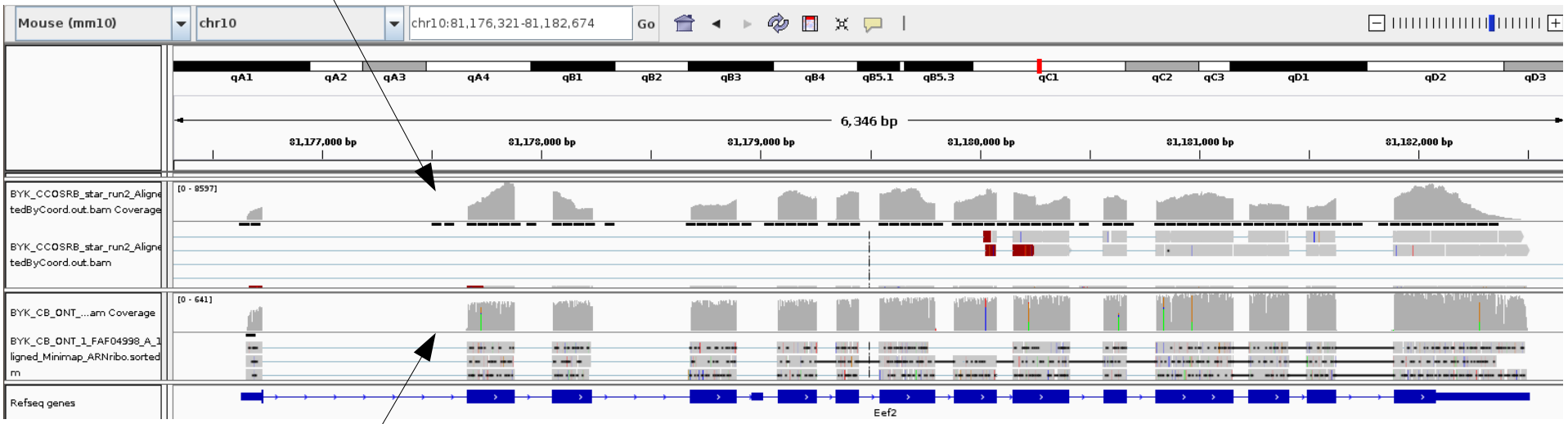
Hard instances for a mapper



Here the solution is to introduce a gap just before the splice site.
These reads could be correctly aligned because we knew the positions of the splice sites
Open question : how to align correctly when no annotations are available ?
Our dataset can be used as a training set

