One year of developments and collaborations around the MinION on the Genomic facility of the IBENS.

Laurent Jourdren (CNRS – IBENS)
Sophie Lemoine (CNRS – IBENS)
Bérengère Laffay (CNRS – IBENS)

13th of December 2017

An on-going project used to validate our protocols and devices

- A mouse model of peripheral nervous system development

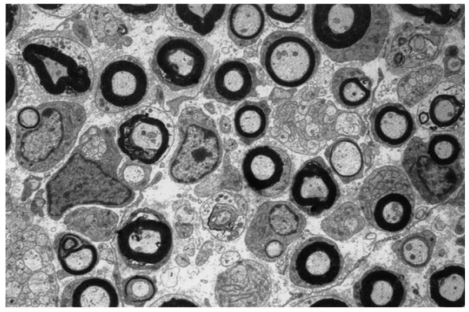
➢ We compare 2 conditions in triplicates
  - Krox20 (Egr2) KO that blocks myelination
  - Wild Type strains

➢ The model is well adapted to splicing event characterisation

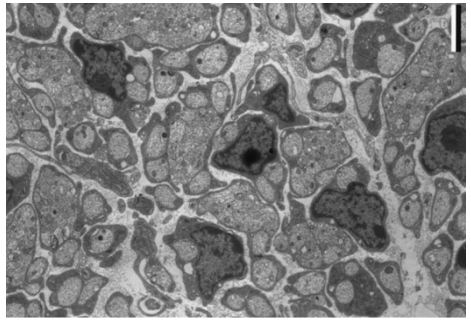
- A molecular biology team directly implied that can verify targets
- The samples are regularly prepared and systematically used to validate all our protocols and devices
  - 17 library preparation protocol tested;
  - 12 runs using Illumina sequencing technology (PE150, SR50, SR75 and PE75).
  - And now ONT…

➢ We have a huge amount of data on this model

Wild Type



Krox20 $^{-/-}$ **Knock Out**

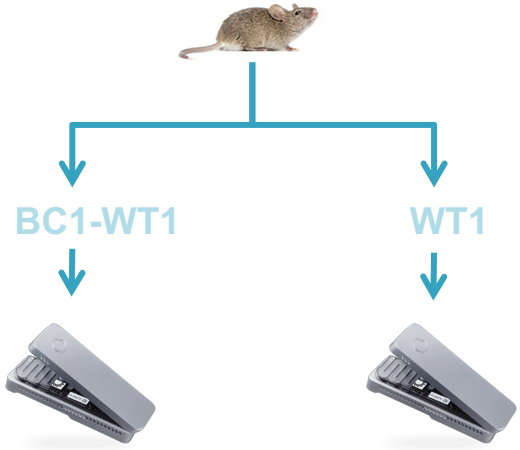


Two test designs to begin with RNA-Seq on MinION

- **Is it possible to run RNASeq on a MinION with multiplexed samples as on an Illumina ?**

We sequenced 2 biological conditions in triplicates.
This design was run 3 times.

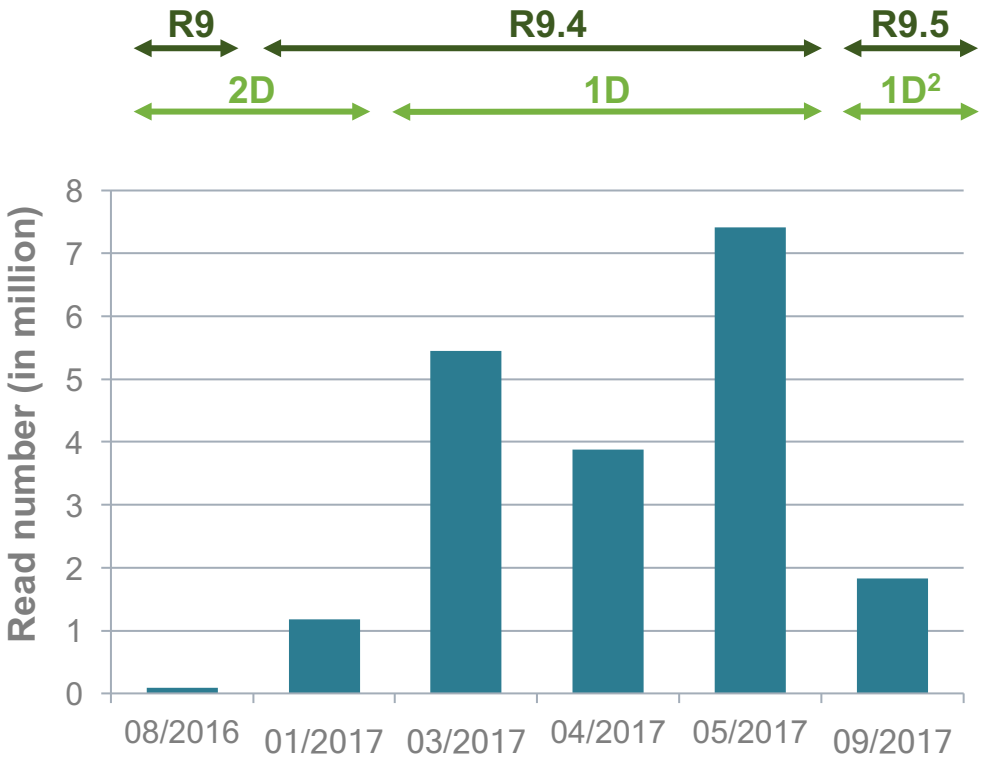
- **What can be the effects of barcodes on libraries and runs ?**



