Analysis of new methods of sequencing applied to metagenomics

Math4genomics – 03/06/2020

Why use new sequencing methods in metagenomics ?

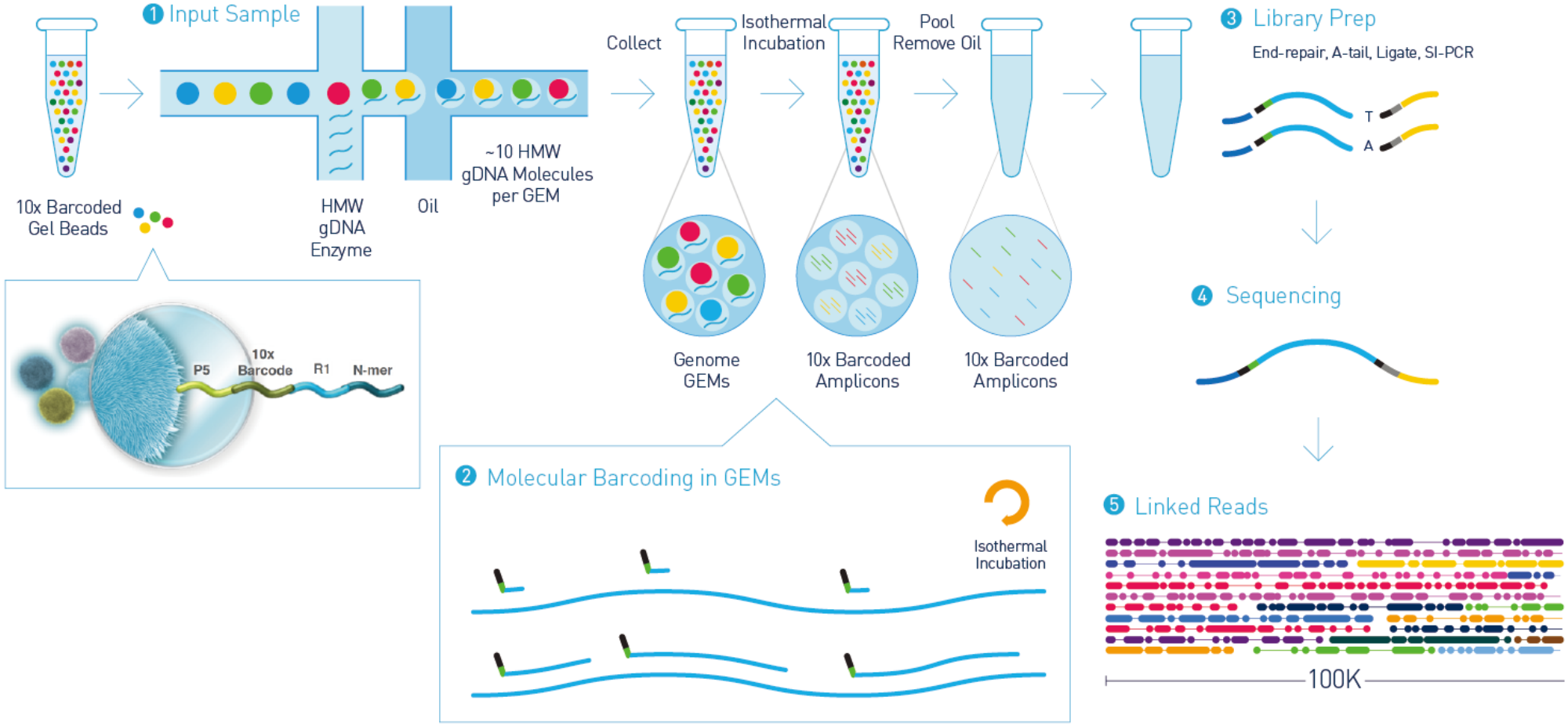
- Generally, the sequencing methods used is short-read sequencing (100-150nt), then these reads are mapped on reference genes
  - → only quantitative information provided
  - → not all the reference genes are associated to a taxonomy

- New sequencing methods to reconstruct genomes
  - → reconstruction of unknown genomes
  - → identify structural variations between strains hosted by different individuals