Comparison with Illumina

Illumina

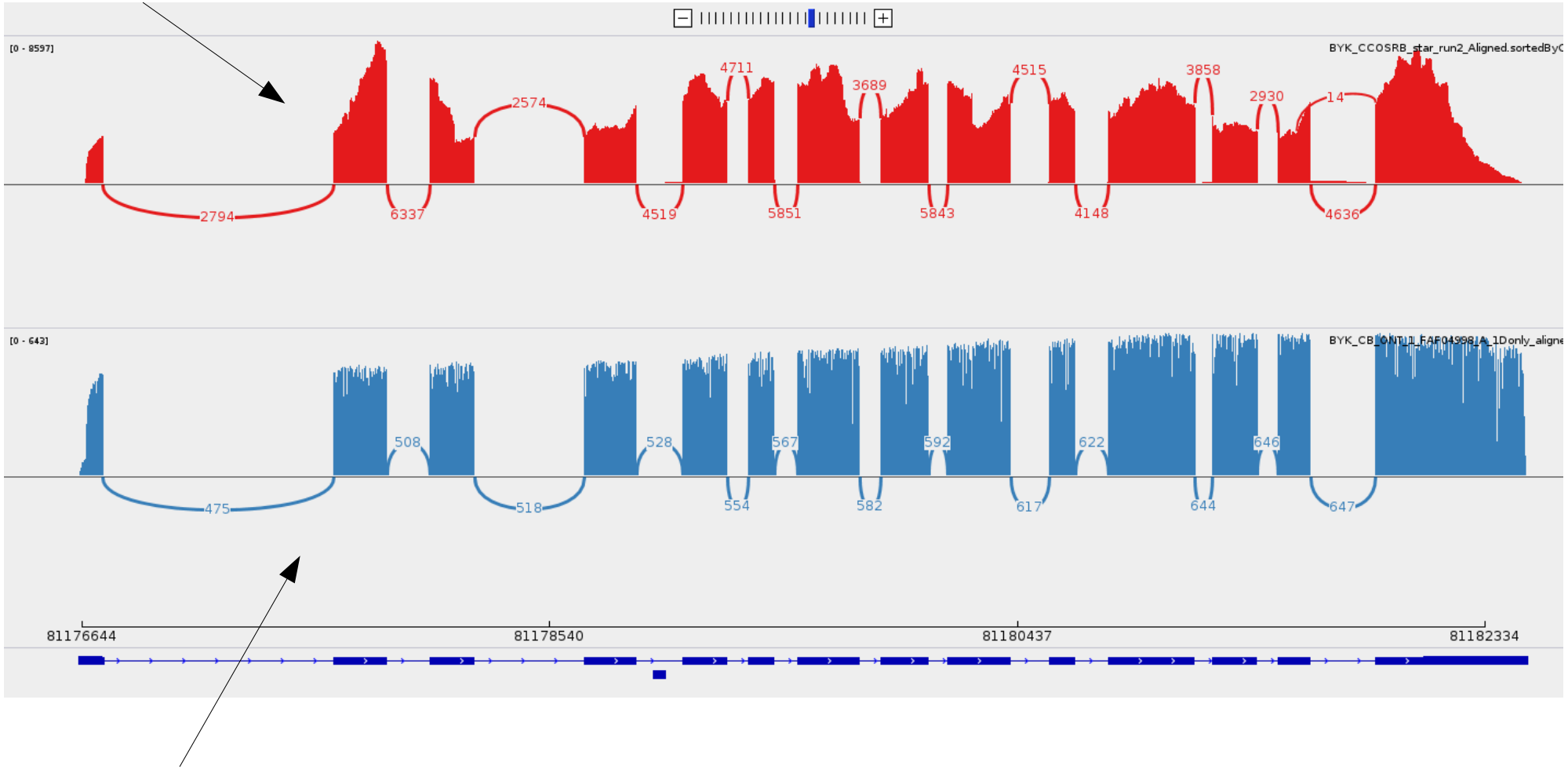


Nanopore

Illumina reads are shorter
There is more local heterogeneity of coverage

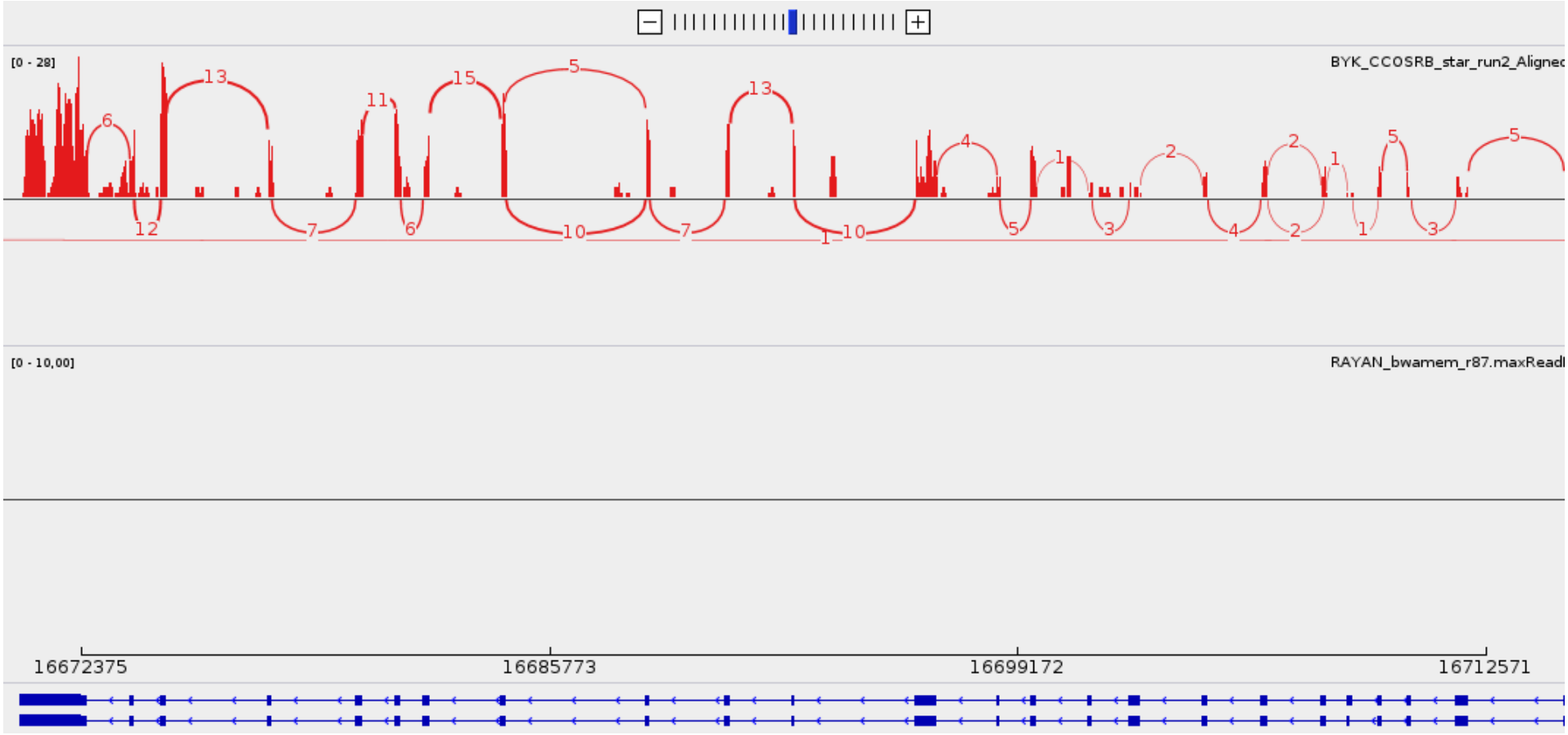
Comparison with Illumina (Sashimi Plot view)

