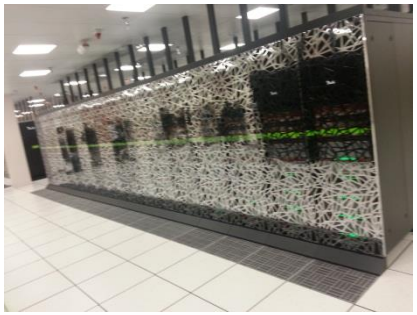
DE LA RECHERCHE À L'INDUSTRIE

www.cea.fr

RNA sequencing with the MinION at Genoscope



Jean-Marc Aury

jmaury@genoscope.cns.fr

@J_M_Aury

Overview

- Genoscope Overview
- MinION sequencing at Genoscope
- RNA-Seq using the Oxford Nanopore technology

Genoscope Overview

http://www.genoscope.cns.fr

- French National Sequencing Center lead by Patrick Wincker, created in 1997 and part of the CEA since 2007

- Provide high-throughput sequencing data to the Academic community, and carry out in-house genomic projects

- Focus on biodiversity : *de novo* sequencing and metagenomic projects (TaraOceans)

- But…. it's not enough to just know one individual's DNA. A single reference genome is not compatible with resequencing approaches

Triticum sp
(wheat)

Musa acuminata
(banana)

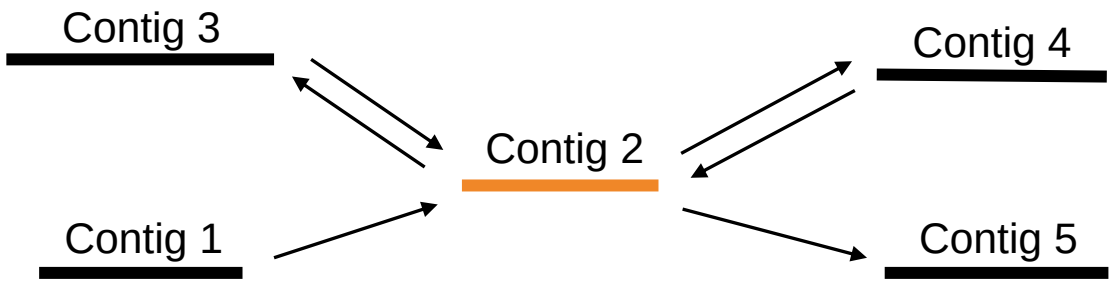
Quercus robur
(oak)

Brasssica napus
(seed rape)

Genome Assembly Difficulties

Genome

Short reads sequencing



Contig graph

=> Repetitive regions lead to fragmented assemblies and under-estimate repeat content

Sequencing capacities

2 Illumina HiSeq 2500
2 Illumina HiSeq 4000

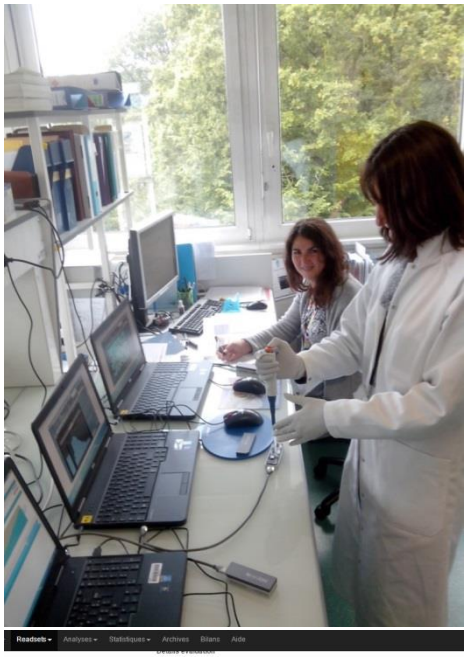
2 MiSeq

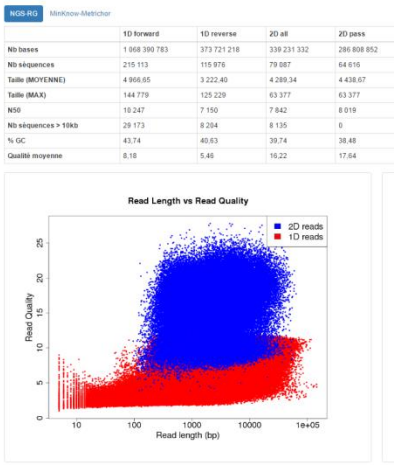
6 Oxford Nanopore MkI
1 PromethION

1 Irys System

MinION sequencing at Genoscope

- 6 MinION devices
- >800 flowcells; >50 different organisms; ~700Gb of ONT reads ; DNA and RNA samples
- *de novo* assembly (22 yeast strains ~12Mb, 4 fungi genomes ~30Mb, several bacterial genomes, >10 plant genomes of 400-700Mb) and gene prediction
- Software development for the automation : management of the data flow, storing metrics in our LIMS
- Benchmark several DNA preparation protocols to obtain longer reads (size-selection using the blue pippin)

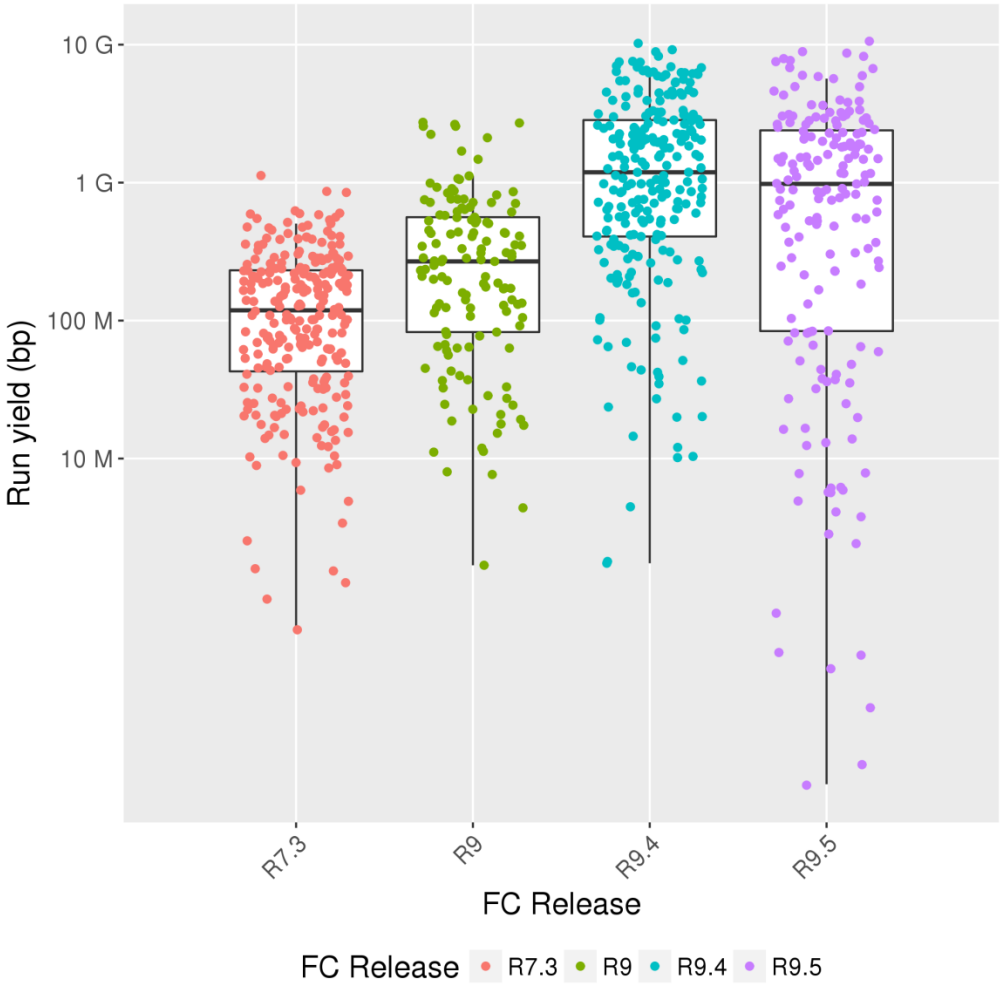




	1D forward	1D reverse	2D all	2D pass
Nb bases	1 068 390 783	373 721 218	339 231 332	286 808 852
Nb séquences	215 113	115 976	79 087	64 616
Taille (MOYENNE)	4 966,65	3 222,40	4 289,34	4 438,67
Taille (MAX)	144 779	125 229	63 377	63 377
N50	10 247	7 150	7 842	8 019
Nb séquences > 10kb	29 173	8 204	8 135	0
% GC	43,74	40,63	39,74	38,48
Qualité moyenne	8,18	5,46	16,22	17,64