enterome
bioscience

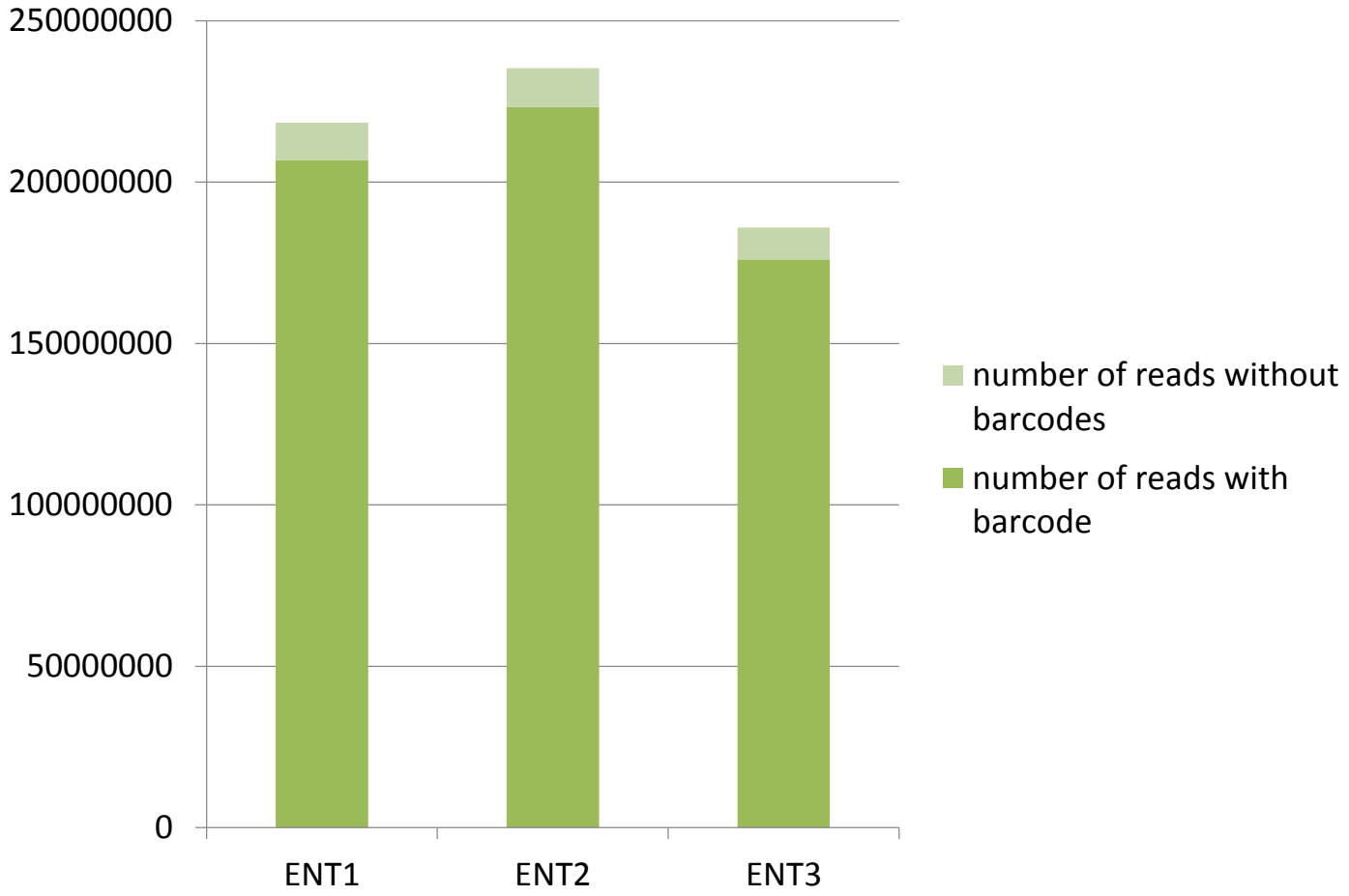
10X technology

10X linked reads technology overview

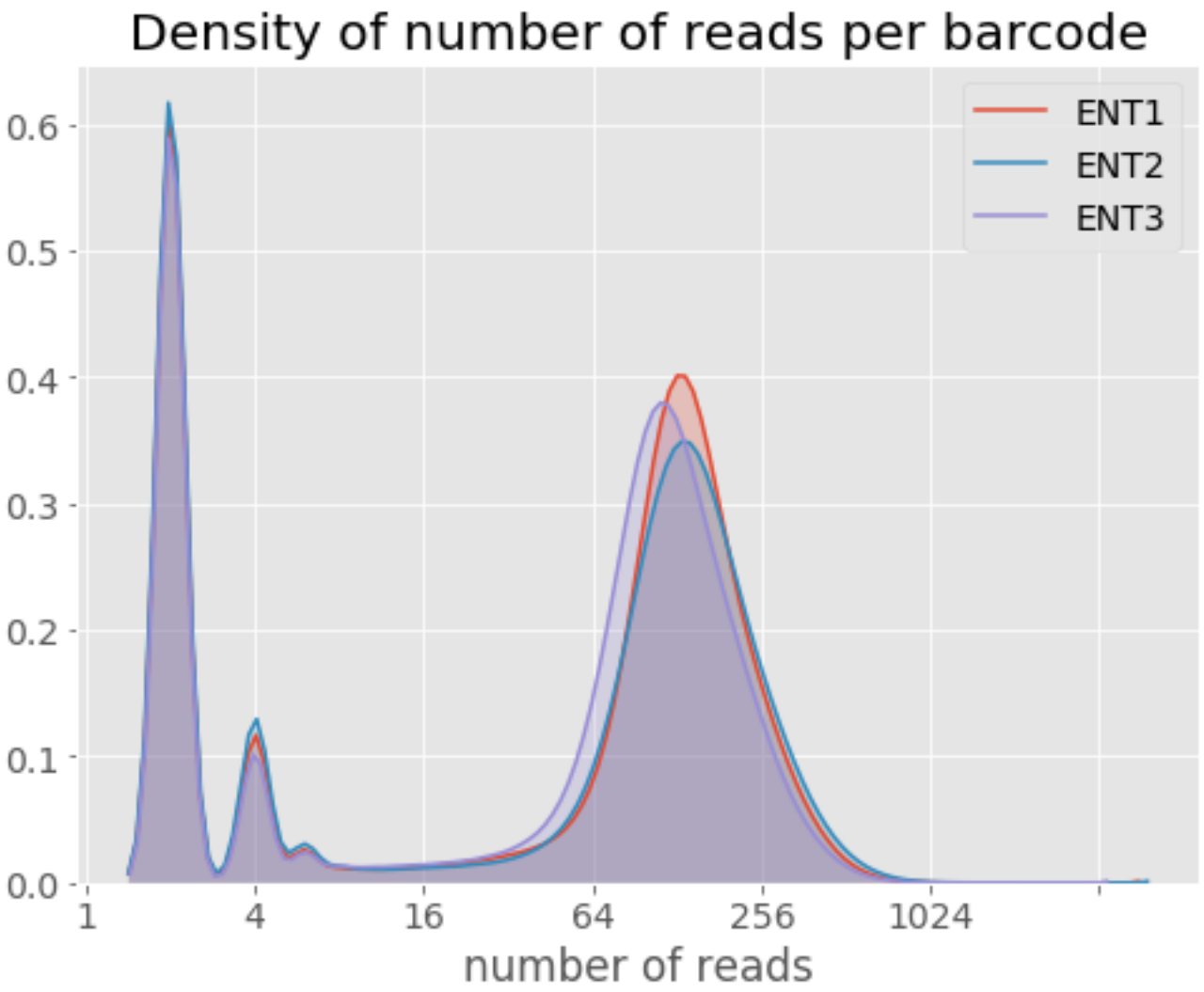


Samples preparation

- 3 samples from feces from 3 healthy donors : ENT1, ENT2 and ENT3
- Number of unique barcodes : from 1.7 to 2 millions