Illumina



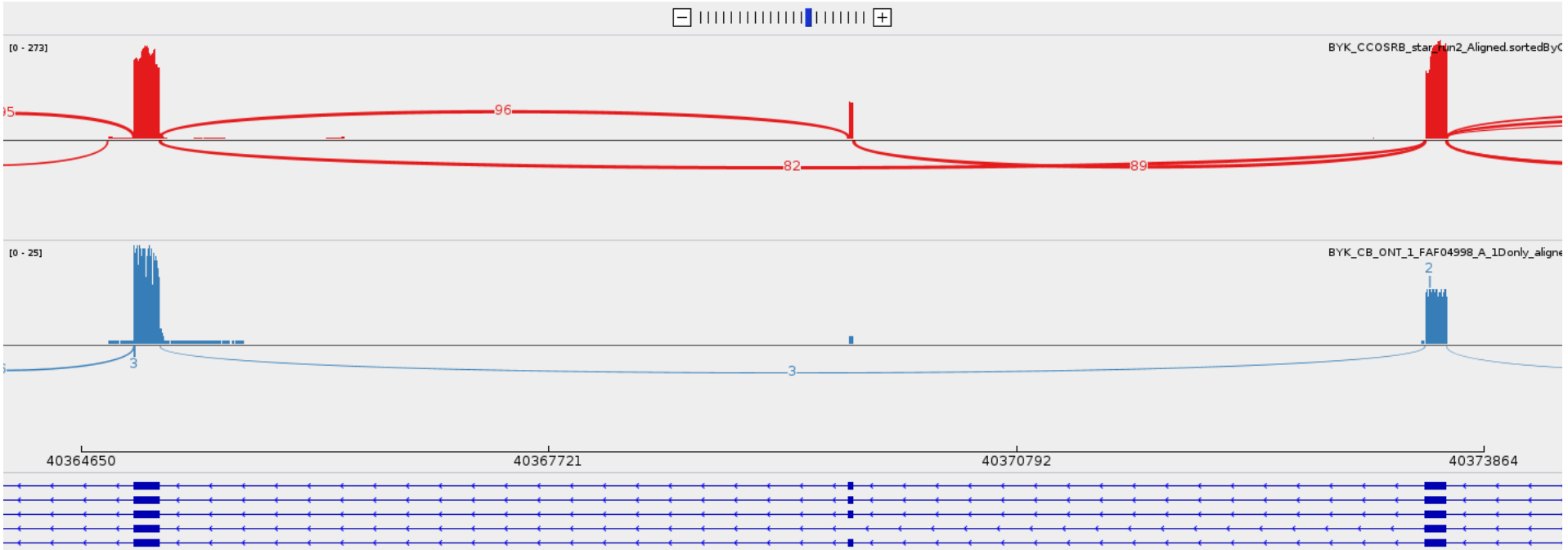
Nanopore

Some genes are not captured at all by Nanopore



Some alternative transcripts are not captured at all by Nanopore

Small exons are harder to find (hard instances for mapping ?)