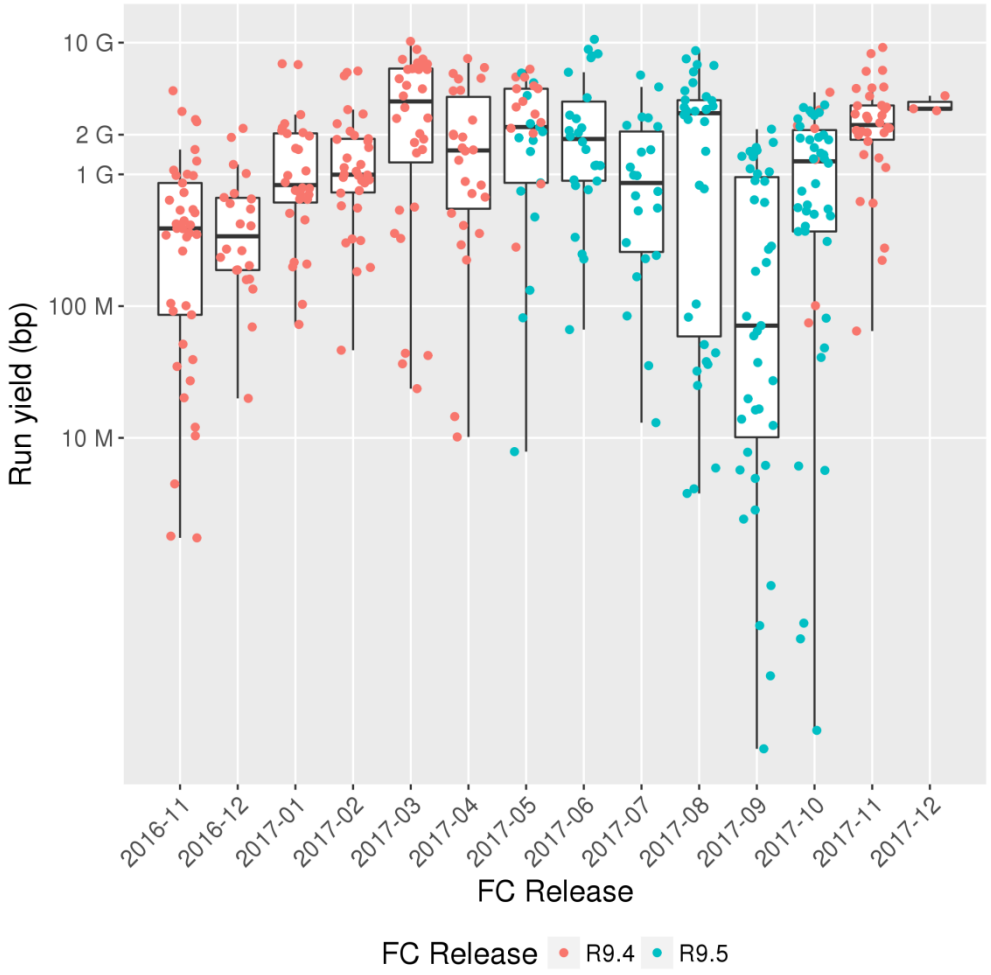
Nanopore : a fast evolving technology

- Yield improvement : ~100Mb to >1Gb but the throughput of R9.5 flowcells seems to be more erratic