We sequenced one wild type sample from our dataset with or without barcode.
This design was run 3 times.

Changes in flowcells and sequencing protocols had a great influence on read throughput

We produce an **average of 5.6 million reads** with R9.4 flowcells and 1D protocol.

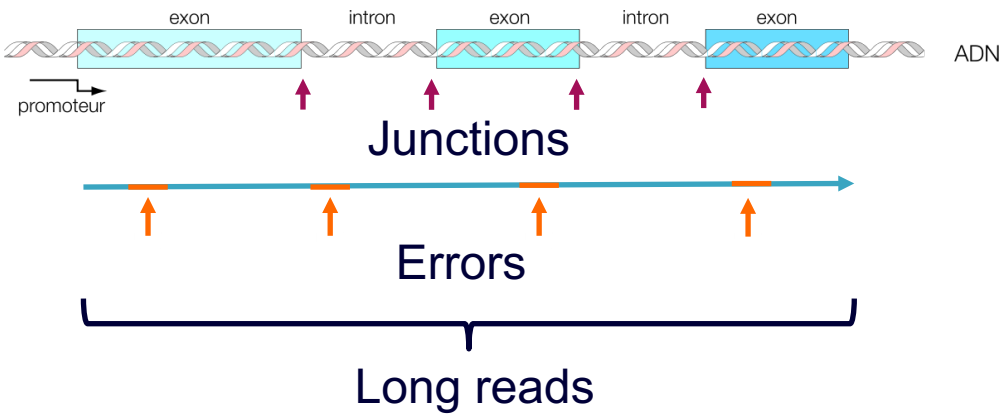


The 1D protocol allowed a great improvement in the read number

- But from 100,000 to up to 7 million reads, the data management was a big issue
  - Fast5 file management
  - Quality control of the run
  - Read alignment

cDNA read alignment

The aligner has to manage :



→ **GMAP**

GMAP + mm10 genome

Consensus 2D reads → 100,000 reads of a multiplexed sample → **Alignment**

1D reads → 500,000 reads of a multiplexed sample → **Alignment** ✗

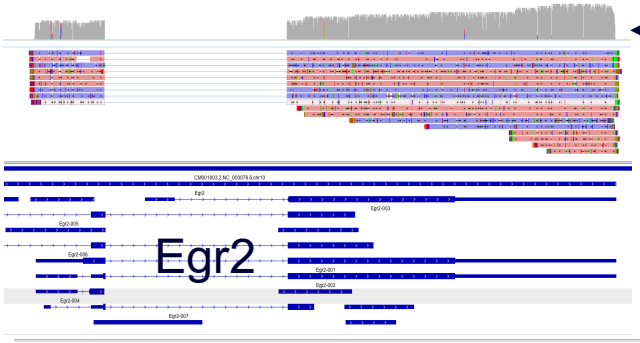
- Heavy read loss
- Shorter Alignments in 1D
- 1D sequencing doubles the error rate 8% to 15%
- Fails most of the time (memory leaks)

GMAP cannot deal with error-prone long reads and junctions together

GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 2005 21: 1859-1875.

Encouraging enough results to go further

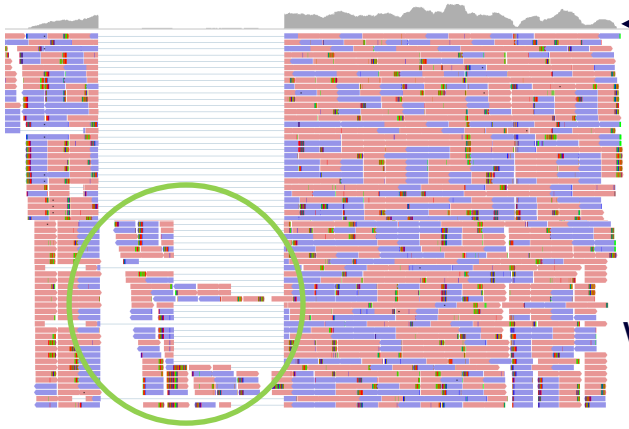
WT 2D Minion



← Homogenous coverage

← Heterogenous coverage

WT SE150 Illumina



Shorter reads make wrong alignment easier

The results are promising : it works !