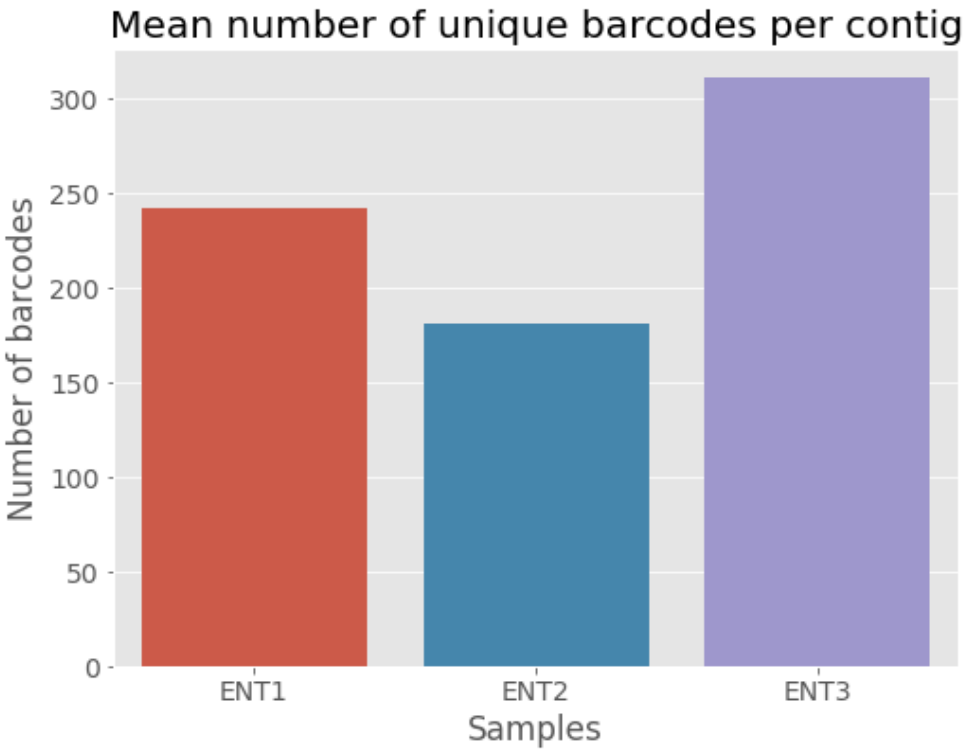
Barcode analysis

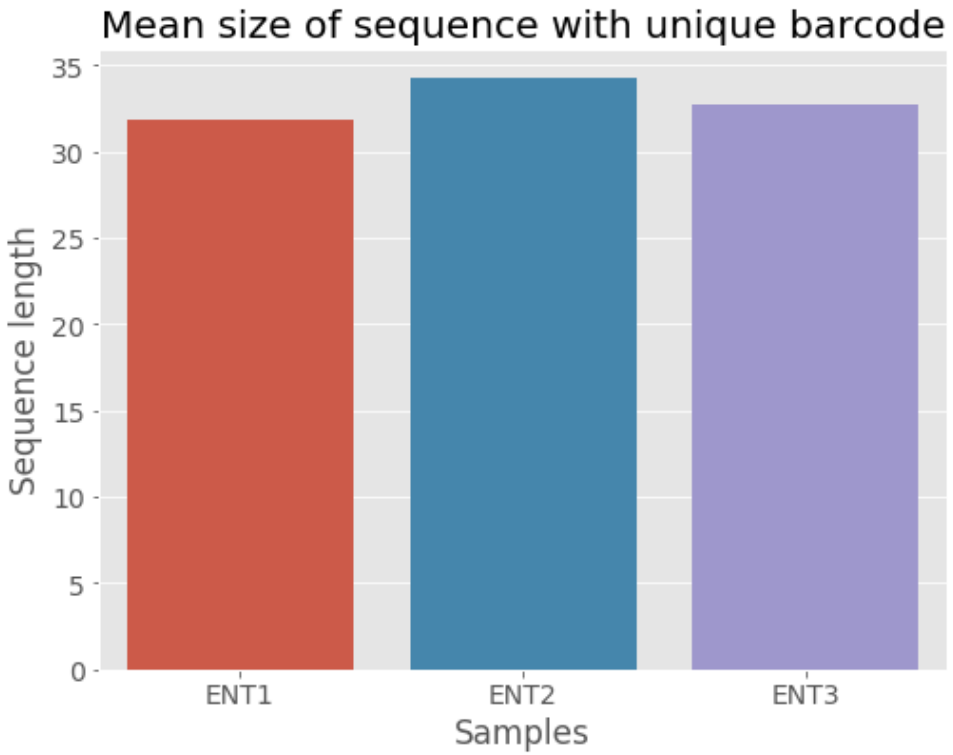


→ Approximately, 25% of barcodes are represented by one pair of reads
→ In general, we have 60 to 250 reads for one barcode

Barcode analysis

- Assembly of reads with Megahit (without the barcode information) then mapping reads on obtained contigs
- For each contig, the number of unique barcode represented is counted :





→ On average, we found one new barcode every 30-35 base.
→ This reflects the fact that multiple similar fragments have been assembled in the same contig