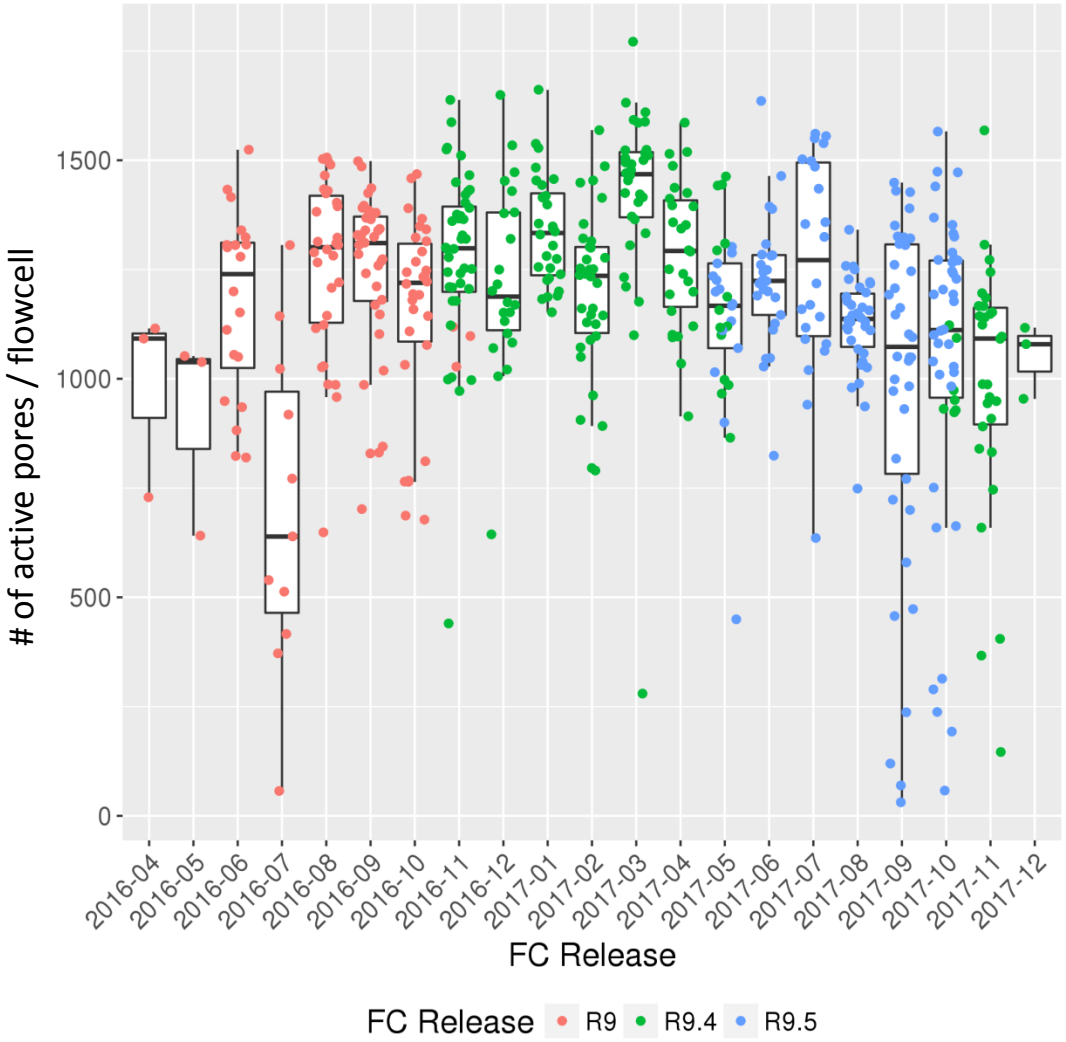
Nanopore : a fast evolving technology

- The throughput dropped off in the last months, we now used R9.4 in production



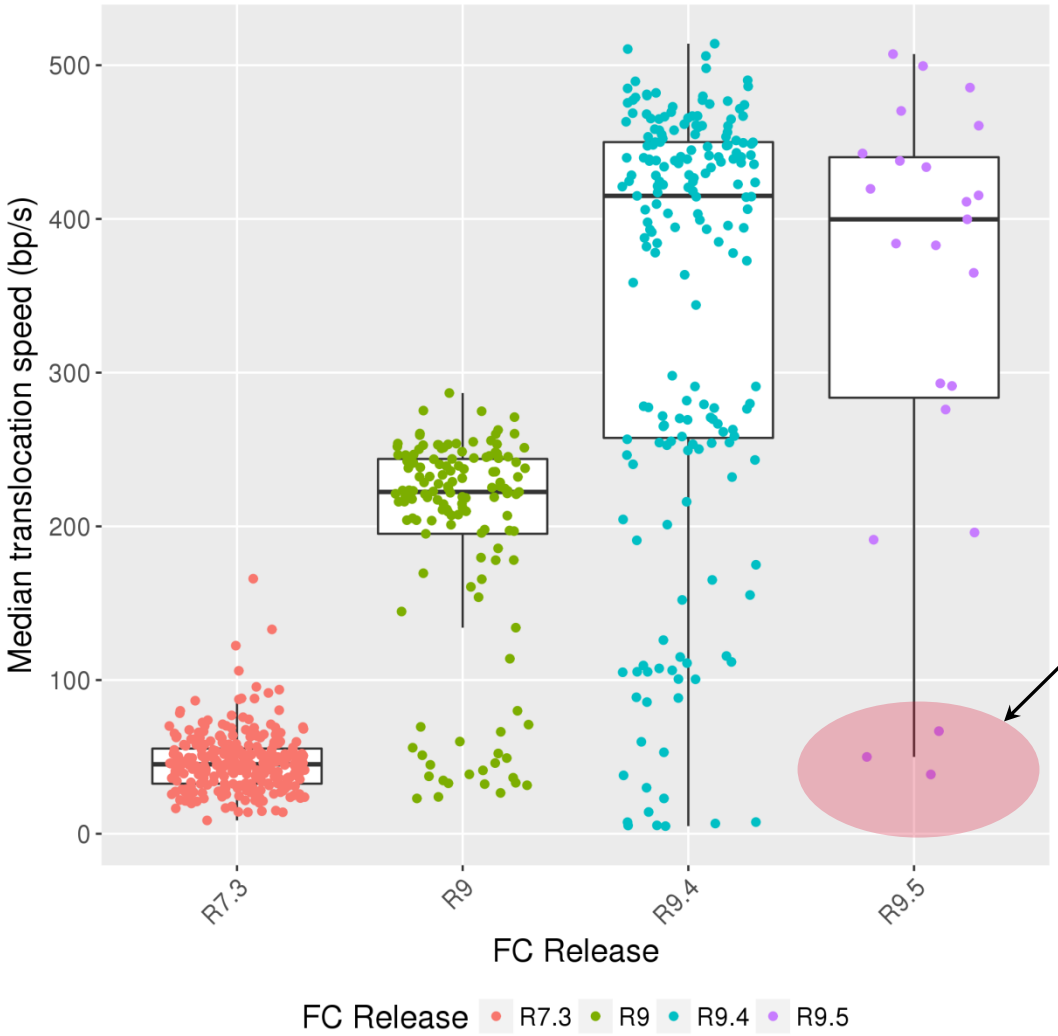
Nanopore : a fast evolving technology

- The flowcell quality seems to be one of the issue



Nanopore : a fast evolving technology

- Improvement of the DNA translocation speed through the pore

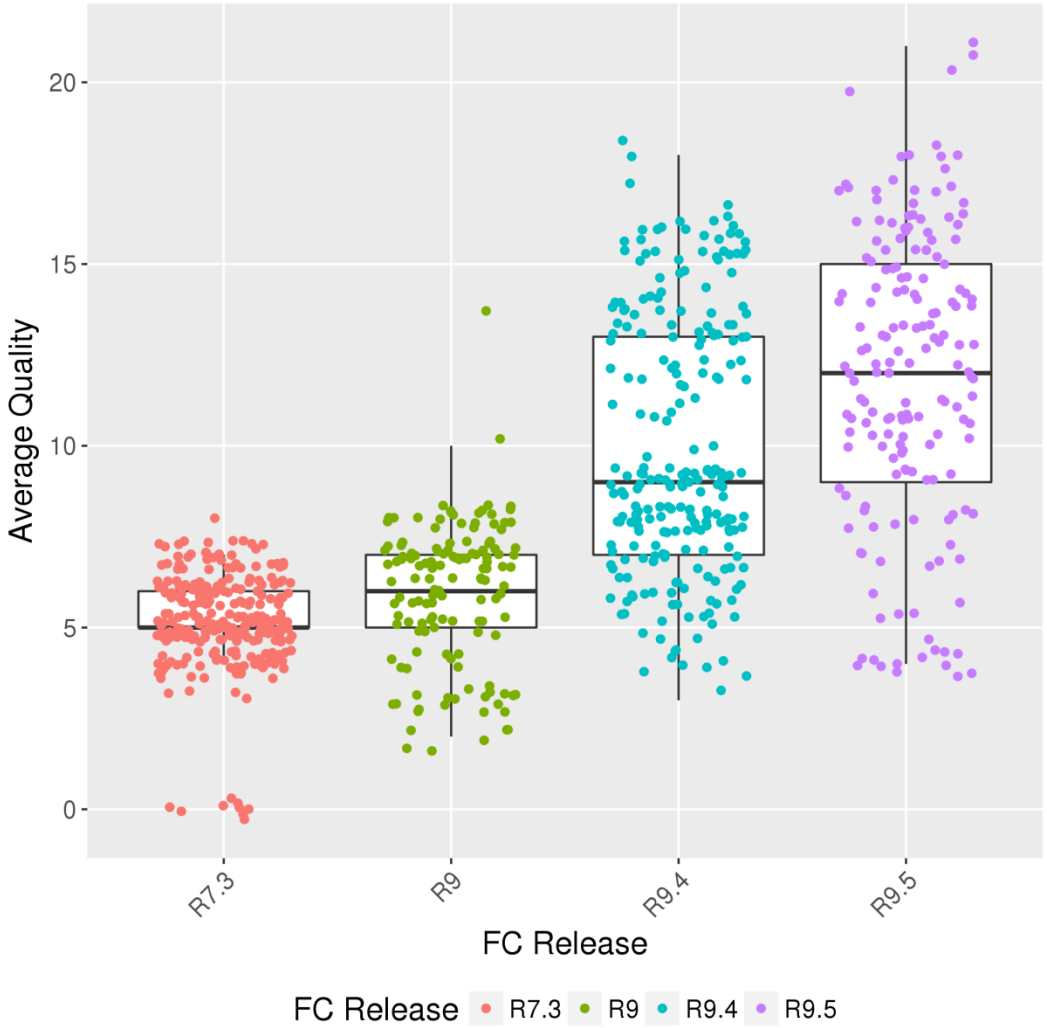


RNA direct sequencing

(~70bp/s)

Nanopore : a fast evolving technology

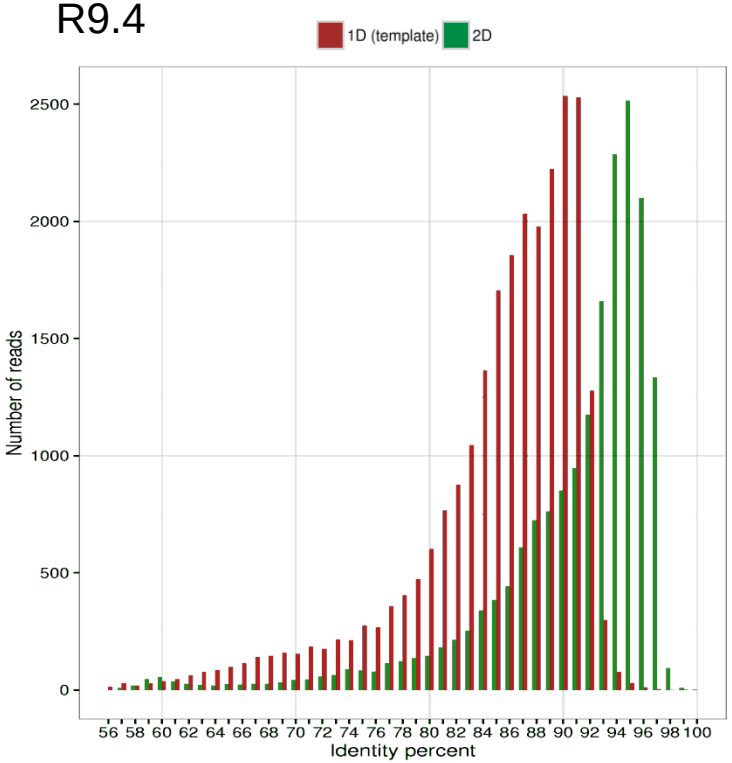
- Average quality and error rate improvement



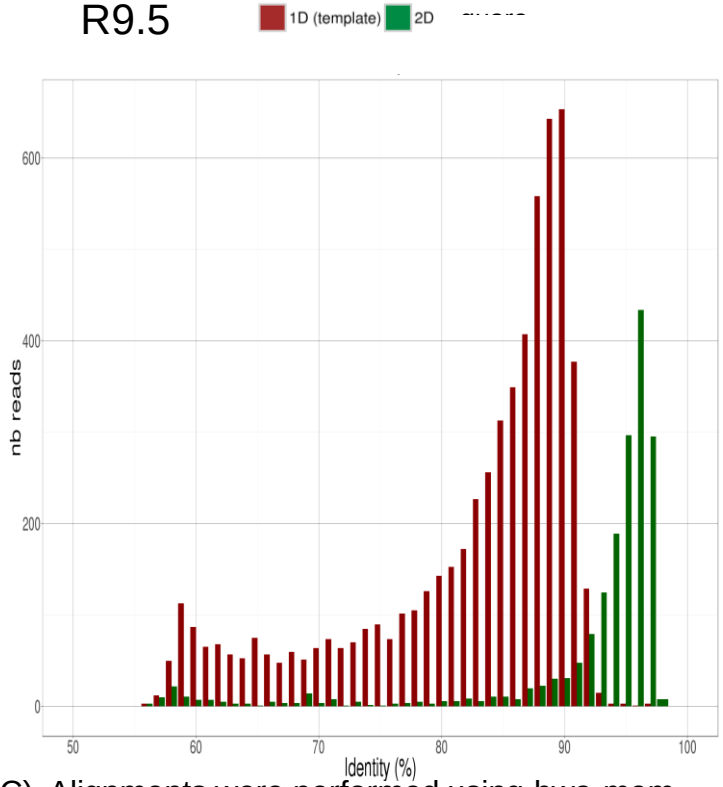
Nanopore : a fast evolving technology

- Today error rate is even lower (in average 14% for 1D reads and <7% for 1D² reads),
  => basecaller is a key component in the error rate drop off

R9.4



R9.5

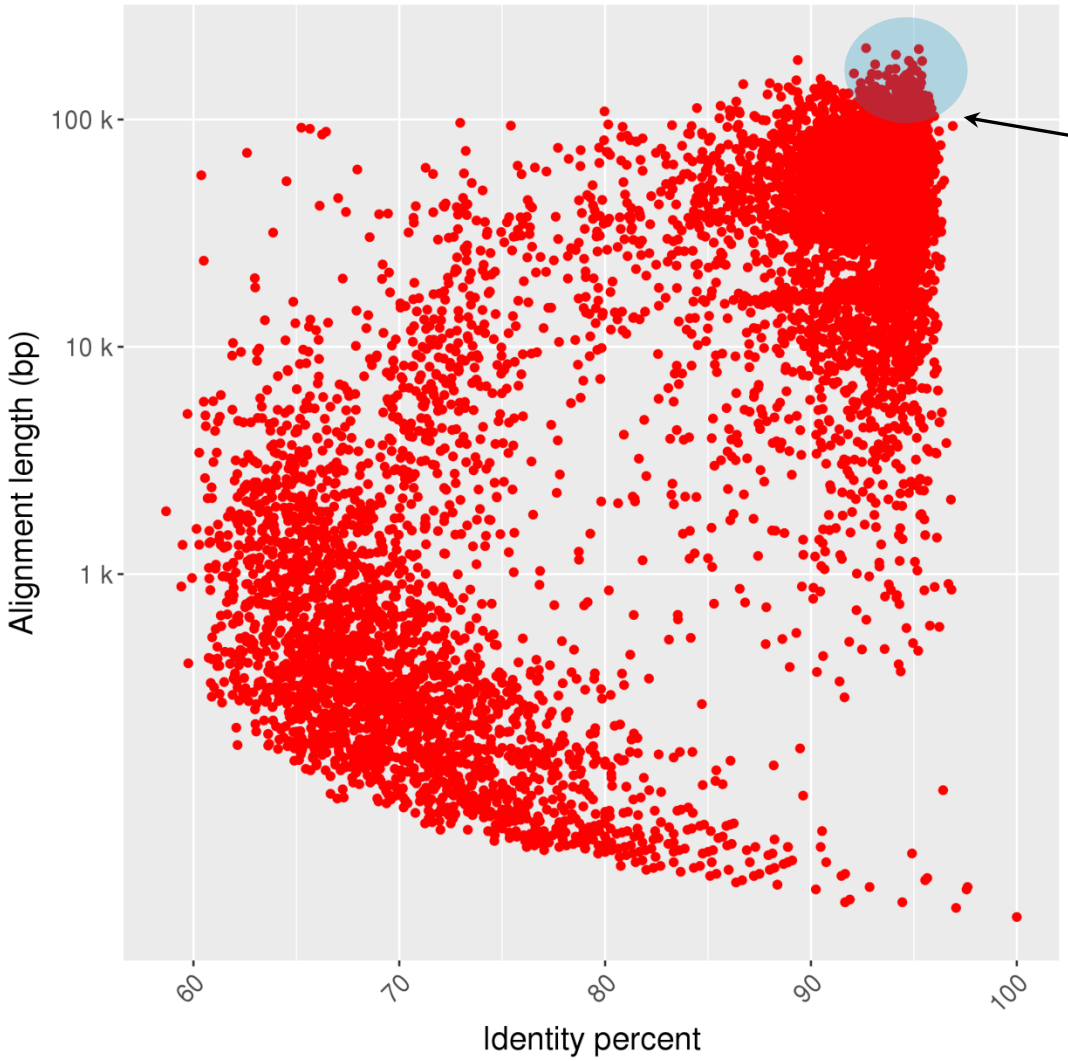


Distribution of identity percent based on yeast reference (S288C). Alignments were performed using bwa-mem

1D² is a real improvement in the error rate, unfortunately we get only up to 30% of 1D² reads

Nanopore : a fast evolving technology

- The device is able to sequence very long DNA fragments (>100Kb)