The bottleneck is the mapping step :

- Error rate in 1D data extend the mapping time
- To improve the mapping step we need to improve quality of 1D data to reach the quality of 2Ds

Read correction to improve the alignment

To align with GMAP, we tried to correct the reads

- We have tons of Illumina reads for the same samples
  - Hybrid correction

Tool	Correction method	Developped for
proovread	**map short reads on long reads**	**PacBio**
AHA	Align short reads on long reads, part of a PacBio assembler	PacBio
ECTools	Multiple sequence alignments	
Lordec	**kmer spectrum**	**PacBio**
LSC	**Multiple sequence alignments**	
NanoCorr	**Multiple sequence alignments**	**Oxford Nanopore**
PBcR	Multiple sequence alignments	PacBio

- **Proovread seems to perform well on high error rated and discontinuous data**
- **Lordec, NanoCorr and LSC are worth being tested**

Laehnemann, D., Borkhardt, A. & McHardy, A. C. Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. Brief. Bioinformatics 17, 154–179 (2016).

Proovread tests on 2D and 1D data

	proovread 2.12*		proovread 2.14*
Protocol	2D	1D	1D
Read number	131 698	1 011 180	1 011 180
Computation time	4 days	**> 1 month**	**18 days**
Untrimmed reads	118 381		790 560
Trimmed reads	88 867		170 903

*Correction with an equivalent Illumina SE 150 illumina sequencing

- Crazy computation time when correcting 1D data
  - ➢ Not reasonable for a platform daily use
- The read quantity decreases a lot along the correction process of 1D data

➢ **Read correction could not be a perspective for a daily use**

Alignments of 1D data with BWA-MEM

BWA-MEM was probably not the best mapper for RNASeq

- But we needed to see our data !

The alignment was performed on mm10 cDNAs

Sample description	Input raw reads	Alignments	% unique alignments
WT01_BC01	2 575 059	3 933 410	35,72
WT01_BC01	4 694 580	6 980 219	38,10
WT01_BC01	1 712 485	2 307 047	33,81
WT01	4 116 471	5 951 589	39,12
WT01	5 369 445	7 340 601	42,72
WT01	5 101 854	6 966 709	43,49

About unique alignments:

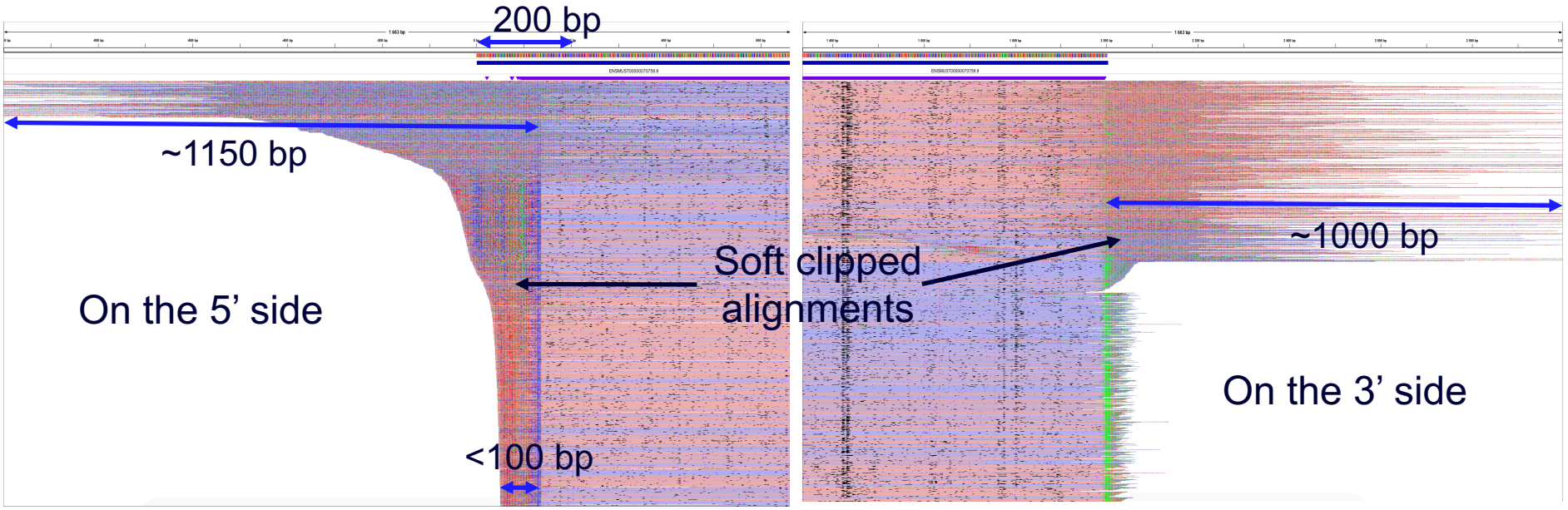
- **Are similar between barcoded and not barcoded runs**
- **Represent only a third of the alignments**

Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. [arXiv:1303.3997v2]

A quick look on the ends of reads (1)

WT1 **without barcode** aligned on mm10 ens88 cDNA

- multimatches are removed
- Mpz-201 (forward strand) is one of the most expressed transcript
- What does it look like on the 5' end?