Assembly process

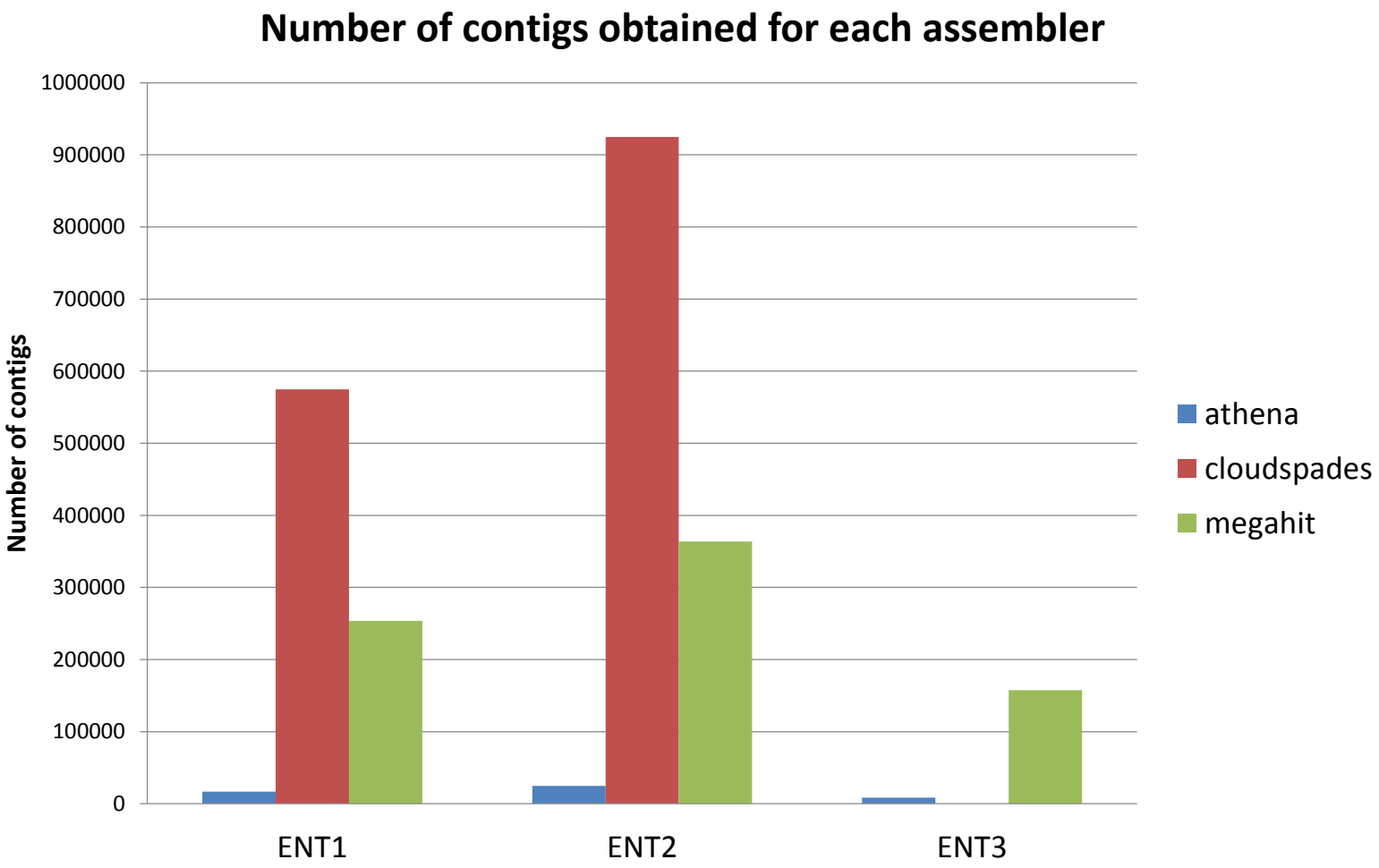
- 3 tested tools :
  - Megahit → classic assembly tool that does not use the information of barcodes
  - Athena and Cloudspades → specific assembly tools that use the information of barcodes

- Cloudspades successfully run for 2 samples

- Athena managed to assembly all the samples but is time consuming

  → approximately 2 days for one run on large computing machine (30 CPUs, 40 Gb of RAM)