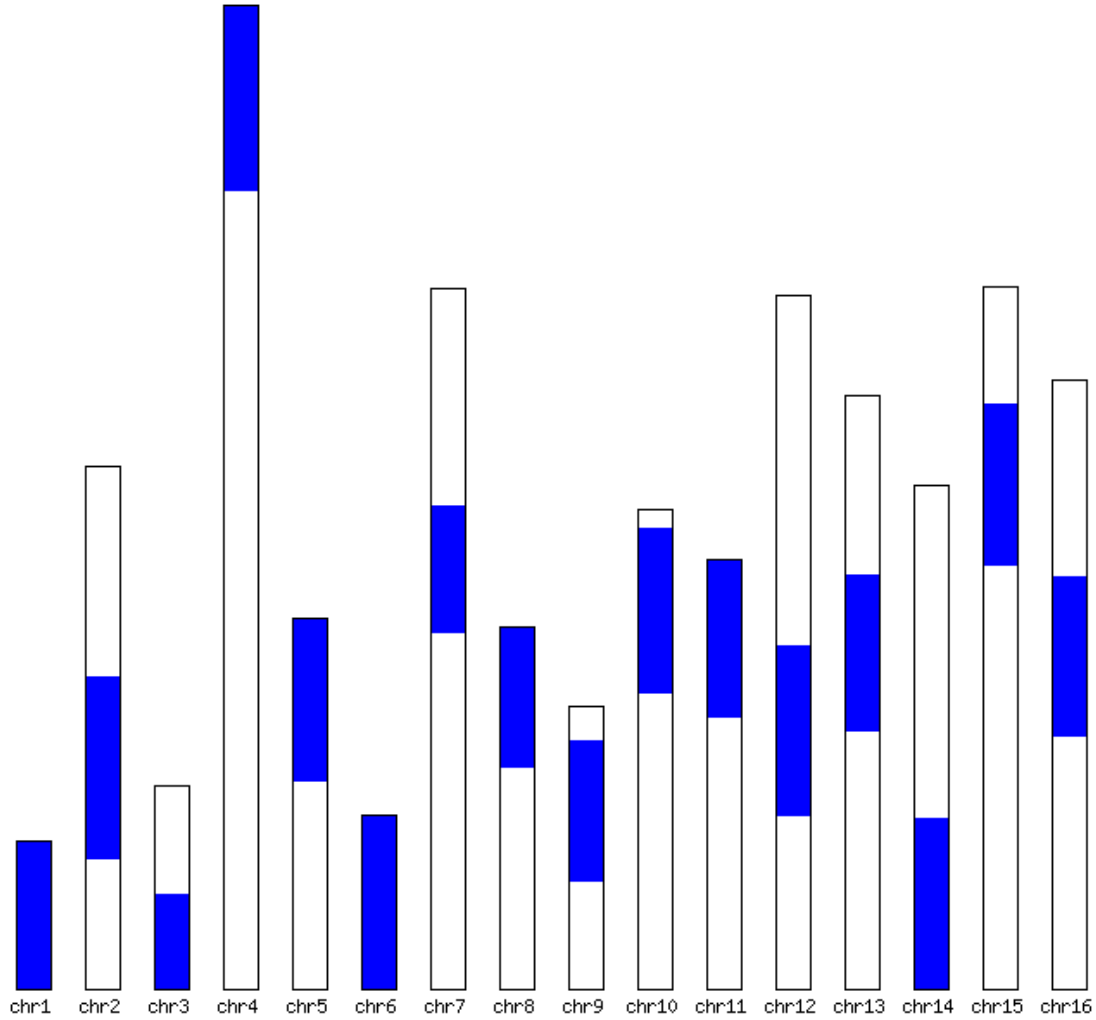


~400 high quality reads with alignment length > 100Kbp

=> ~4X of yeast genome

Nb bases	2 036 675 349
Nb sequences	137 109
Max length (bp)	461 529
N50 (bp)	50 800
Nb seq. > 50kb	11 695
Nb seq. > 100kb	3 275

Nanopore : a fast evolving technology

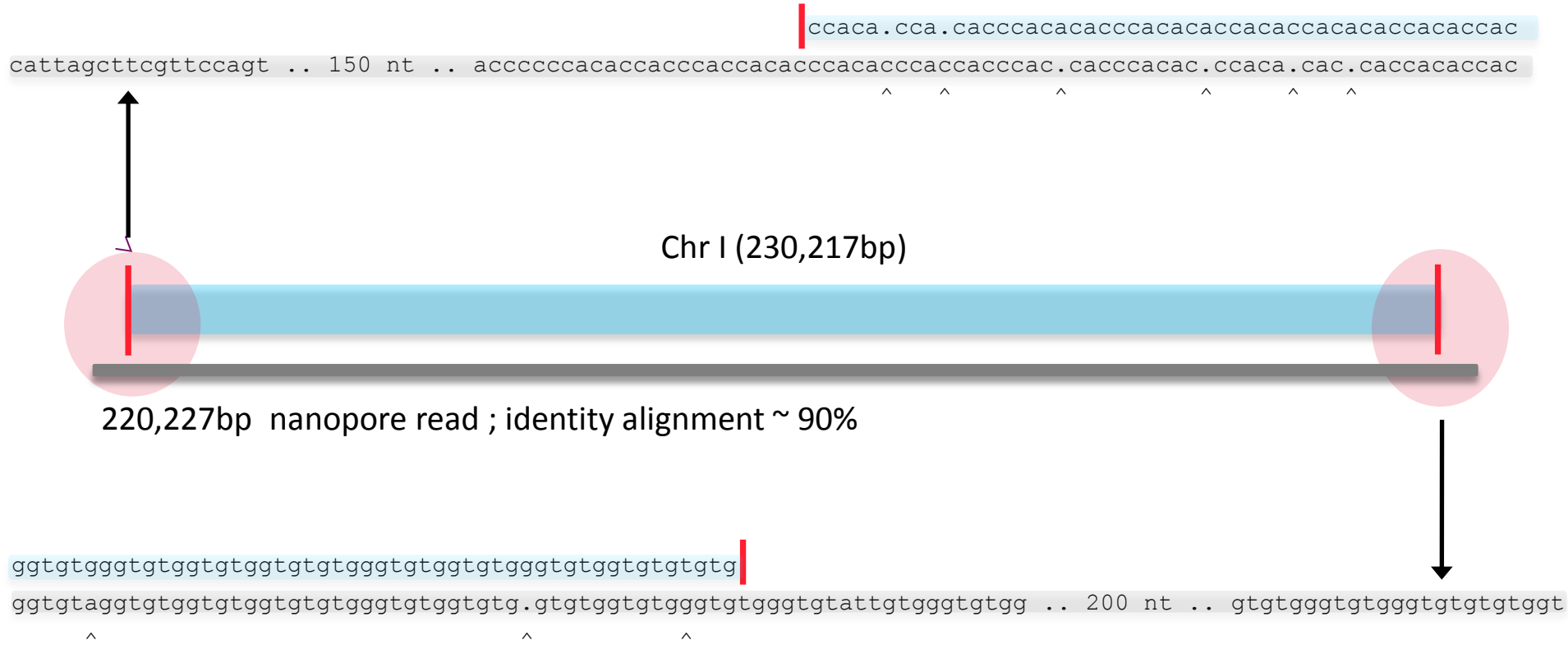


read with the longest alignment for each chromosome

Smallest chromosomes 1 and 6 are obtained in a single nanopore read !

Nanopore : a fast evolving technology

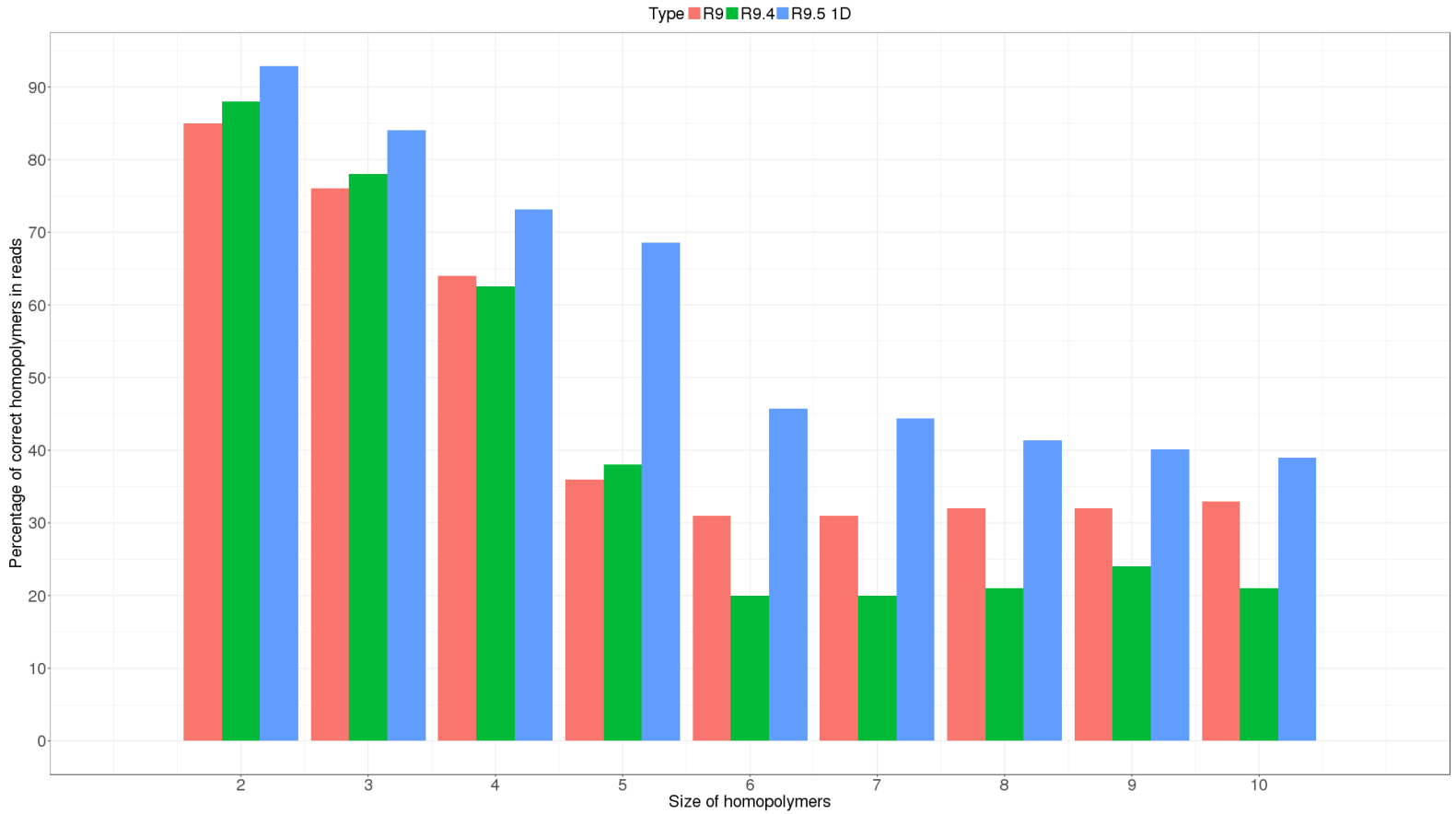
- Chromosomes can be captured entirely, the example read span chromosome 1 from telomere to telomere



- The nanopore read is smaller than the chromosome due to deletions

Nanopore : a fast evolving technology

- High error rate in homopolymers is still an issue for de novo sequencing projects, however the R9.5 release (and scrappie) really improve the basecalling of homopolymers.