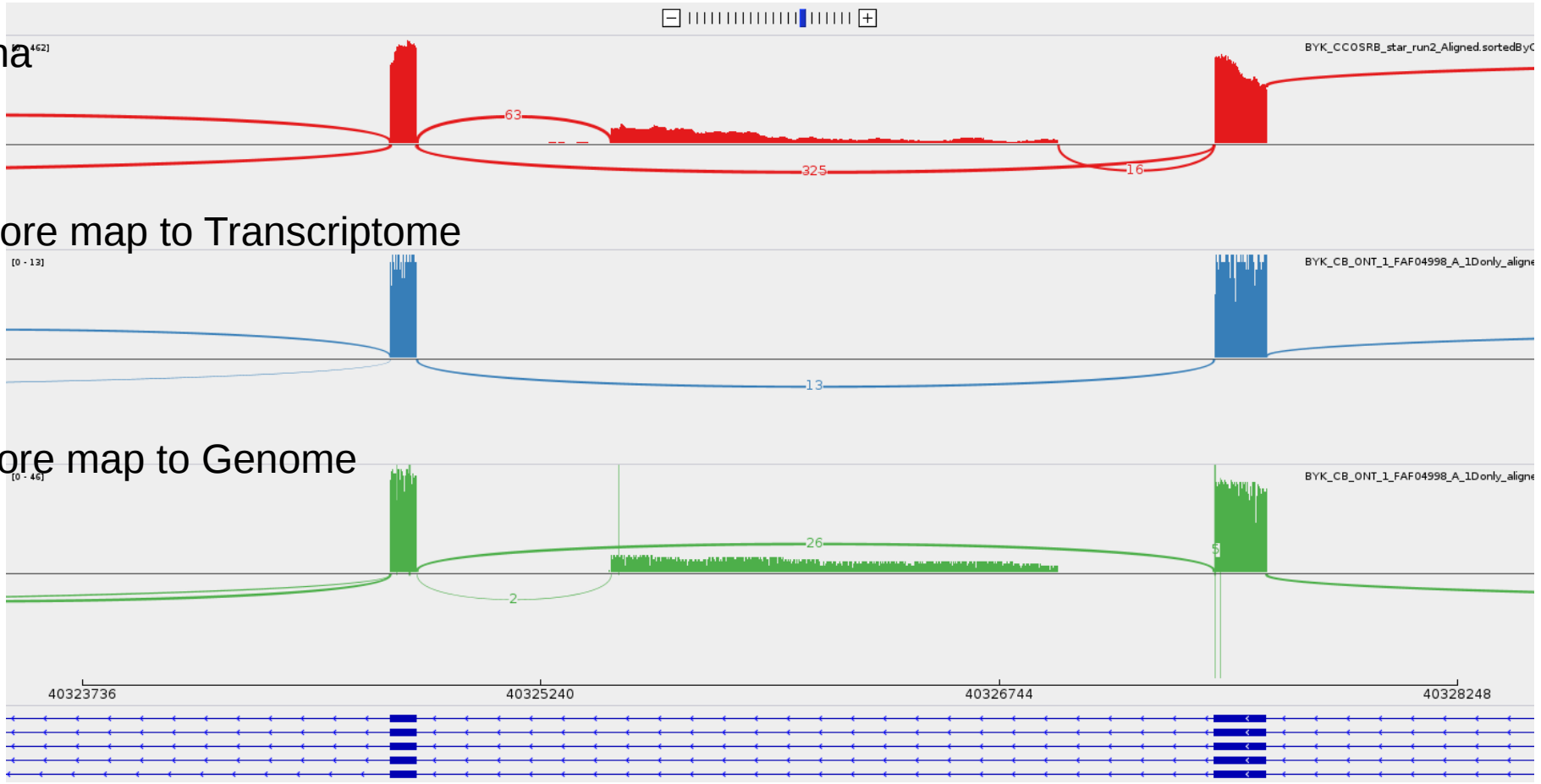
Exon size : 30nt

Novel exons are harder to find (hard instances for mapping ?)

Illumina

Nanopore map to Transcriptome

Nanopore map to Genome



Currently, no long read mapper correctly handles annotation

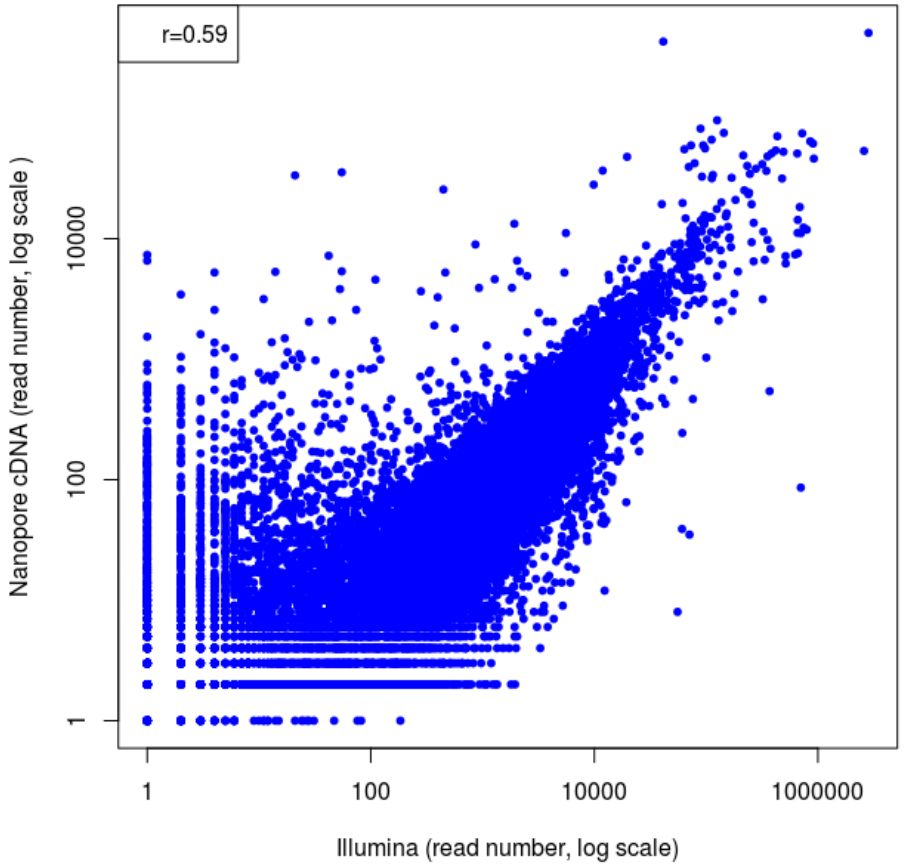
Summary on mapping

- There are still improvements to propose to map long reads, especially when no annotation is available
- However, the difference of depth between technologies (~50-100 fold) leads to missing many isoforms/genes

