# State of the art and comparison of long reads technologies

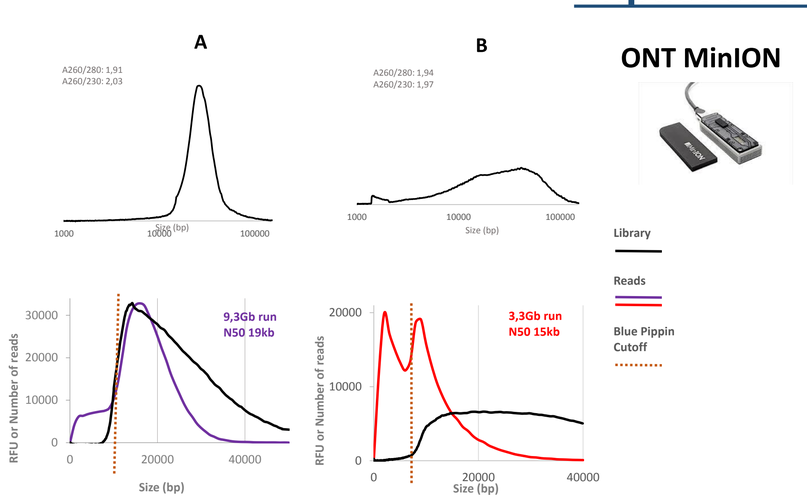
Anais POIRAUDEAU[1*], Maxime MANNO[1*], Céline VANDECASTEELE[1], Alain ROULET[1], Celine ROQUES[1], Marie VIDAL[1], Catherine ZANCHETTA[1], Pauline HEUILLARD[1], Fabrice ROUX[2], Baptiste MAYJONADE[2], Jérôme GOUZY[2], Yann GUIGUEN[3], Christophe KLOPP[4], Pierre FRASSE[5], Mohamed ZOUINE[5], Cécile DONNADIEU[1], Olivier BOUCHEZ[1], Gérald SALIN[1] and Claire KUCHLY[1]

*1 INRA, US 1426, GeT-PlaGe, Genotoul, Castanet-Tolosan, France*
*2 LIPM, UMR 441 INRA, 31326, Castanet-Tolosan, France*
*3 INRA, UR 1037 LPGP Laboratoire de Physiologie et Génomique des Poissons, 35000, Rennes, France.*
*4 MIAT, UR875 INRA, Plateforme bioinformatique, 31326, Castanet-Tolosan, France*
*5 ENSAT, UMR 990 INRA/INPT-ENSAT, Laboratoire Génomique et Biotechnologie des Fruits, 31326, Castanet-Tolosan, France*

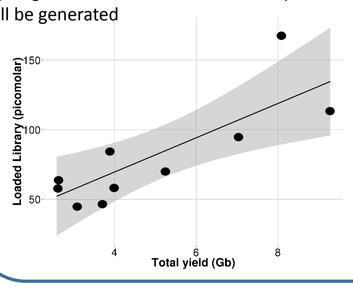
* Equal contribution to this work

**Thanks to its experience on short reads sequencing using the Illumina technologies, the GeT-PlaGe core facility began to evaluate and use long read technologies since the beginning of 2015: Pacific BioSciences RSII, Oxford Nanopore Technology MinION and 10XGenomics Chromium. Genomic issues such as complex genome assembly, structural variant discovery or phasing can be addressed by those long read technologies.**
**As DNA quality is the most important requirement to obtain an efficient sequencing, sample requirements for each technology, and quality controls performed on GeT-PlaGe will be detailed. For all the technology presented, DNA sample needs high quality and purity. For ONT MinION and 10XGenomics Chromium, the reads length have theoritically no limits compare to PacBio RSII (max around 50 kb) but the input DNA size is the key for all of them. The amont of DNA required for sequencing can be huge and challenging to obtain. The DNA quality is the cornerstone of a good bioinformatic analysis particulary for assemblies.**
**We are presenting current projects concerning *de novo* assembly results obtained using multiple Long Read technologies, for several genomes (bacteria, fungus, tomato and fish).**