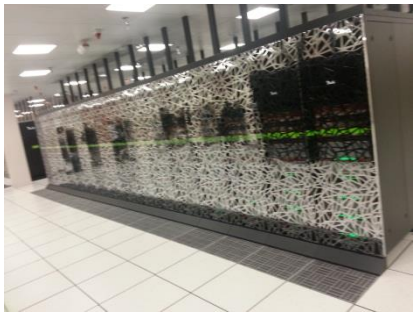


It is still impossible to generate high quality consensus using nanopore only strategy.

DE LA RECHERCHE À L'INDUSTRIE

www.cea.fr

cDNA-Seq and RNA-Seq using the Oxford Nanopore technology



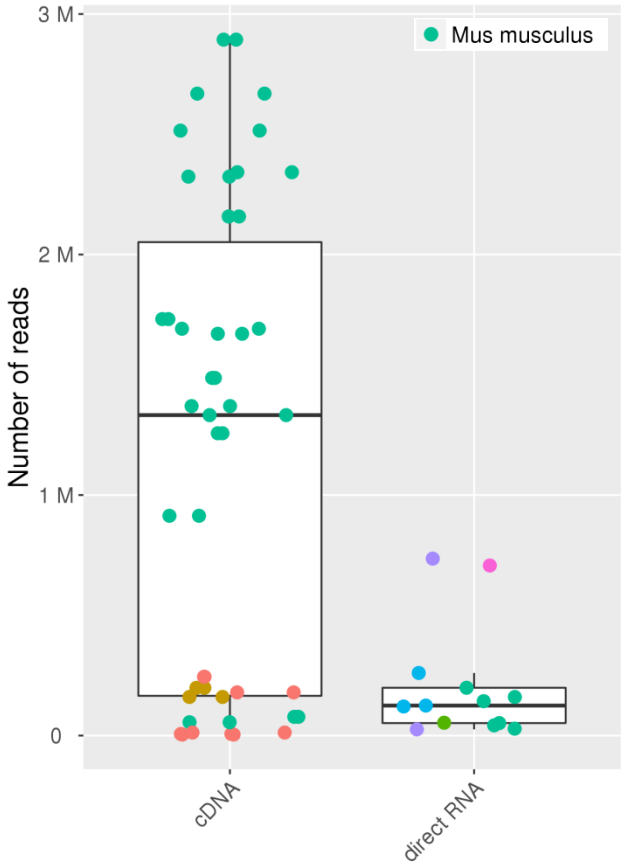
Jean-Marc Aury

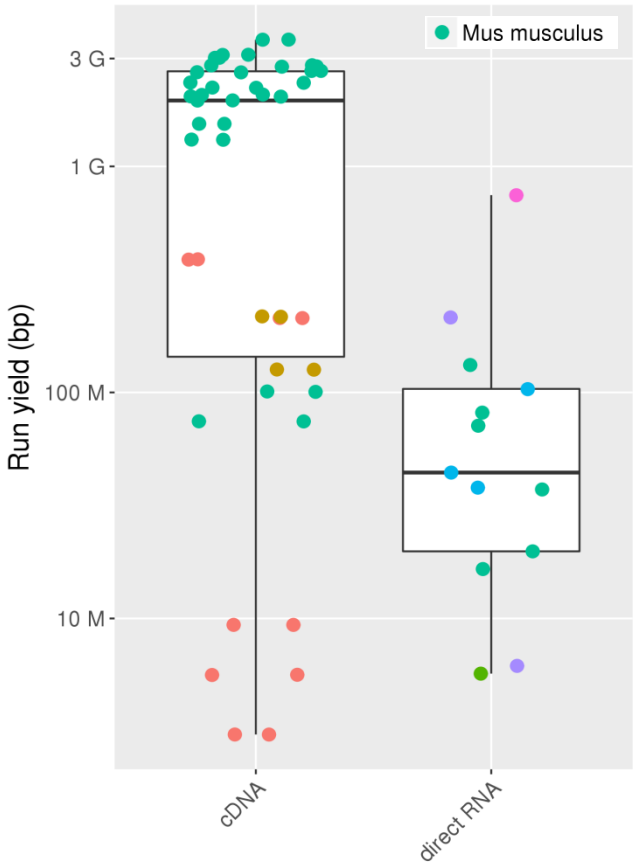
jmaury@genoscope.cns.fr

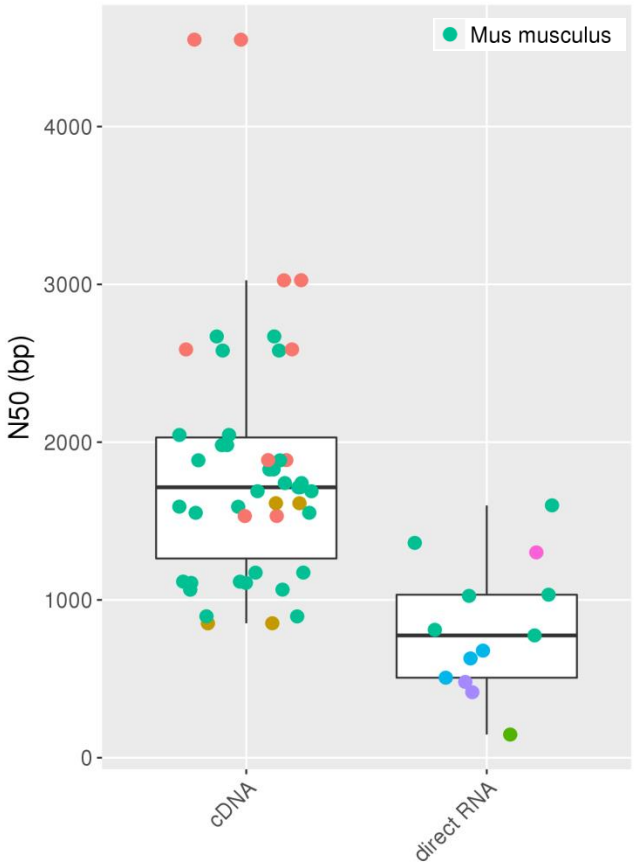
@J_M_Aury

Comparison Nanopore / Illumina

A typical cDNA-Seq experiment generates around 2M of reads, in comparison RNA-Seq experiments generate less reads (450bp/s vs 70bp/s)







Nanopore : a fast evolving technology

Dataset used to perform comparisons

Brain sample

FC release	R9.4
Nb sequences	1 256 967
Nb bases	2 074 348 139
N50 (bp)	1 885

Brain sample

FC release	R9.5
Nb sequences	160 450
Nb bases	81 508 561
N50 (bp)	1 033

Brain sample

FC release	HiSeq 4000
Nb sequences	59M
Nb bases	17Gb
N50 (bp)	150

Liver sample

FC release	R9.4
Nb sequences	1 369 927
Nb bases	1 956 452 499
N50 (bp)	1 591

Liver sample

FC release	R9.5
Nb sequences	198 708
Nb bases	131 963 731
N50 (bp)	1 026

Liver sample

FC release	HiSeq 4000
Nb sequences	45M
Nb bases	13Gb
N50 (bp)	150

cDNA sequencing

Direct RNA

Nanopore : a fast evolving technology

Mapping of reads against RefSeq genes (refseq109) and the mouse genome (GRCm38)

Alignment against GRCm38 using minimap2 (36 cores)

	Number of reads	Mapped reads	Mapped bases (of aligned reads)	Elapsed time (sec)
1D cDNA	1 256 967	90.7%	89.6%	396
RNA direct	160 450	33.8%	82.8%	20

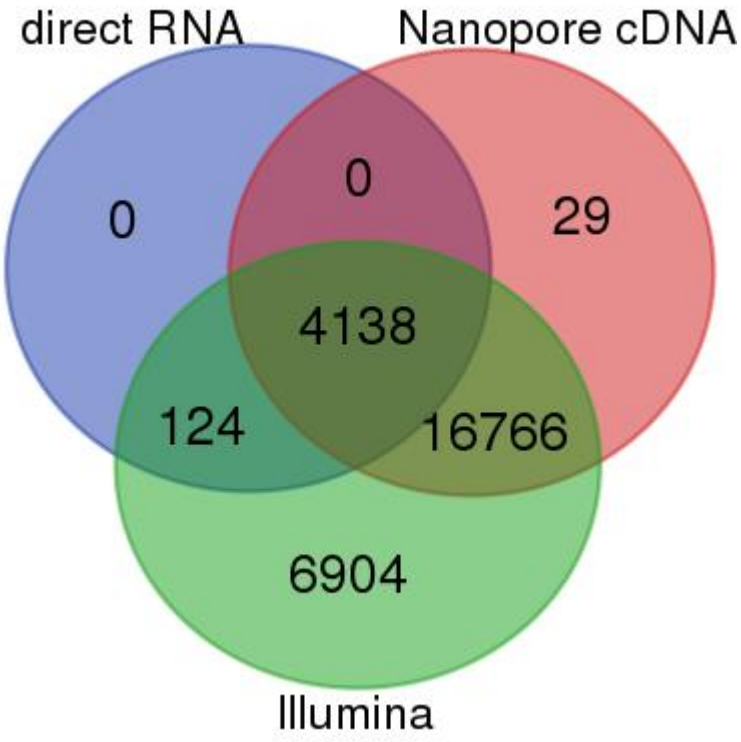
Alignment against RefSeq 105 using bwa-mem (8 cores)