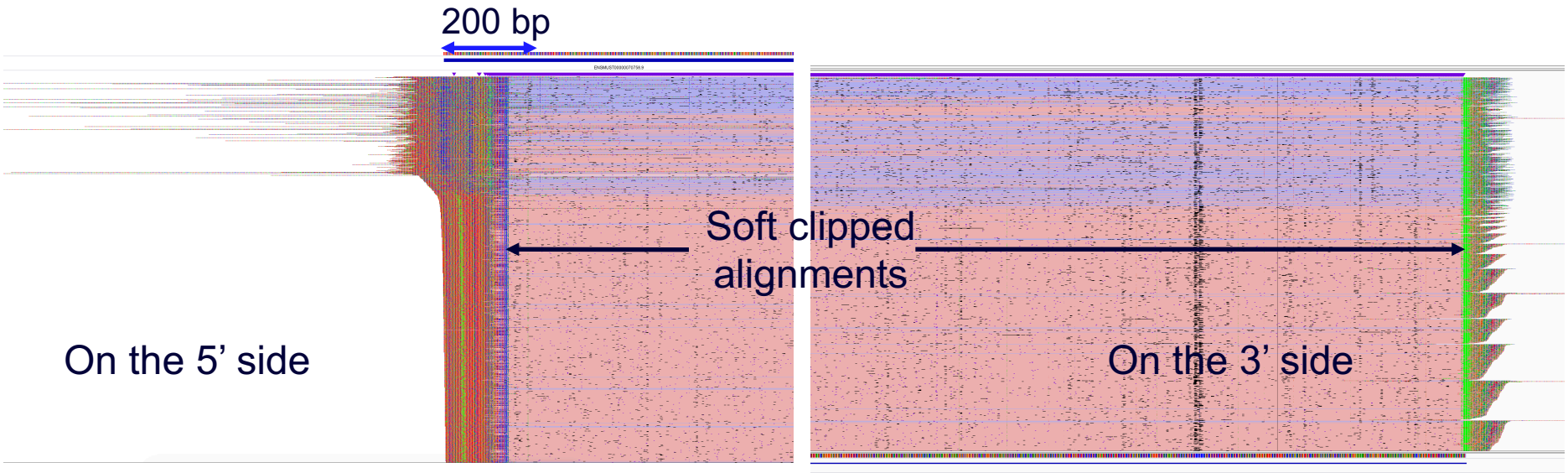
The 5' and 3' ends are very dirty

- A good explanation for the hybrid correction failure and the mapping issues

A quick look on the ends of reads (2)

WT1 **with barcode** aligned on mm10 ens88 cDNA

- multimatches are removed
- Mpz-201 (forward strand) is one of the most expressed transcript
- What does it look like on the 5’ and 3’ end?



The nonsense sequence looks different in 5’ on a barcoded sample :

- Maybe smaller ?
- It’s still dirty

The ends of reads need to be cleaned before the mapping step

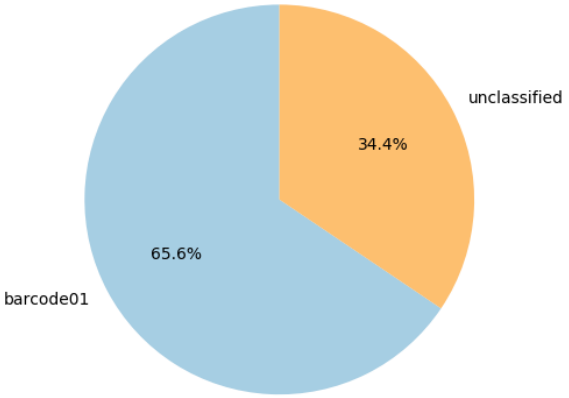
- Both 5’ and 3’ extremities have misaligned sequences
- These misalignments are soft-clipped and penalise dramatically the global alignment quality (RNAs are short sequences)

If reads are cleaned before mapping we expect :

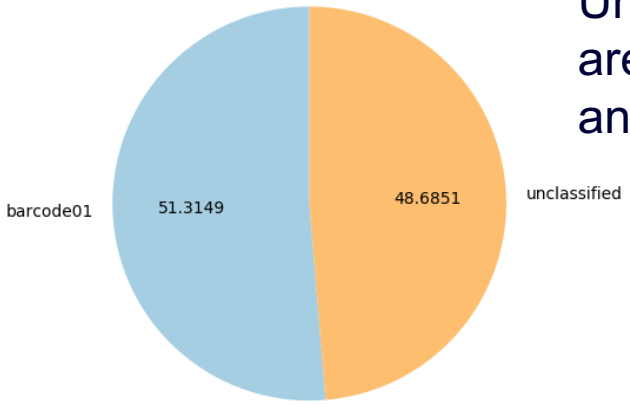
- More reads aligned
- Better alignments