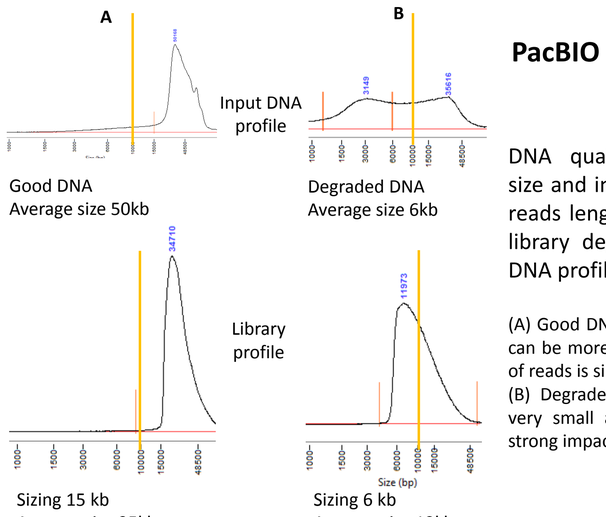
## Impact of the DNA quality on the long read sequencing efficiency

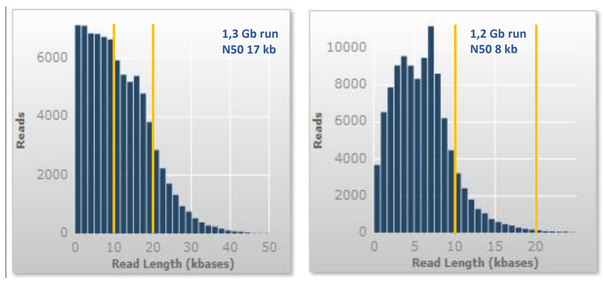



Impact of DNA quality (i.e. degradation) on sequencing efficiency
(A) Non degraded DNA: when a suitable library preparation is performed, the read profile will be very similar to the library profile
(B) Degraded DNA: we will be able to perform a library preparation but a high amount of small reads will be generated




As we work in number of molecules and as we want the longest molecules, a high amount of DNA will be required to perform a library preparation with a very high molecular weight DNA.




### PacBIO RSII

DNA quality impacts library size and indeed on sequencing reads length. The sizing of the library depends on the input DNA profile.

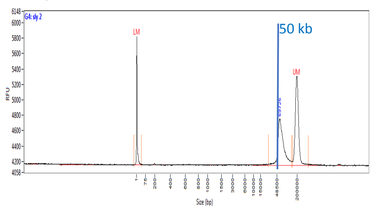
(A) Good DNA: the size on the library can be more than 10 kb and the N50 of reads is similar (17 kb).
(B) Degraded DNA: the size can be very small around 6 kb and has a strong impact of the N50 reads (8 kb)

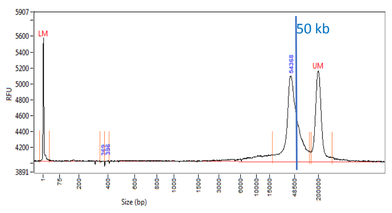
### 10X Genomics Chromium

The 10X Chromium technology requires fragments greater than 50 kb. For *de novo* assemblies, it recommends 100kb fragments.