Assembly analysis

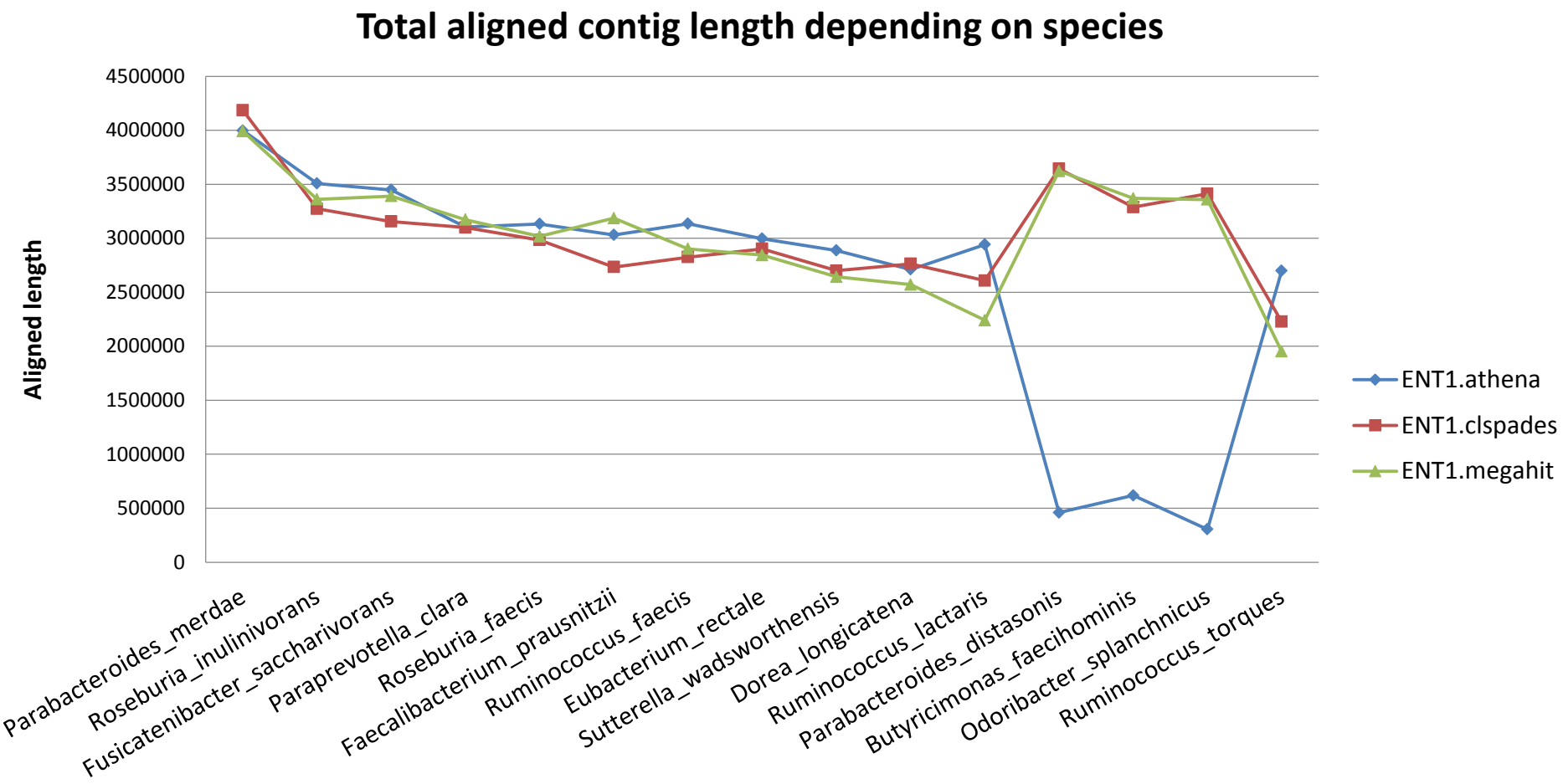


Metaquast comparison : ENT1

Assembly	ENT1.athena	ENT1.cloudspades	ENT1.megahit
Number of contigs (>500 bp)	**16 738**	**121 389**	**143 184**
Total length	**211 362 115**	**331 941 134**	**330 432 418**
Total aligned length	48 771 118	63 743 109	64 945 104
Unaligned length	**162 275 467**	**264 970 470**	**263 926 727**
Genome fraction (%)	**48,7**	**64**	**63,6**
Largest contig	1 021 174	876 749	687 903
Largest alignment	221 614	204 217	151 089
# misassembled contigs	855	1 144	2 548

→ The genome fraction is lower for Athena but the contigs obtained are longer for a reduced number of contigs

Metaquast comparison : ENT1



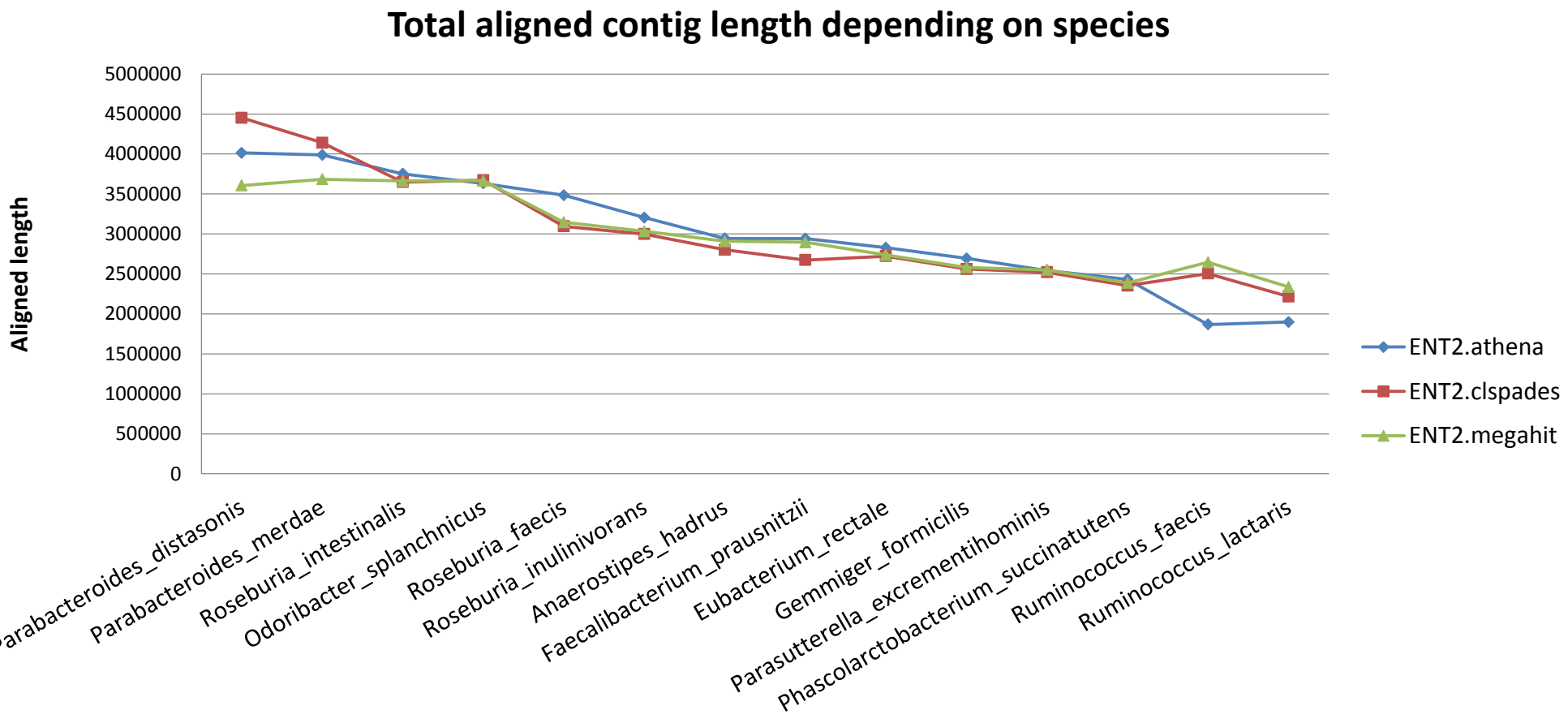
→ The aligned length is comparable for the most of species

→ But for some species, the length of assembly with athena is drastically falling

Metaquast comparison : ENT2

Assembly	ENT2.athena	ENT2.cloudspades	ENT2.megahit
Number of contigs (>500 bp)	**24 855**	**174 228**	**196 290**
Total length	**282 787 620**	**461 018 073**	**449 745 372**
Total aligned length	52 005 807	57 383 930	57 750 825
Unaligned length	**230 341 439**	**400 326 806**	**390 607 541**
Genome fraction (%)	**53,8**	**60,1**	**59,6**
Largest contig	1 519 458	989 393	457 602
Largest alignment	201 261	179 866	174 554
# misassembled contigs	886	959	2 098

Metaquast comparison : ENT2



Quality of reconstructed genomes

- Binning of contigs with MetaBAT : based on tetranucleotide frequency and coverage.