	Number of reads	Mapped reads	Mapped bases (of aligned reads)	Elapsed time (sec)	rRNA	Mitochondrial
1D cDNA	1 256 967	84.7%	64.2%	4 481	21.6%	15.8%
RNA direct	160 450	25.9%	75.2%	65	0.1%	18.5%

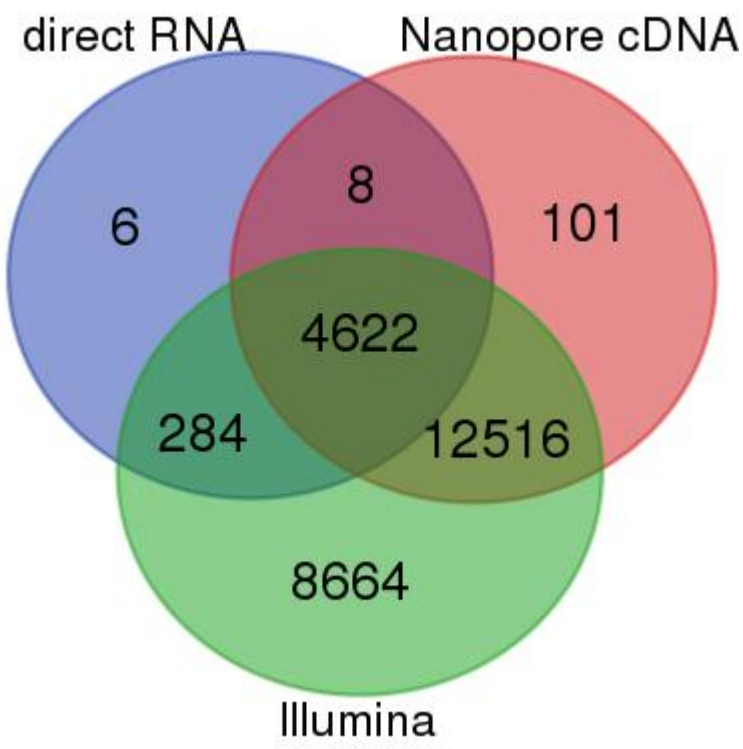
Comparison Nanopore / Illumina

Number of RefSeq genes seen by each sequencing technology

Brain sample

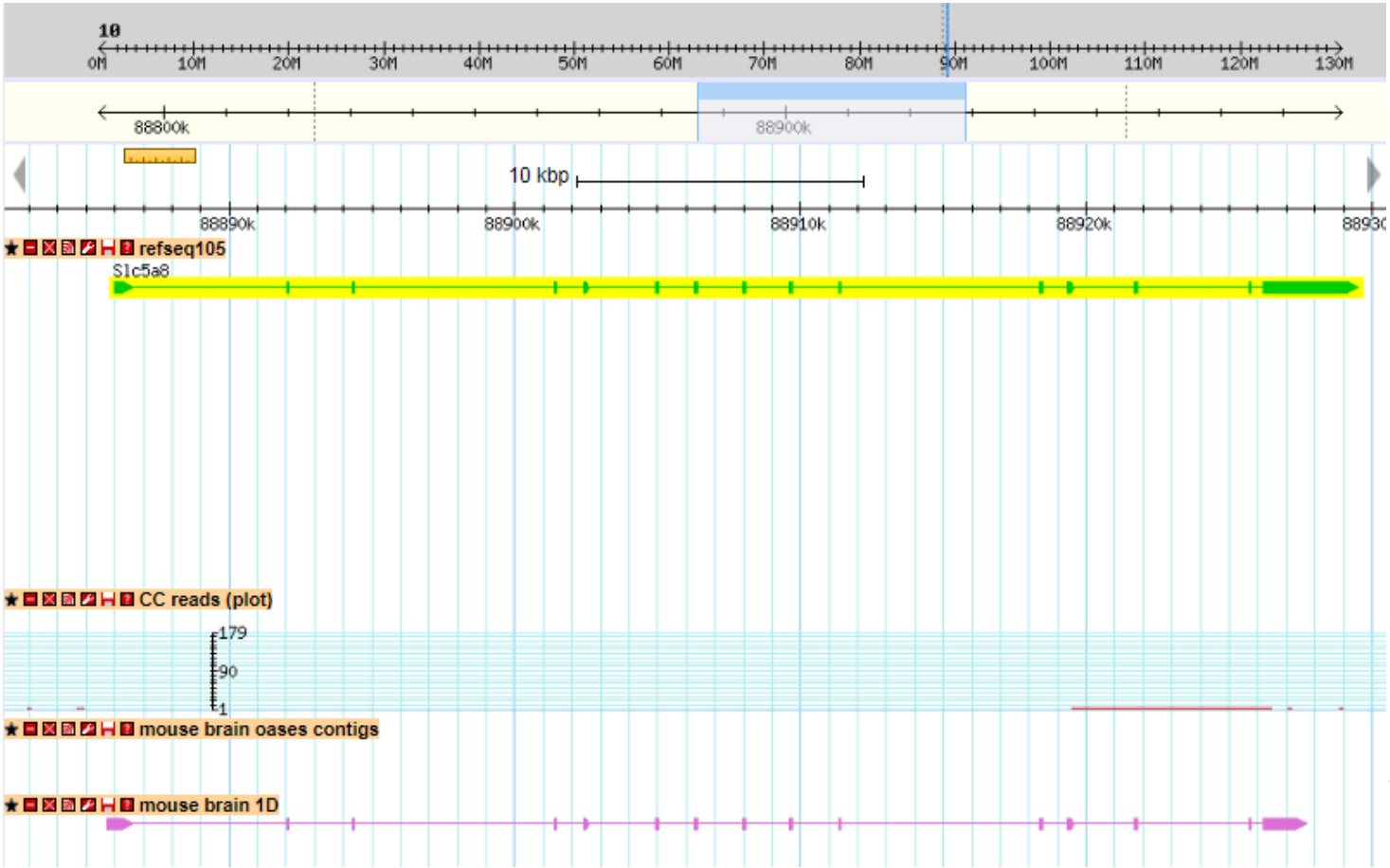


Liver sample



Comparison Nanopore / Illumina

A gene can be covered entirely by a single read

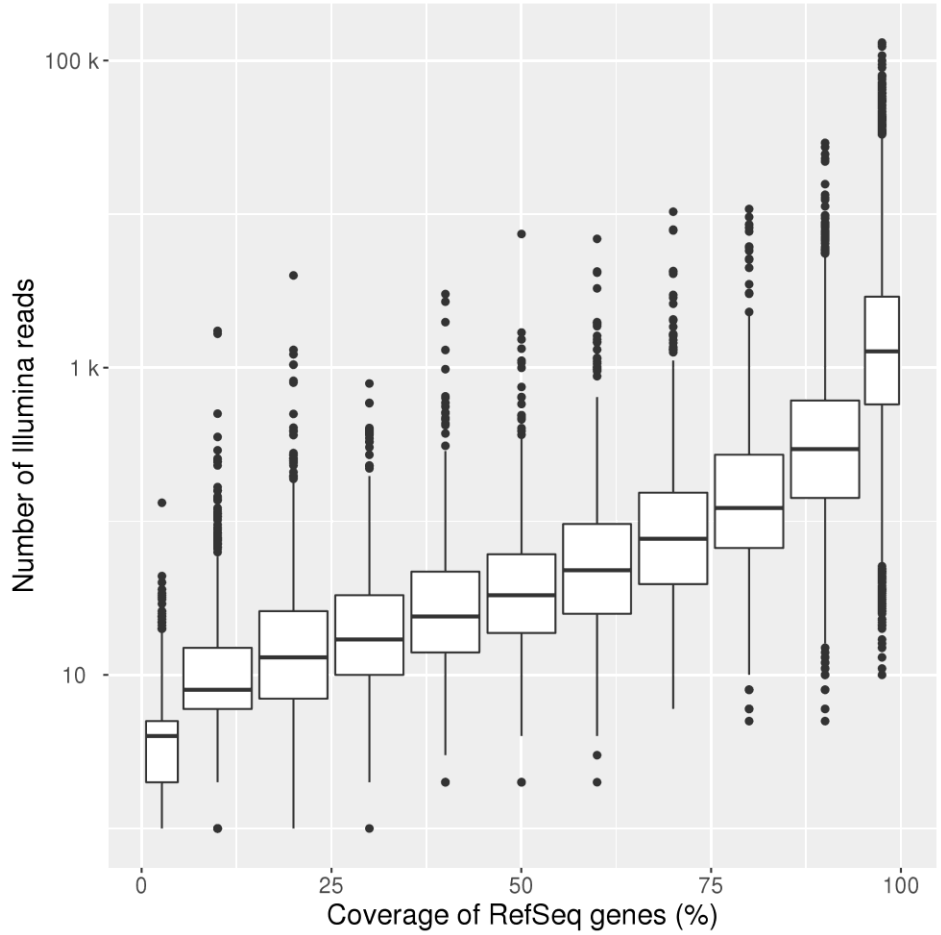


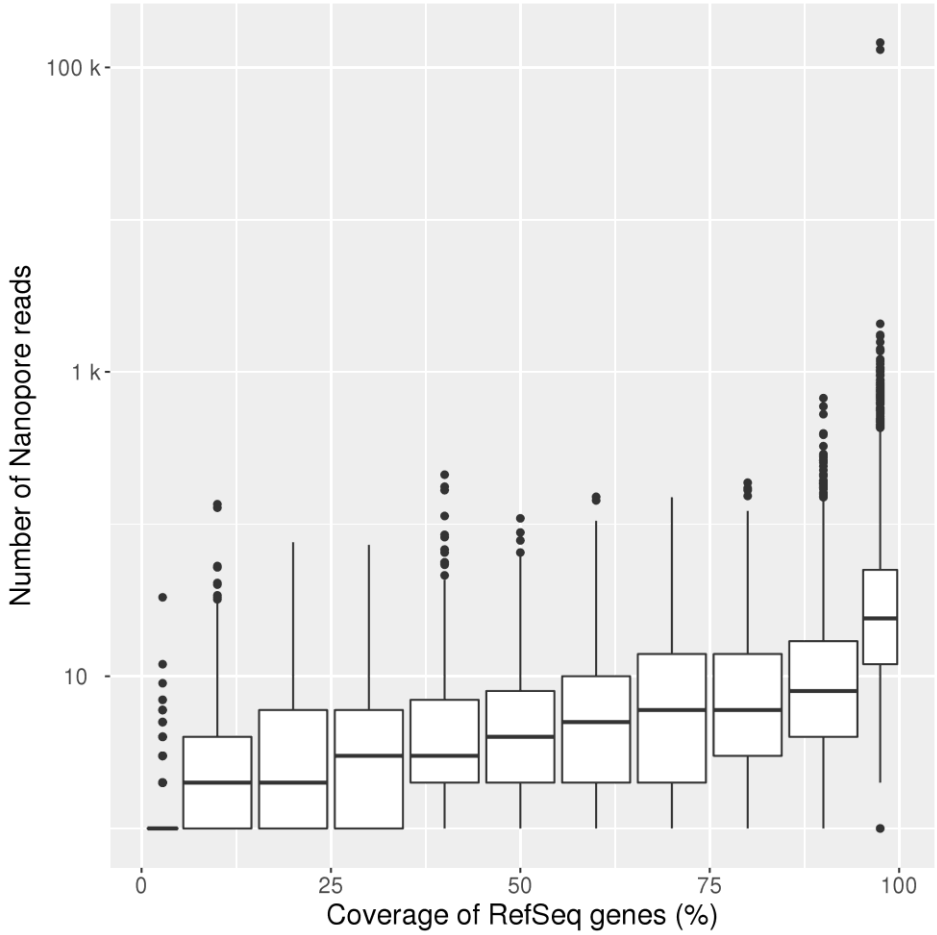