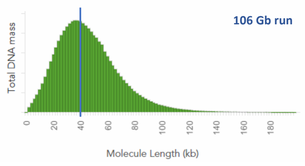
DNA concentration: 3,70 ng/µL
83,5% size > 50 kb
No purification

DNA concentration: 11,6 ng/µL
37,4% size > 50 kb
Purification 1,5 by Ampure XP beads

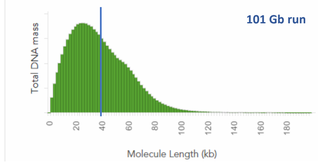



| | |
|---|---|
| SNPs Phased | 98.8% |
| Longest Phase Block | 8,011,755 bp |
| N50 Phase Block | 1,758,571 bp |
| Molecule Length | µ 40,187 bp |
| DNA in Molecules >20kb | 87.5% |
| DNA in Molecules >100kb | 3.15% |
| Corrected Estimated of DNA Loaded | 1.42 ng |




| | |
|---|---|
| SNPs Phased | 97.8% |
| Longest Phase Block | 2,091,473 bp |
| N50 Phase Block | 137,044 bp |
| Molecule Length | µ 33,215 bp |
| DNA in Molecules >20kb | 75.4% |
| DNA in Molecules >100kb | 1.51% |
| Corrected Estimated of DNA Loaded | 1.05 ng |




With the 10X Chromium, DNA quality and large fragment size are essential for a good use of the technology and impact the size of the reconstructed fragments and the size of the phased blocks.

## ONT MinION vs PacBio RSII: let's fight !

Concerning the problematic of *de novo* assembly using long reads, PacBio technologies are challenged by ONT's. Can we compare them on the same input DNA (Bacteria and Fungus) ?

### Bacterial raw data accuracy

ONT raw data accuracy seems to be better than PacBio's, but errors are not random (long homopolymer bias). It might be an obstacle to obtain a good assembly.
Can we use the ONT technology alone or combined with Illumina data to get a genome with the equivalent quality than PacBio ?

| Raw data sets | Accuracy[1] |
|---|---|
| HiSeq[2] | 99.7 % |
| RSII[3] | 81.3 % |
| MinION[4] | 86.3 % |

[1] Accuracy base on alignment to the PacBio genome reference
[2] Illumina MiSeq total raw data
[3] PacBio RSII total raw data
[4] ONT MinION total raw data

### Bacterial genome assembly (5 Mb)

| Assembly sets | Input Cov[1] | nbrContigs | totNuc |
|---|---|---|---|
| MinION-CANU[2] | 1156 | 1 | 5045252 |
| MinION-CANU-PILON | 71[3] | 1 | 5122723 |
| RSII-HGAP3[4] | 81 | 1 | 5078836 |

[1] Raw data coverage (X)
[2] ONT raw data (1D protocol), qscore > Q10 and size > 3 kb
[3] Illumina data coverage
[4] PacBio raw data, polymerase qscore > 0.80 and subread size > 3 kb

**Assembly Pipeline for ONT data**

**SEQUENCING – MinKNOW 1.3.25**
→ Output: Raw Data (.fast5)
Quantity: 226 Go
Duration: 2 days

**BASECALLING - Albacore 1.0.1**
→ Output: Raw Data + Calling (.fast5)
Quantity: 1.3 To
Duration: 3 days
Ressources: 24 threads (2G mem/thread)

**ASSEMBLY - Canu 1.3**
→ Output: (.fasta/q)
Quantity: 0,5 Go
Duration: 3 hours
Ressources: 8 threads (1G mem/thread)

**POLISHING - PILON 1.21**
→ Output: (.fasta/q)
Quantity: 0,5 Go
Duration: 0.5 hour
Ressources: 8 threads (10G mem/thread)

The ONT read length allow to get just one contig for bacteria like the other sets. ONT Assembly has been perfomed with all raw data but we didn't assess with less coverage.
From basecalling to assembly, the PacBio RSII pipeline is more efficient than the ONT pipeline which is frequently updated.

### Bacterial assembly completness[1] metrics

| Assembly sets | Complete genes | Fragmented genes | Missing genes | Total genes groups searched |
|---|---|---|---|---|
| MinION-CANU | 25 (16.9%) | 53 (35.8%) | 70 (47.3%) | 148 |
| MinION-CANU-PILON | 141 (95.3%) | 0 | 7 (4.7%) | 148 |
| RSII-HGAP3 | 141 (95.3%) | 0 | 7 (4.7%) | 148 |

[1] Completness computed with BUSCO V2 on the bacteria_odb9

The best BUSCO results are given by PacBio and ONT+Illumina data: 95.3% of complete genes. While ONT data allow us to generate only one contig, most of the genes are missing or fragmented.
With these versions of MinKNOW and Albacore, ONT data have to be combined with Illumina's to get the same results as PacBio RSII.