Quantification

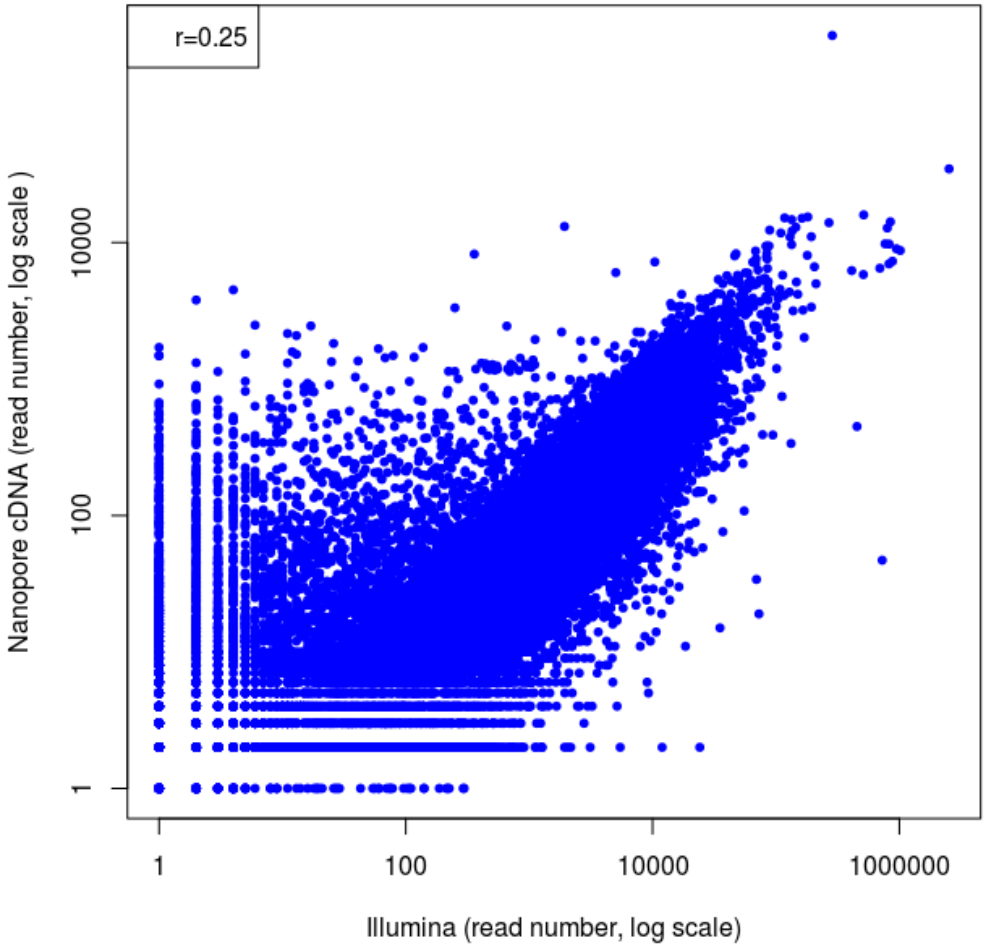
- Each read corresponds to an individual mRNA molecule.
- Counting the number of reads is a proxy for the number of mRNAs
- There are 60X more reads with Illumina. Hence we sample 60X more mRNAs.

Quantification Illumina Vs Nanopore (mouse liver)



Correlation is quite weak. R^2=17 %.
This means that 85 % in Nanopore read counts is not explained by Illumina.
Some genes are detected as poorly expressed by Illumina
and highly expressed by Nanopore
Who is right ?

Quantification Illumina Vs Nanopore (mouse brain)