➢ It could also be a strategy to rescue reads that were not demultiplexed properly (sequencing errors also affect barcodes)

Run 1



Run 2



Unclassified reads are lost for further analysis

Very few tools are available to clean the reads

We cannot use cutadapt or trimmomatic to cut ends :

- Size of sequence to cut varies
- Quality is lower than illumina standards

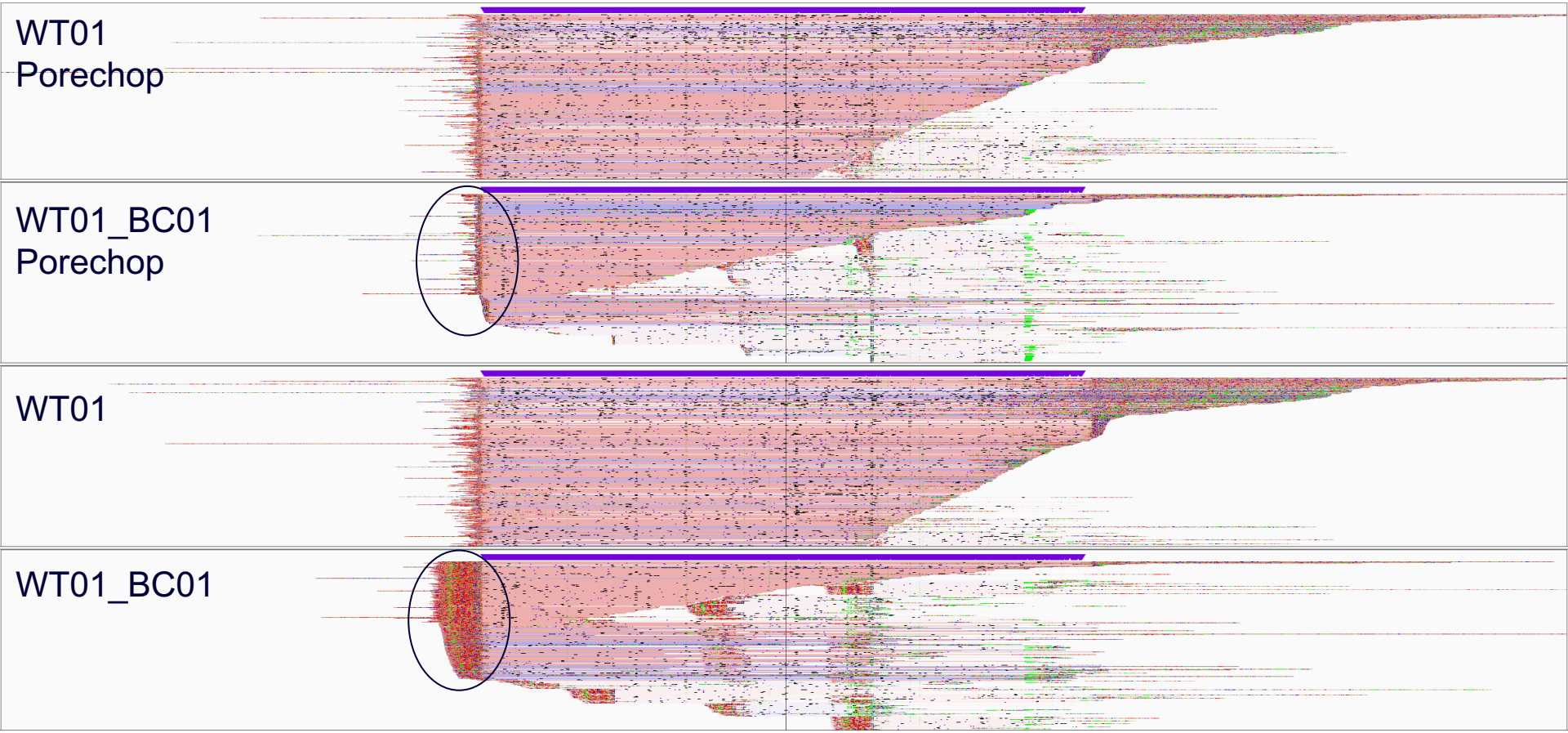
is currently the best available tool to clean nanopore reads

Samples	Raw read	% reads after PoreChop	% unique alignments	% multiple alignments	% unmapped
BC samples	3 634 820		37	62	2
NonBC samples	4 742 958		41	51	8
BC samples+ PoreChop	**3 634 820**	**98,9**	**56**	**42**	**2**
NonBC samples+ PoreChop	**4 742 958**	**99,7**	**49**	**43**	**9**