50 illumina reads are aligned and partially cover the gene

The entire gene is covered by a single nanopore read

Comparison Nanopore / Illumina

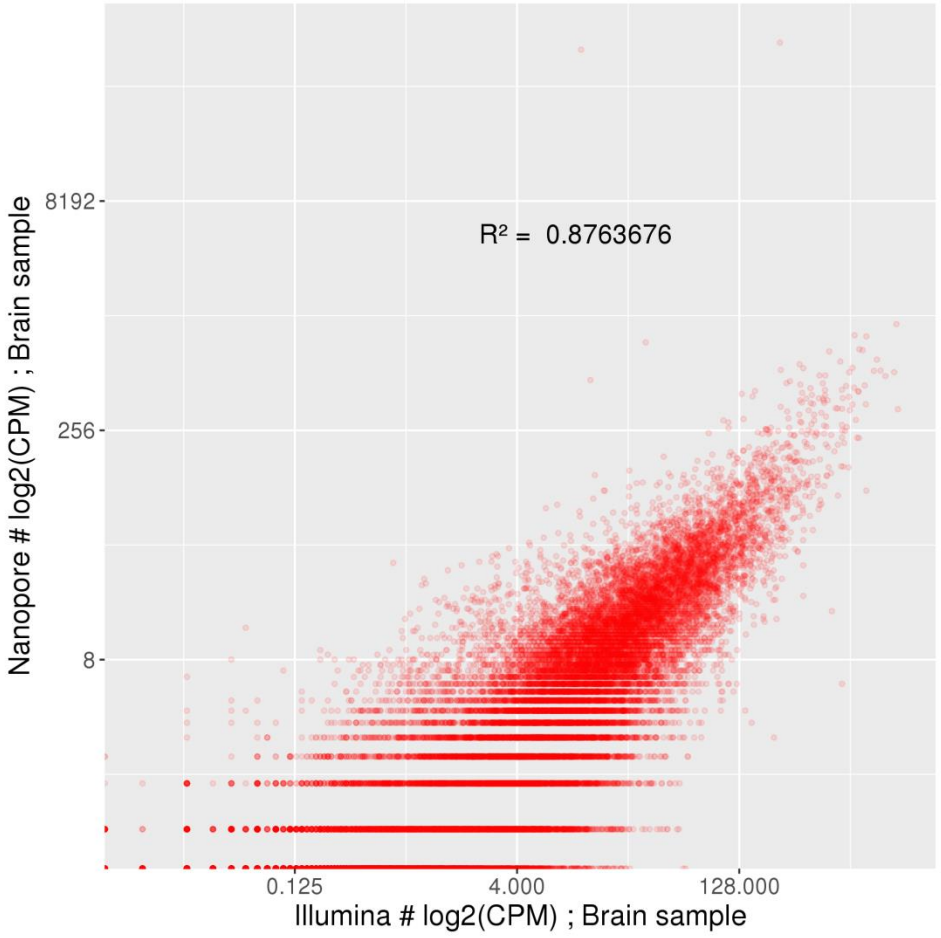
As expected, less nanopore reads are needed to cover RefSeq genes, when we need at least 500 illumina reads to cover 75% of a given gene, 10 nanopore reads are sufficients

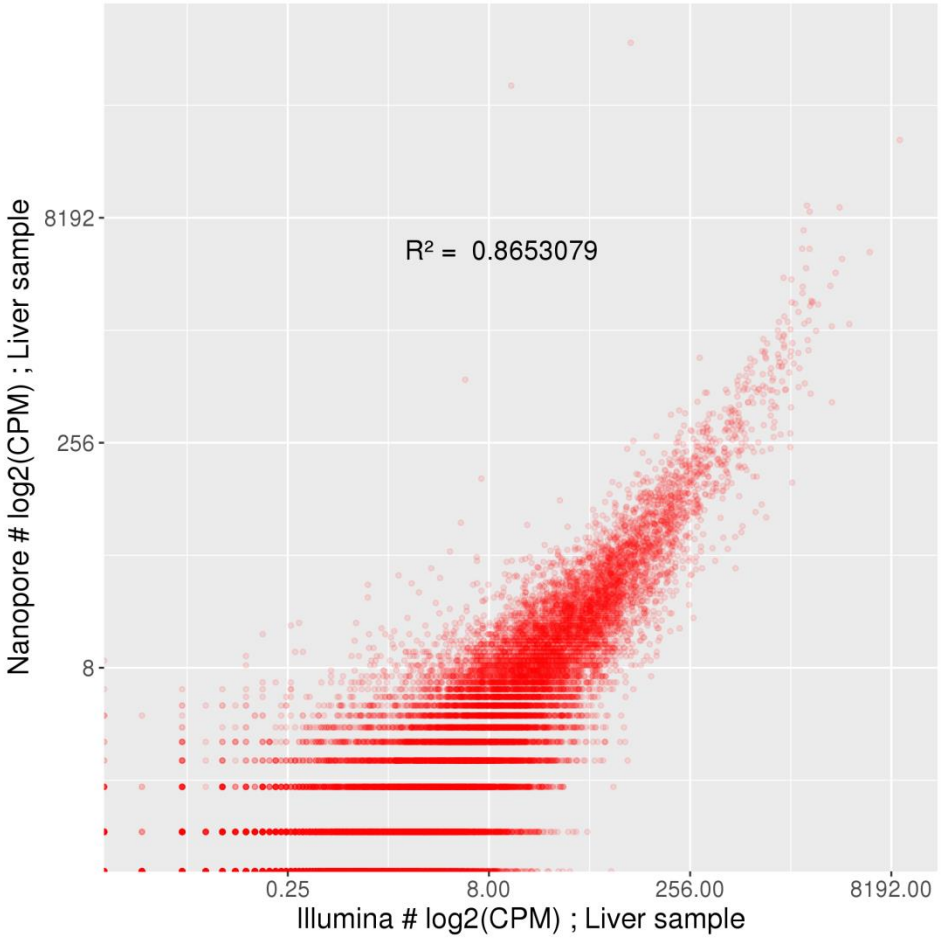




Comparison Nanopore / Illumina

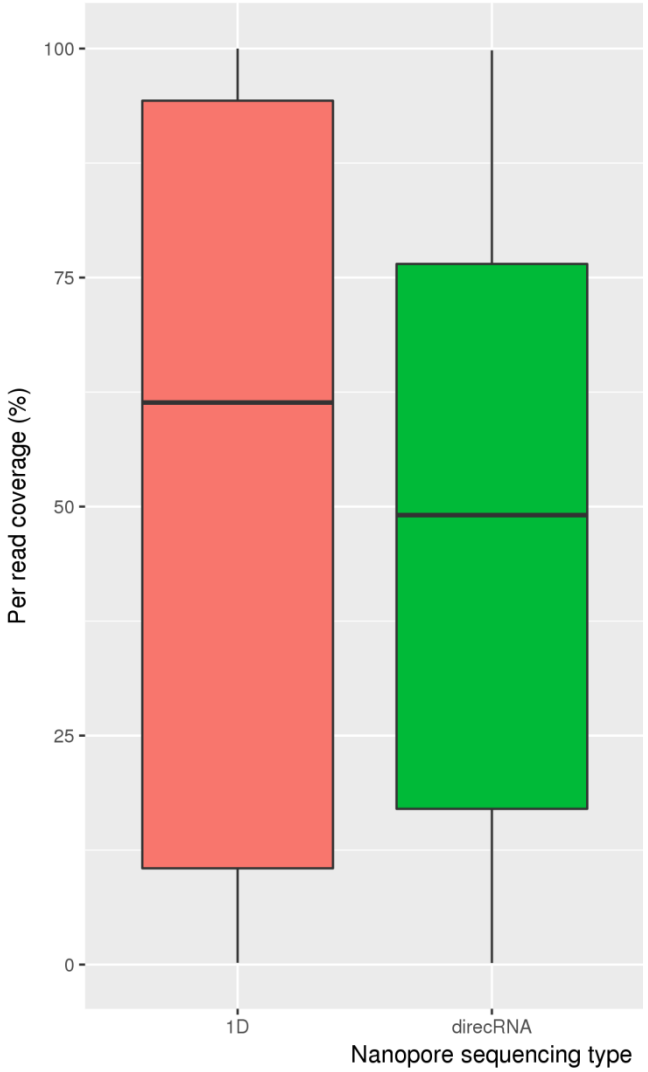
Expression levels (brain and liver samples) are correlated between Illumina and Nanopore experiments





Are all reads full-length RNA ?