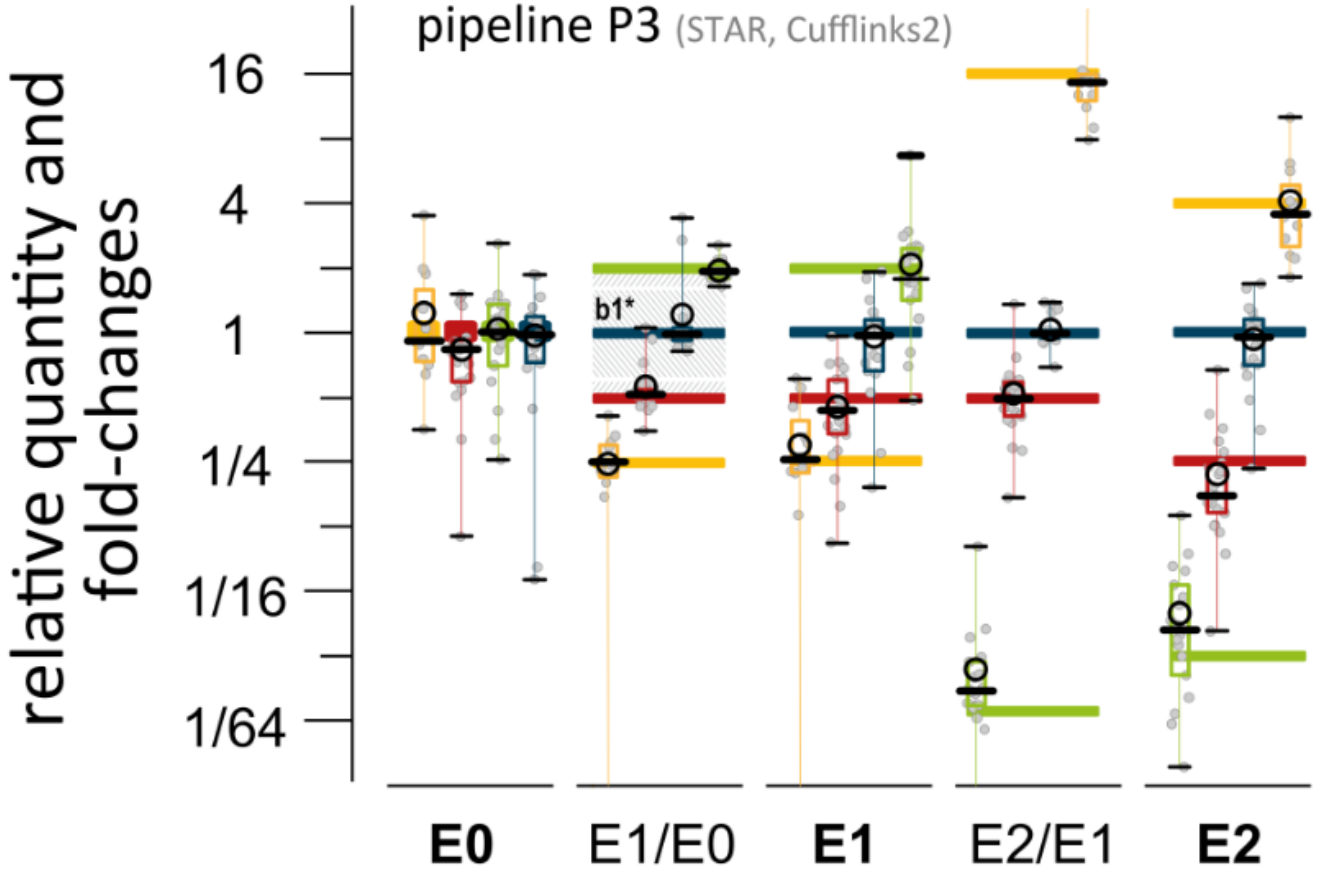


The correlation is even weaker in brain, where more genes are poorly expressed

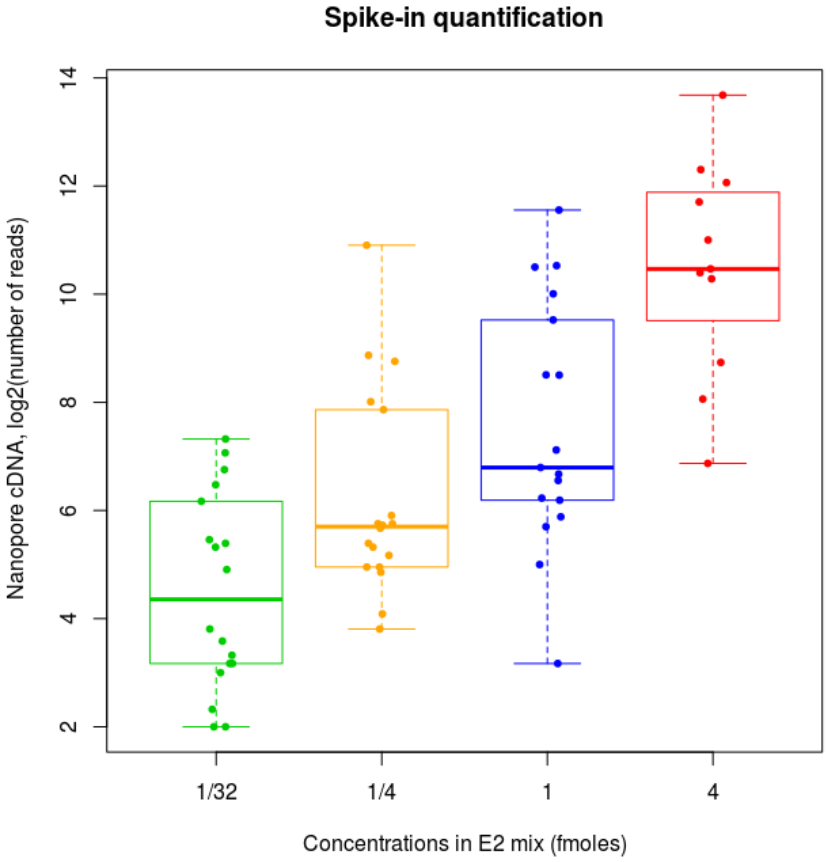
Spike-in data

- In order to know which technology gives the best quantification, we introduced in our samples transcripts in predefined quantities
- SIRV : Spike-In RNA Variants
- Lexogen E2 mix : 7 genes, 10 transcripts per gene, abudance varying from 1/32 to 1

Spike-ins
(Illumina data from Lexogen)