- No influence on the percentage of unmapped reads
- Decrease of multimapped reads (mapping on cDNAs= a lot of multiple alignments)
- Increase of unique reads, especially on barcoded samples

https://github.com/rrwick/Porechop

A quick look on the ends of reads (3)



- The gain of PoreChop is visually unclear on the non barcoded library
- It is stricking on the barcoded library

PoreChop , pros and cons

- ✓ The sequences are cleaner
- ✓ The reads align better
- ✓ The loss of reads is insignificant

- ❖ It takes several hours per sample
- ❖ The sequences are still dirty
- ❖ The adaptor and barcodes sequences used in the protocols are unclear
  - ➢ The theoretical sequences do not cope with the observed sequences…
    - ➢ Could we have something better Santa Nanopore ??

- ❖ The sequences are part of the code what makes the configuration uneasy

- ➢ **PoreChop cannot be integrated yet in our analysis pipeline**

As we are not specialized in algorithms,
we began to work with the LIRMM in Montpellier on the demultipexing and trimming steps

Minimap2 can perform much better than BWA-MEM

A versatile pairwise aligner for genomic and spliced nucleotide sequences

- Can be used for long and short reads
- Performs Splice-aware alignment of PacBio Iso-Seq or Nanopore cDNA or Direct RNA reads
- Does not mind a ~15% error rate

6 x1D barcoded samples	Reads /sample	% Unmapped reads /sample	% Reads With Unique Alignment /sample	% Unique reads on exons
run1	493 119	64	34	34
run2	**403 425**	**7**	**90**	**87**
run3	829 644	29	52	54

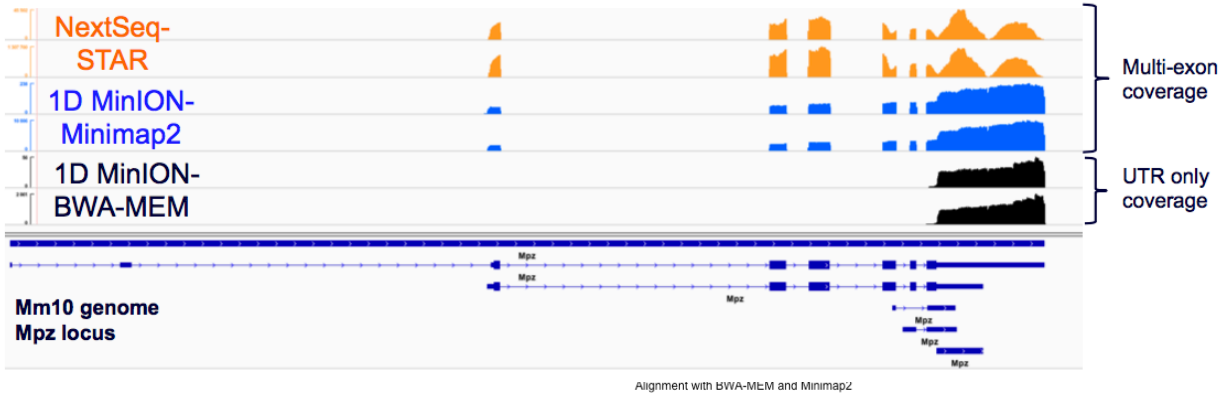
- Runs can be very heterogeneous
- The more you get does not mean the more pertinent you have
- Alignment percentage can reach better level than STAR on short reads

Li, H. (2017). Minimap2: fast pairwise alignment for long nucleotide sequences. arXiv:1708.01492

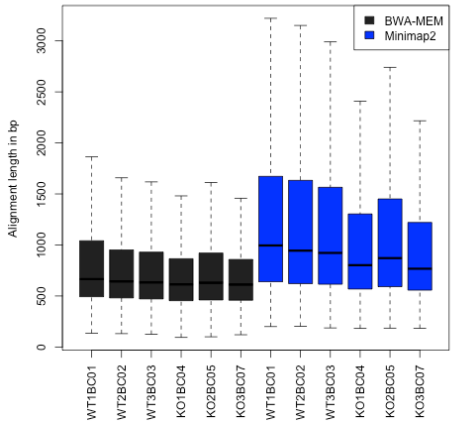
Minimap2 versus BWA-MEM

6 x 1D barcoded samples (1 run)	Mean read number	Mean alignment number	Mean Unmapped reads	Mean uniquely mapped read number	% of uniquely mapped reads
BWA-MEM	403 425	1 277 317	3 830	128 968	32
Minimap2	403 425	658 190	31 695	365 183	87

- Minimap2 outclasses BWA-MEM in number of reads uniquely mapped
- BWA-MEM does not align well over junctions, it cannot be used to identify isoforms
- Minimap2 behaves well over junctions