- Quality assesment of bins, inspired from A. Bishara *et al*.
  - checkM, based on genes markers → score of completeness and contamination
  - Aragorn → detection of tRNA
  - Barrnap → prediction of rRNA

- Level of quality :
  - Complete : > 90 % completeness, < 5 % contamination, presence of one copy of each rRNA, at least 18 tRNA
  - High : > 90 % completeness, < 5 % contamination
  - Medium : > 50 % completeness, < 10 % contamination

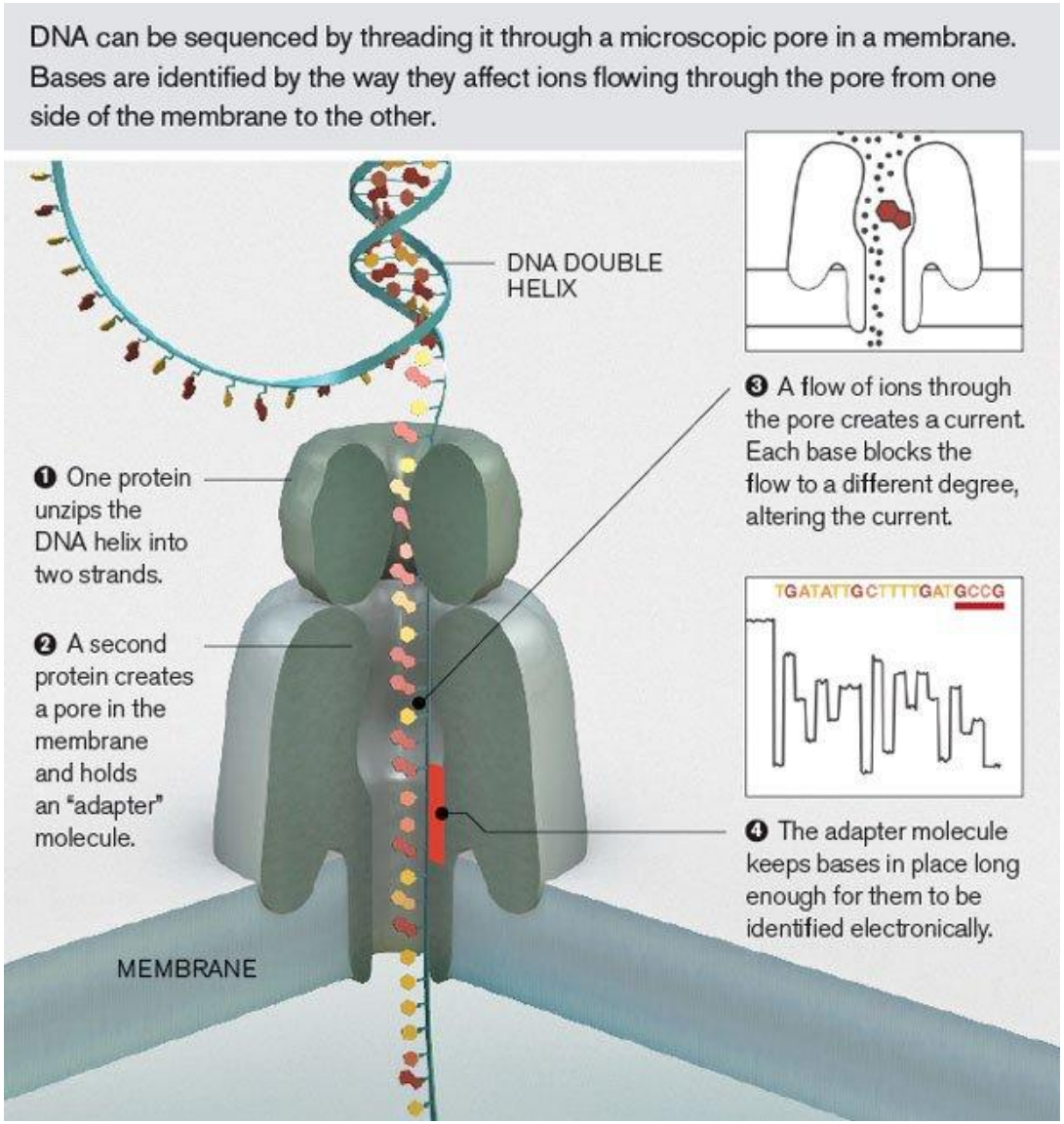
Genomes quality of samples

	ENT1		ENT2		ENT3
	Athena	Cloudspades	Athena	Cloudspades	Athena
low	75	72	76	92	23
	67 %	60,5 %	55,1 %	57,5 %	49 %
medium	24	33	42	52	16
	21,4 %	27,7 %	30,4 %	32,5 %	34 %
high	8	13	12	15	4
	7,1 %	10,9 %	8,7 %	9,4 %	8,5 %
complete	5	1	8	1	4
	4,5 %	0,8 %	5,8 %	0,6 %	8,5 %
total	112	119	138	160	47

→ Athena gives more complete genomes.
→ Cloudspades assemblies are more present in the medium and high qualities