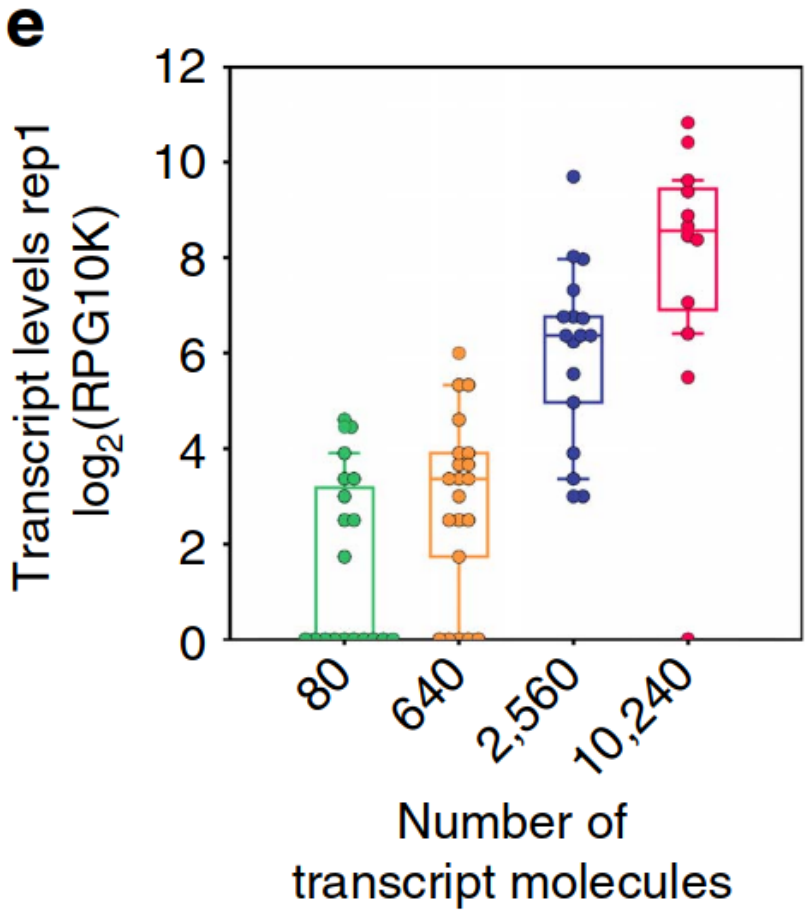


Spike-in results (our cDNA Nanopore data)

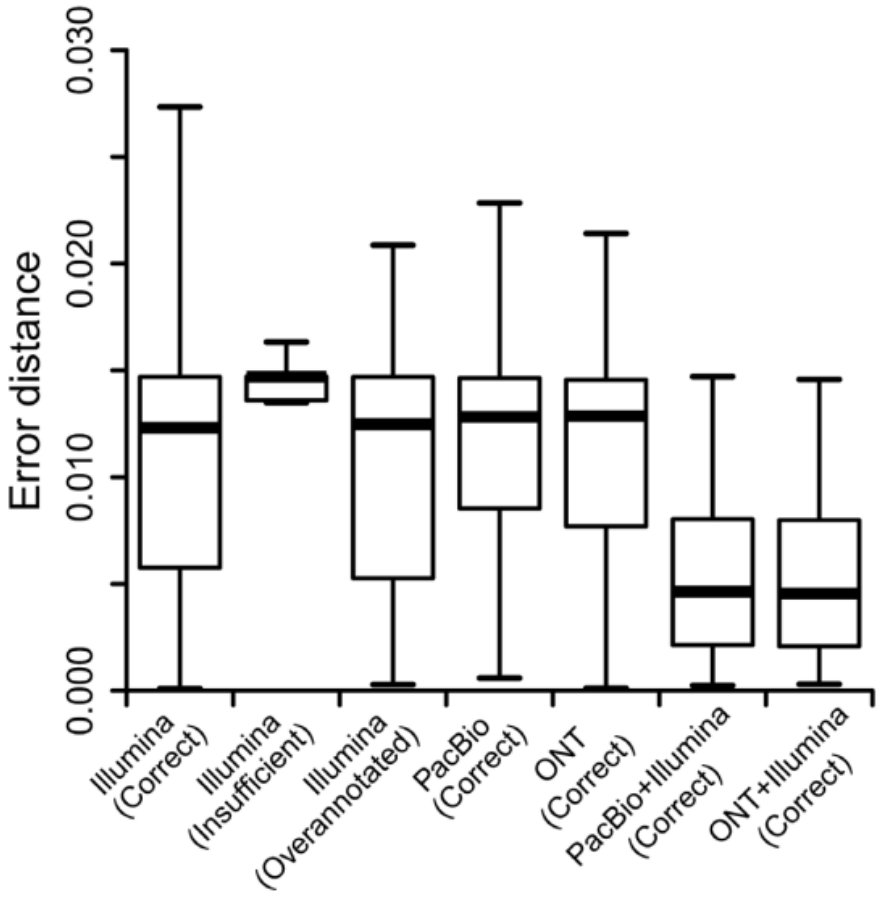


$R=0.55, R^2= 30$ %, this means that 70 % of the variance is unexplained

Spike-in results
Byrne et al. 2017 Nat Comm



Spike-in results
Weirather et al. F1000



Quantification summary

- Illumina and Nanopore do not provide the same quantification
- The quantification by Nanopore is not so reliable, in particular for rare transcripts
- We are waiting for our spike-in Illumina data to have a full comparison
- RNA direct yet provides another quantification