- Minimap2 alignments are much longer than BWA-MEM alignments

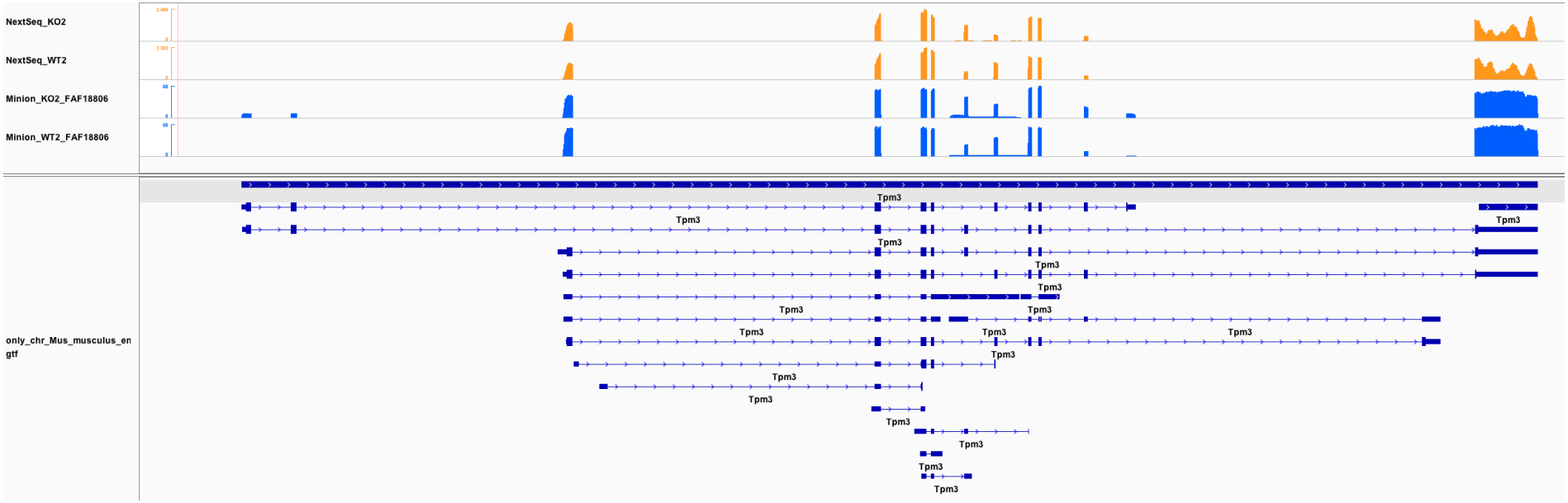


Minimap2 is now integrated to Eoulsan, our analysis pipeline

Jourdren L, Bernard M, Dillies MA, Le Crom S, Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses. Bioinformatics. 2012 Jun 1;28(11):1542-3

Detection of splicing events really works

- Collaboration with GenoSplice to **detect new splicing events** by comparing ONT with Illumina reads.
- We found Tropomyosine (Tmp3) transcripts not seen using short reads.

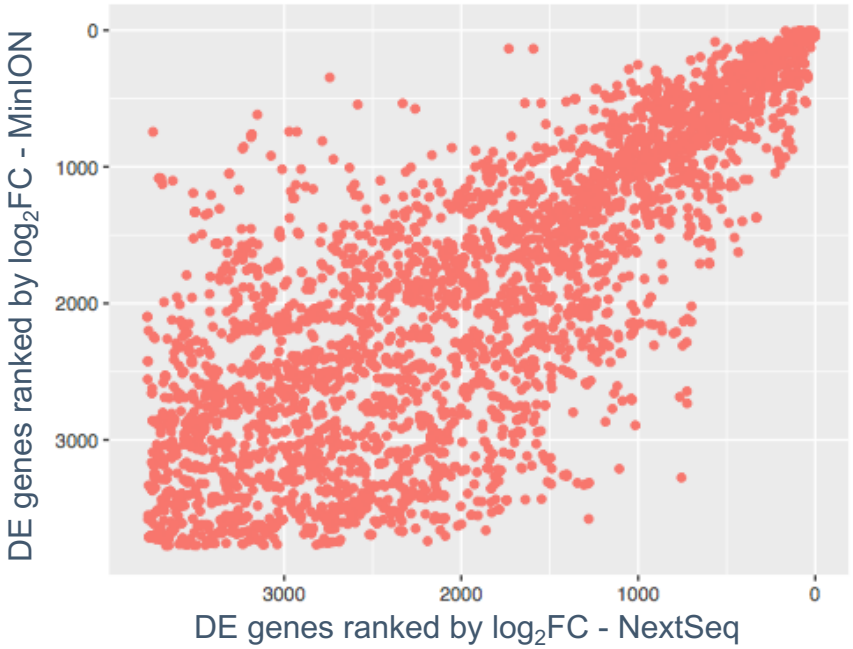


MinION 1D reads can be used for differential analysis

- We performed differential analyses on the multiplexed design:

  3 x KO Egr2 versus 3 x WT

- We get 6,551 differentially expressed transcripts (adjusted p-value < 0.01) with 300,000 alignments by sample
- 86% of these transcripts are shared with Illumina analysis

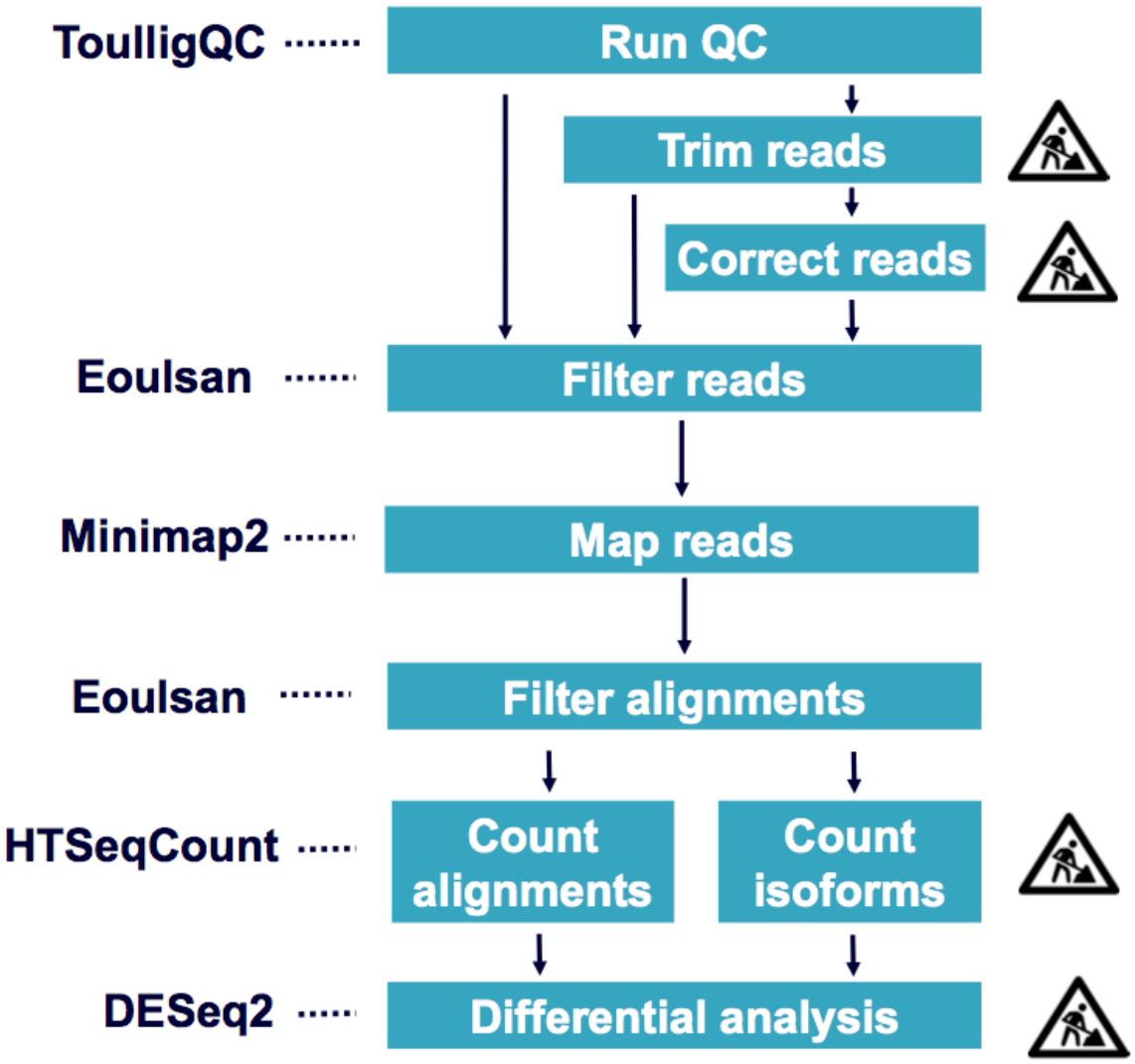


The GO enrichment of the MinION data is what we expected:

- myelin assembly
- fatty acid biosynthetic process
- Lipid biosynthesis…

Our controls behave the way they should (Mpz, Pmp22, Mbp, Prx….)

Eoulsan includes new tools dedicated to Nanopore data



- Eoulsan is now updated to perform differential analyses on MinION reads
- The specific isoform steps are under development (Bérengère Laffay-Master2 internship during 2 years)
- The improvement of the demultiplexing phase is crucial to get a higher coverage

The IBENS genomics facility team

https://genomique.biologie.ens.fr

genomique@biologie.ens.fr Genomique_ENS