Oxford Nanopore technology

Oxford nanopore technology overview



→ Detection of local electrical potential

→ Base-calling from these values to obtain the DNA sequences

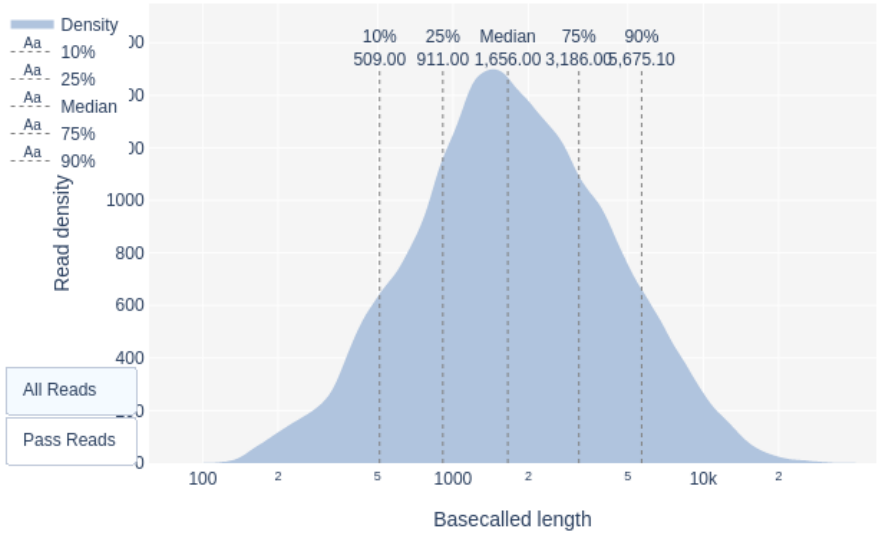
http://www2.technologyreview.com/news/427677/nanopore-sequencing/

Samples sequencing

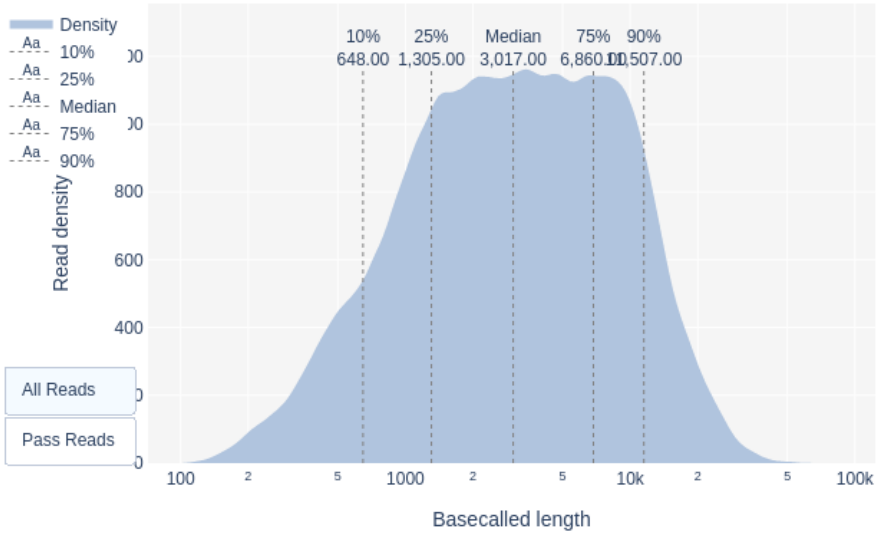
- 2 samples analyzed : ENT2 and ENT3

	ENT2	ENT3
Number of reads	4 922 475	3 053 785
Median read length	1 664	3 025
Total length	12 538 529 930	14 950 090 294

ENT2 density read length



ENT3 density read length



Assembly process

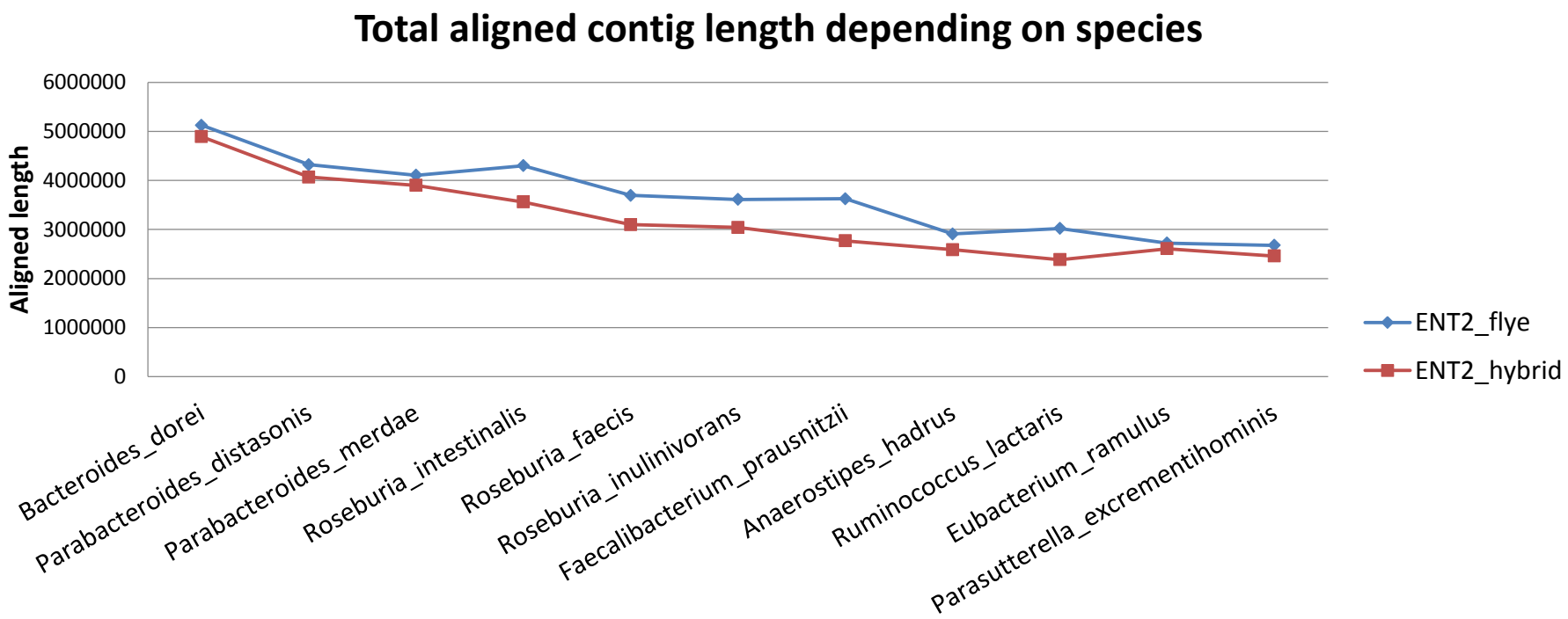
- 2 assembly strategies :
  - Flye assembly :
    - specific tool for long reads assembly with metagenomic option
    - Only use of nanopore reads
  - Hybrid assemby :
    - with metaspades, specific for metagenomic assembly
    - use of nanopore reads and short reads from same individual