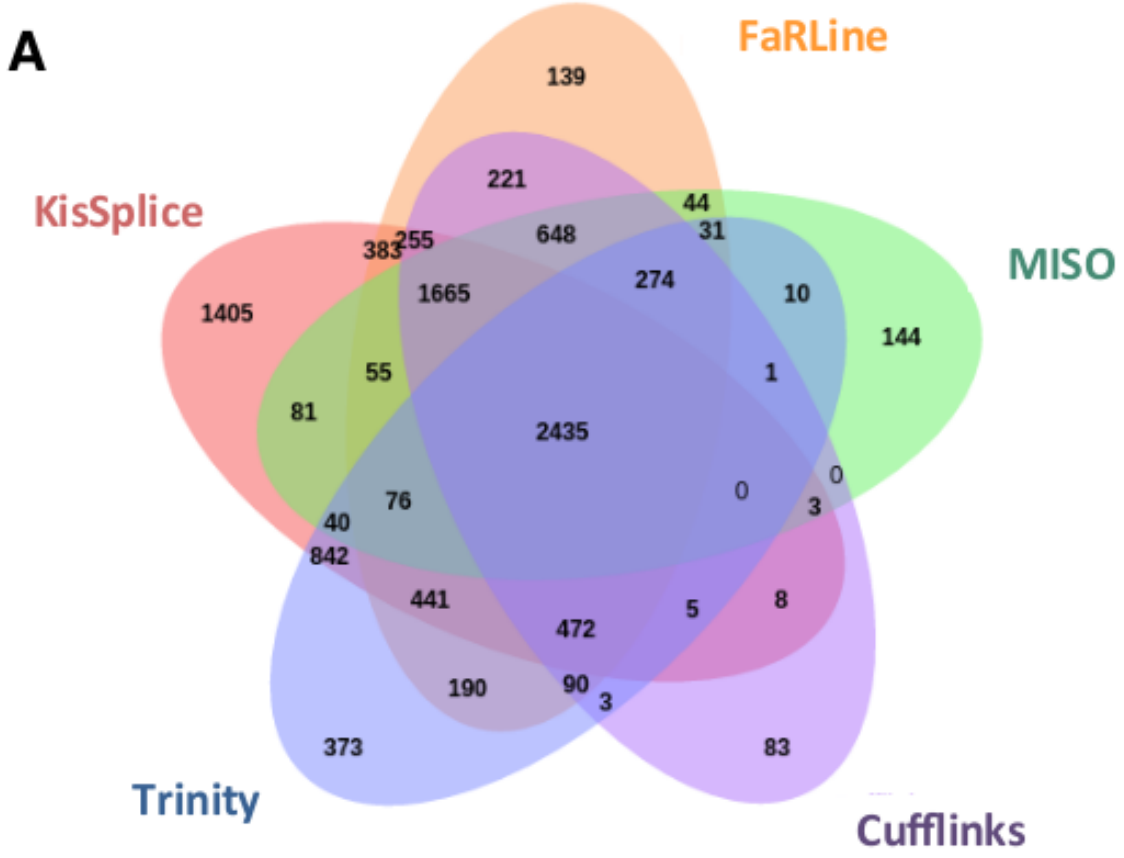
Illumina Vs Nanopore

- Illumina is stronger for
  - Discovering Splice sites
  - Differential analysis (higher read counts --> more power)
- Nanopore is stronger for
  - Phasing exons

Summary Bioinformatics Developments

- Technology moves very fast
- Not clear how much time we should spend on bioinformatics development
- Many questions are still open on bioinformatics of splicing with Illumina data
- We aim at developping methods which take advantage of Illumina depth and Nanopore length
- How to efficiently use annotations is not easy

Various methods to find exon skipping from Illumina data



Bibliography

nature COMMUNICATIONS

ARTICLE

Received 24 Apr 2017 | Accepted 23 May 2017 | Published 19 Jul 2017

DOI: 10.1038/ncomms16027 OPEN

Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells

Ashley Byrne[1,2], Anna E. Beaudin[3,†], Hugh E. Olsen[2,3], Miten Jain[2,3], Charles Cole[2,3], Theron Palmer[3], Rebecca M. DuBois[3], E. Camilla Forsberg[3,4], Mark Akeson[2,3] & Christopher Vollmers[2,3]

Bioinformatics, 2017, 1–7

doi: 10.1093/bioinformatics/btx668

Advance Access Publication Date: 23 October 2017

Original Paper

Sequence analysis

Evaluation of tools for long read RNA-seq splice-aware alignment

Krešimir Križanović[1], Amina Echchiki[2,3], Julien Roux[2,3,†] and Mile Šikić[1,4,*]

F1000Research F1000Research 2017, 6:10(

Check for updates

RESEARCH ARTICLE

REVISED Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis [version 2; referees: 2 approved]

Jason L Weirather [1*], Mariateresa de Cesare [2*], Yunhao Wang [1,3,4*], Paolo Piazza[2], Vittorio Sebastiano[5,6], Xiu-Jie Wang[4], David Buck[2], Kin Fai Au [1,7]

Other resources

- https://github.com/nanopore-wgs-consortium/NA12878/blob/master/RNA.md
- Minimap2 Vs gmap
  - http://complex.zesoi.fer.hr/index.php/en/blog-en/56-gmap-vs-minimap2

Acknowledgments

- All members from the Aster Project