### Fungus genome assembly (50Mb)*

| Assembly sets[1] | N50 raw data | Quantity of raw data | Nbr contigs | N50 contigs | Total size of contigs | Complete genes |
|---|---|---|---|---|---|---|
| RSII | 9.7Kb | 2.9 Gb | 915 | 166 kb | 84 Mb | 92% |
| MinION LR[2] | 19Kb | 2.3 Gb | 408 | 312 kb | 68 Mb | 89% |
| MinION LR+8 Kb[3] | 13.5Kb | 3.8 Gb | 525 | 160 kb | 64 Mb | 89% |

[1] ONT and PacBio assembly with CANU 1.5 + PILON
[2] ONT 1D Long Reads protocol, sizing from 15 Kb to 90 Kb.
[3] ONT 1D Long Reads protocol, sizing from 15 Kb to 90 Kb + ONT 1D protocol, sharing at 8 Kb, size > 8 Kb
[4] Completness computed with BUSCOv2

**Pipeline softwares**

**SEQUENCING – MinKNOW 1.6.11**
**BASECALLING - Albacore 1.1.0**
**ASSEMBLY - Canu 1.5**
**POLISHING - PILON 1.21**
**COMPLETNESS - BUSCO v2**

With this more complex genome, it is difficult to conclude about the quality of the assembly. Regarding the PacBio data, we obtained a higher number of contigs, but also more complete genes. ONT data allow us to reduce the number of contigs despite of the integrity of genes. Moreover, most of the short ONT reads can be removed to improve the assembly metrics.

In case we have some Illumina data, ONT can be equivalent to PacBio concerning the *de novo* assembly. It strongly depends on the biological query and the genome complexity. Today, we have to combine ONT data with an other data set to exploit them, but that can change in a short term future.

* Acknowledgments: Thanks to Frédéric Breton *et all* and Tan Joon Sheong *et all* to allow us to present their results

## PacBio RSII, 10X Chromium: a love story ?

**These two approaches are long reads technologies but unlike PacBio RSII, 10X Genomics use synthetic reads. For assembly purposes, are PacBio RSII and 10X Chromium results similar or is there a benefit to use them together?**

The analysis have been performed on fish and tomato genomes. For PacBio *de novo* assemblies, we assessed 5 long reads assembly softwares: Miniasm [v0.2], Falcon[v0.4.2], Canu[v1.4] and SmartDeNovo[v1.0.0]. We obtained the best results with Canu and SmartDeNovo, so only those will be presented here.

### Genome assemblies metrics

| Fish genome assembly | Cov (X) | # | L50 contig | N50 contig |
|---|---|---|---|---|
| **10x Supernova[1]** | 104x | 43476 | 3641 | 51157 |
| **PacBio SmartDeNovo** | 71x | 701 | 55 | 4137750 |
| **PacBio Canu** | 71x | 4062 | 126 | 1341974 |
| **10x + SmartDeNovo[2]** | | 534 | 44 | 5470783 |

| Tomato genome Assembly | Cov (X) | # | L50 contig | N50 contig |
|---|---|---|---|---|
| **10x Supernova[1]** | 116x | 24579 | 1995 | 104298 |
| **PacBio SmartDeNovo** | 81x | 857 | 112 | 2062400 |
| **PacBio Canu** | 81x | 508 | 47 | 4973822 |
| **10x + Canu[2]** | | 284 | 19 | 13603907 |

[1] The assemblies are realized by Supernova (v.1.1.2) a 10xGenomics software.
[2] To combine PacBio RS II and Chromium 10x assemblies, we used ARCS software.