A small proportion of reads are full-length RNA, in average a cDNA and RNA read cover 55% and 47% respectively of a RefSeq gene



TeloPrime amplification kit

We tested the TeloPrime amplification kit from Lexogen

Based on Lexogen´s unique Cap-Dependent Linker Ligation (CDLL) and long reverse transcription (long RT) technology, it is highly selective for full-length RNA molecules that are both capped and polyadenylated.

2 sequencing runs from brain and liver samples

Brain sample

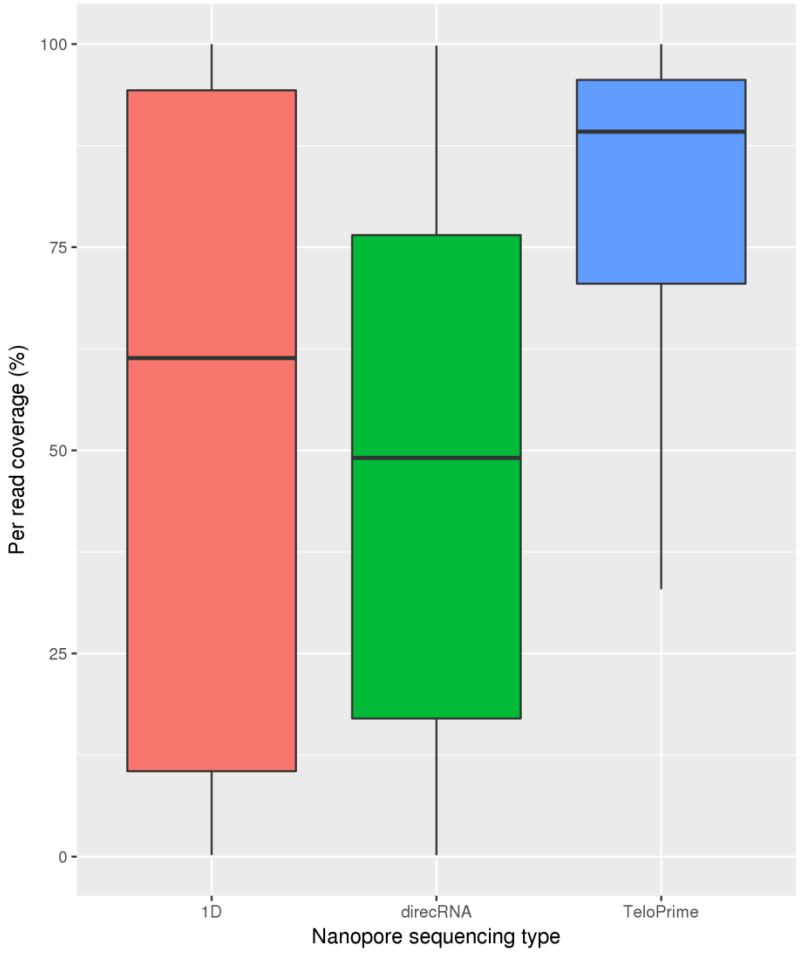
FC release	R9.5
Nb sequences	2 668 975
Nb bases	2 641 896 941
N50 (bp)	1 116

Liver sample

FC release	R9.5
Nb sequences	1 691 454
Nb bases	1 312 184 503
N50 (bp)	896

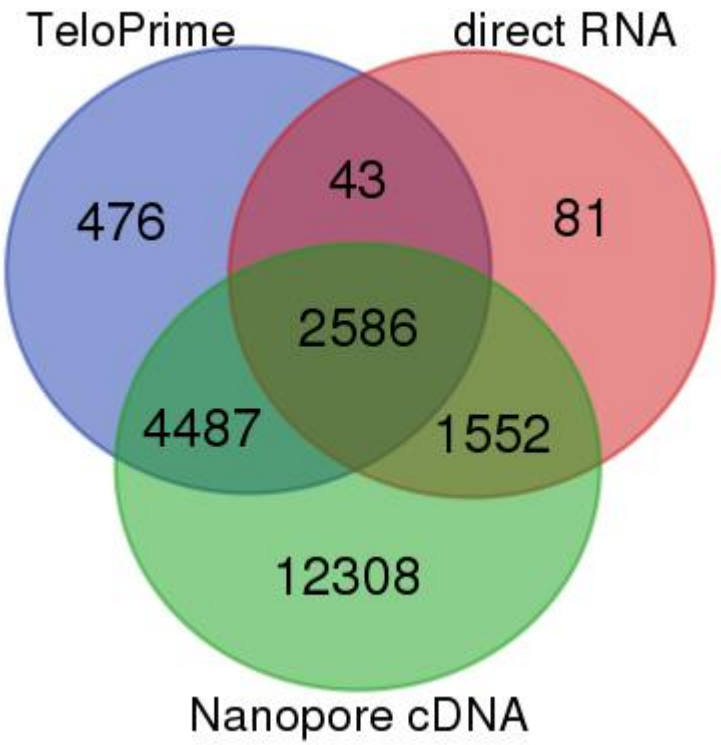
TeloPrime amplification kit

TeloPrime reads better cover RefSeq genes, compared to cDNA and RNA sequencing. in average a TeloPrime read cover 80% of a RefSeq gene



TeloPrime amplification kit

Even with a higher number of reads, TeloPrime reads spread over a limited number of genes (~8k vs ~21k using 1D protocol)



Nanopore cDNA

FC release	R9.4
Nb sequences	1 256 967
Nb bases	2 074 348 139
N50 (bp)	1 885

Direct RNA

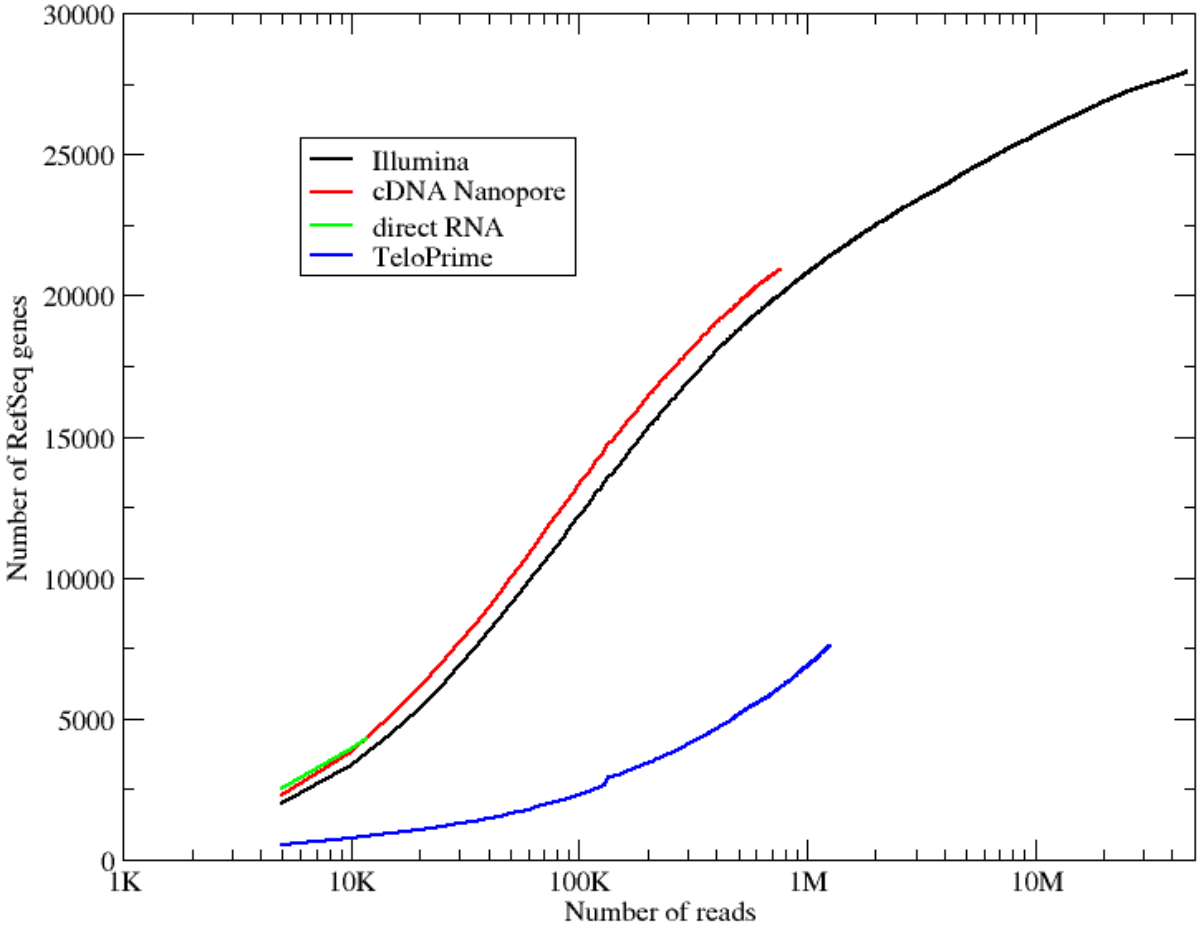
FC release	R9.5
Nb sequences	160 450
Nb bases	81 508 561
N50 (bp)	1 033

TeloPrime

FC release	R9.5
Nb sequences	2 668 975
Nb bases	2 641 896 941
N50 (bp)	1 116

TeloPrime amplification kit

We need to sequence at a higher depth with the TeloPrime amplification kit to be able to catch a high proportion of RefSeq genes



Conclusion

- Today the throughput of the MinION device is sufficient for profiling eukaryotic gene expression, gene prediction can take advantage of long reads to avoid transciptome assembly

- The potential of the device to sequence long reads is impressive, sequencing of large eukaryotic genomes is now possible even with the MinION device

- Error rate is acceptable for de novo sequencing projects (a high proportion of reads with less than 10% of errors), still an issue with homopolymers

- Need to improve the “wetlab part” to increase the proportion of full-length reads, TeloPrime kit seems to bring a real improvement

Acknowledgements

R&DBioSeq Team
www.genoscope.cns.fr/rdbioseq

jmaury@genoscope.cns.fr

@J_M_Aury

- Genoscope labs
  - Bioinformatic : Corinne Da Silva, Stefan Engelen, Benjamin Istace and Marion Dubarry
  - Nanopore Sequencing : Corinne Cruaud, Odette Beluche, Emilie Payen, Thomas Guérin and Arnaud Lemainque
- Members of the ASTER project
- Funding agencies : CEA, Genoscope, France Génomique and ANR