Metaquast comparison : ENT2

Assembly	ENT2.flye	ENT2.hybrid
Number of contigs (>500 bp)	**6 569**	**69 726**
Total length	**371 695 694**	**388 325 709**
Total aligned length	53 962 179	49 047 920
Unaligned length	**317 626 904**	**337 528 845**
Genome fraction (%)	**65,8**	**64,5**
Largest contig	2 856 717	2 610 394
Largest alignment	325 246	213 301
# misassembled contigs	372	812

→ The flye assembly, with only use of nanopore reads, gives a slightly better genome fraction for 10-fold less contigs than hybrid assembly.

Metaquast comparison : ENT2



enterome
bioscience

Comparison between 10X and nanopore assemblies

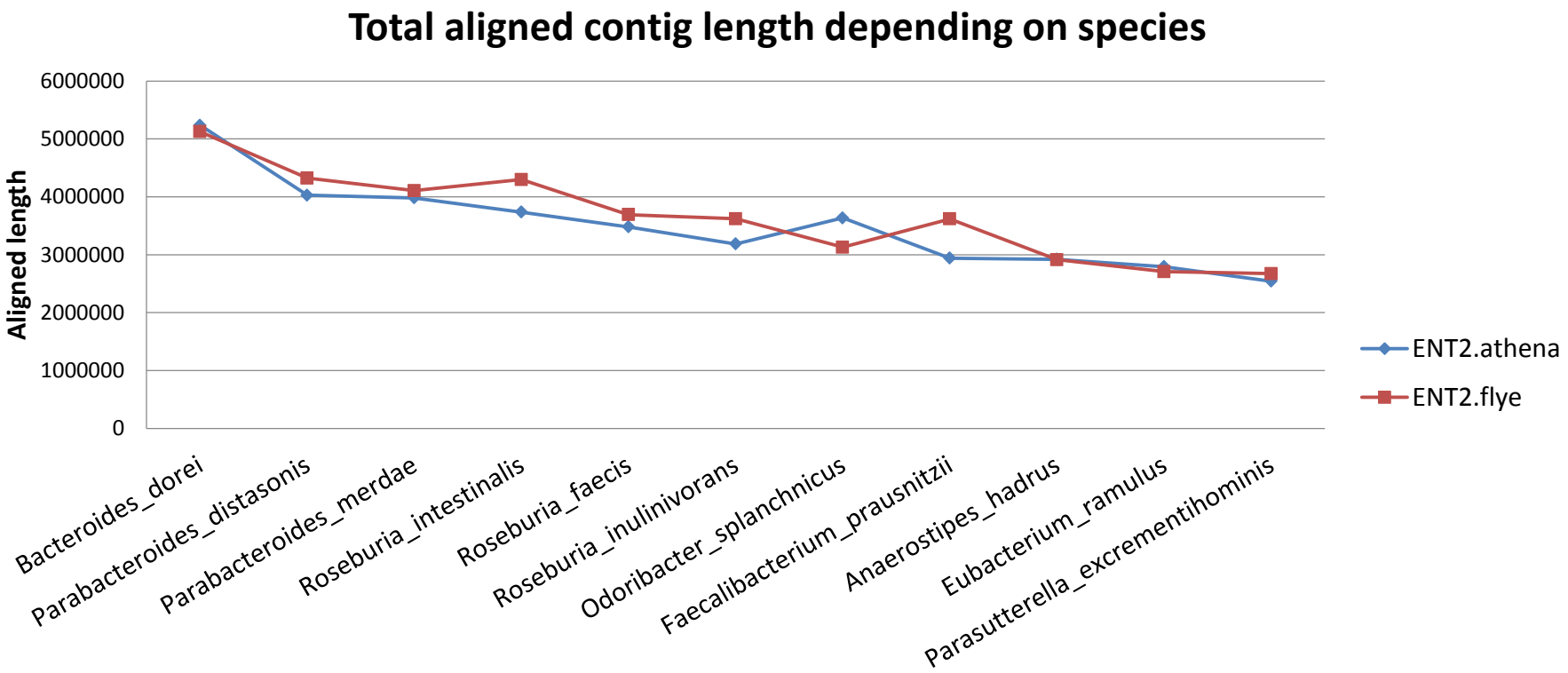
Metaquast comparison for ENT2 sample

Assembly	ENT2.athena 10X	ENT2.flye Nanopore
Number of contigs (>500 bp)	**24 855**	**6 569**
Total length	**282 787 620**	**371 695 694**
Total aligned length	53 197 335	58 181 442
Unaligned length	**229 158 438**	**313 294 956**
Genome fraction (%)	**62,4**	**66,1**
Largest contig	1 519 458	2 856 717
Largest alignment	208 591	325 246
# misassembled contigs	884	448

→ In this example, the assembly with nanopore reads seems to outperform the results obtained with reads from 10X technology

→ But we have only one sample to make the comparison, we cannot generalize

Metaquast comparison for ENT2 sample



Conclusion

Conclusion

- 10X technology :
  - Samples preparation seems to be less difficult than for long reads sequencing
  - No assembler that clearly outperforms the other in all the cases, but these tools are still under developpement

- Oxford nanopore technology :
  - Samples preparation is certainly the bottleneck for metagenomic analysis : DNA extraction step is very difficult to master
  - Tools for hybrid assembly (use of short and long reads) need improvement for metagenomic

enterome
bioscience

Thank you!