According to assemblies metrics, the best results are obtained with the PacBio-SmartDeNovo for fish assembly and PacBio-Canu for tomato assembly. The use of ARCS software improves a lot the metrics.

**Pipeline softwares**

**Canu + Quiver:**
Duration = 2 days + 15 days
Threads = 8 (80G/Threads) + 16 (40G/Threads)

**SmartDeNovo + Quiver:**
Duration = 2 days + 9 days
Threads =16 (80G/Threads) + 16 (40G/Threads)

**10x Supernova :**
Duration = 5 days
Threads = 20 (300G/Threads)

**ARCS software:**
Duration = 3 days
Threads = 25 (30G/Threads)

### Genome assembly completness[1] metrics

| Fish Assembly[2] | Complete | Fragmented | Missing | Total groups searched |
|---|---|---|---|---|
| **10x Supernova** | 2131 (82,4%) | 328 (12,7%) | 127 (4,9%) | 2586 |
| **PacBio SmartDeNovo** | 2277 (88,1%) | 224 (8,7%) | 85 (3,2%) | 2586 |
| **ARCS SmartDeNovo** | 2278 (88,1%) | 223 (8,6%) | 85 (3,2%) | 2586 |

| Tomato Assembly[3] | Complete | Fragmented | Missing | Total groups searched |
|---|---|---|---|---|
| **10x Supernova** | 1303 (90,5%) | 44 (3,1%) | 93 (6,5%) | 1440 |
| **PacBio Canu** | 1353 (94,0%) | 25 (1,7%) | 62 (4,3%) | 1440 |
| **ARCS Canu** | 1357 (94,2%) | 23 (1,6%) | 60 (4,2%) | 1440 |

[1] Completness computed with BUSCO V2
[2] Lineage data: vertebrata_odb9
[3] Lineage data: embryophyta_odb9

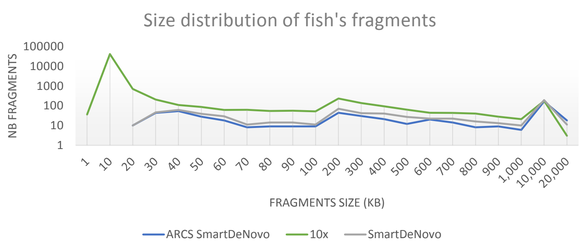
In the BUSCO analysis, the number of ancestral genes found is better for PacBio assemblies.
The BUSCO score increases a little bit for the combination of 10X Chromium and PacBio assemblies: there are more genes found and less fragmented and missing genes.

**Pipeline software**

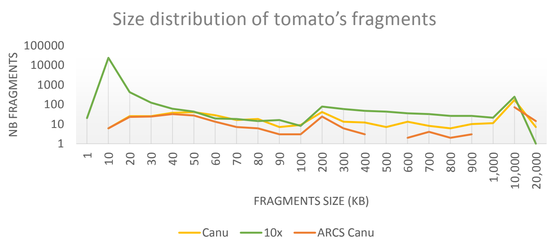
Duration = 4h
Threads = 4 (60G/Threads)

### Fragment distribution

**Fish Results**




**Tomato Results**




In these two graphs, the fragment distribution of both samples are represented. The distributions follow the same trend except for 10X Chromium which are over represented at 10 kb. With ARCS software, we obtained less contigs but most of them are at 10.000 kb. Furthermore, there are more fragments at 20.000 kb in ARCS assemblies than the others. Indeed, the use of ARCS permits to obtain longer fragments than PacBio or 10X Chromium individual assemblies.

After the end of the analysis, we can not say there is a general « best assembler » for PacBio RSII because the result varies according to the samples. These differences can be explained by the genome complexity.
The 10X Chromium Supernova results do not indicate better metrics than PacBio RSII. The combination of both improves significantly the PacBio quality assembly metrics.