IMPLEMENTATION AND EVALUATION OF THE MINION

□ Catherine ZANCHETTA & Maxime MANNO, Céline VANDECASTEELE, Claire KUCHLY, Olivier BOUCHEZ, Christophe KLOPP, Alain ROULET, Gérald SALIN, Céline ROQUES, Cécile DONNADIEU
Unité INRA 1426, GenoToul Genomics Platform (GeT-PlaGe & Bioinfo), 31326 Castanet-Tolosan, FRANCE

□ Baptiste MAYJONADE, Jérôme GOUZY, Fabrice ROUX
UMR CNRS-INRA 441-2594, LIPM, 31326 Castanet-Tolosan, France
□ Guillaume CROVILLE, Jean-Luc GUERIN
UMR 1225 ENVT-INRA, IHVV, 31076 Toulouse, France
□ Caroline CALLOT, Stéphane CAUET, Hélène BERGES
Unité INRA 1258, CNRGV, 31326 Castanet-Tolosan, FRANCE

Toulouse « Genome and Transcriptome » core facility (GeT-PlaGe - France) enables scientists to design and carry out their studies using multiple Next Generation Sequencing technologies. GeT-PlaGe acquired the MinION (Oxford Nanopore Technologies) allowing to sequence unique DNA or RNA molecules by electroporation. DNAs of three different species have been sequenced in order to assess the technology and see if it meets the users needs. The ability to produce high molecular weight libraries improve the yield. Depending on the biological question, ONT data have to be completed with Illumina data.

EVALUATION OF BASIC PARAMETERS

KEY POINTS

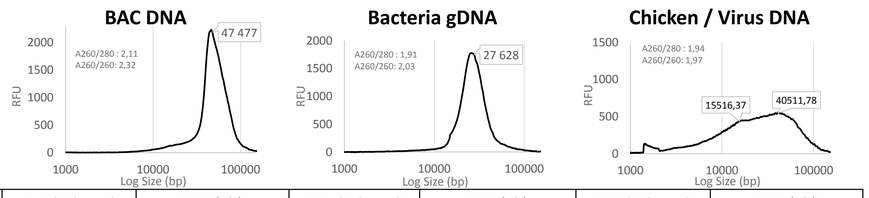
- **For complex genome assembly, if a shearing and a size selection are performed on long fragments, very long reads can be obtained (40 kb shearing → N50 = 24.6, up to 9 Gb).**
- **For small genomes, the quantity of data allows multiplexing which will reduce the cost of the sequencing.**
- **For diagnosis and production fields, the Rapid and the 1D protocols allow a fast library preparation (30 minutes – 3 hours).**
- **An important amount of DNA is required which is a technological limitation for some studies (minimum of 1.5 µg to produce 2Gb).**

THE MINION FOR A BACTERIAL GENOME ASSEMBLY

KEY POINTS

- **For small genomes, MinION combined to Illumina data generate high quality genome assemblies (0% of fragmented genes).**
- **Albacore 1.0.1. version reduces the error rate on Raw data (13.68%) but seems to create false homopolymers.**
- **The informatics pipeline from the sequencing to the assembly can be all completed in local (12 days duration).**
- **Important informatics resources are necessary, especially for the base calling (9.3 Gb run : 60 threads and 1.4 To workspace).**

1 / DNA profiles and metrics obtained



Preparation kit used and Amount of DNA required	Total Filtered (Gb)	N50 (Kb)	Total Filtered (Gb)	N50 (Kb)	Total Filtered (Gb)	N50 (Kb)
Rapid 200 ng	0.1	37.9	0.5	10.1	0.3	33.5
1D 1,5 µg	2.6	9.6	4.0	8.7	3.9	9.4
1D+ $\frac{1{,}5\ \mu g\ *\ mean\ size}{8\ Kb}$	8.1	24.6	9.3	19.0	3.3	15.1

These metrics have been obtained using R9.4 flowcells, MinKNOW 1.1 and 1.3 and Albacore 1.0.1 versions. It has been verified that the amount of active pores per flowcell were similar and that all flowcells were able to get 450b/s.

- For small genomes, the Rapid protocol can produce long reads and up to 500 Mb.
- The 1D protocol (8 kb shearing) is the one giving the most similar results among DNAs.
- The 1D+ protocol (including Megaruptor shearing and BluePippin size selection ; adapted for each DNA) is the one giving the highest amount of data and N50.

INFORMATIC PIPELINE FOR LOCAL SOLUTIONS

Sequencing of a bacterial genome : *Xanthomonas campestris* 5 Mb 1D+ protocol

SEQUENCING – MinKNOW
→ Output : Raw Data
Type : .fast5
Quantity : 226 Go
Duration : 2 days

BASECALLING (and DEMULTIPLEXING) - Albacore
→ Output : Raw Data + Calling
Type : .fast5
Quantity : 1.3 To
Duration : 6 days
Ressources : 24 threads (2 G mem/thread)

EXTRACTION - Poretools
→ Output : Extract reads
Type : .fastq
Quantity : 18 Go
Duration : 2 hours
Ressources : 10 threads (1 G mem/thread)

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

MINION POLISHING – Nanopolish 0.6.1
→ Output : Polished.assembly
Type : .fasta(q)
Duration : 3 days
Ressources : 8 threads (1 G mem/thread)

ILLUMINA POLISHING – Pilon
→ Output : Polished.assembly
Type : .fasta(q)
Duration : 1 day
Ressources : 8 threads (1 G mem/thread)

4 / Assessment of the *Xanthomonas campestris* (*Xc*) genome assembly using BUSCO

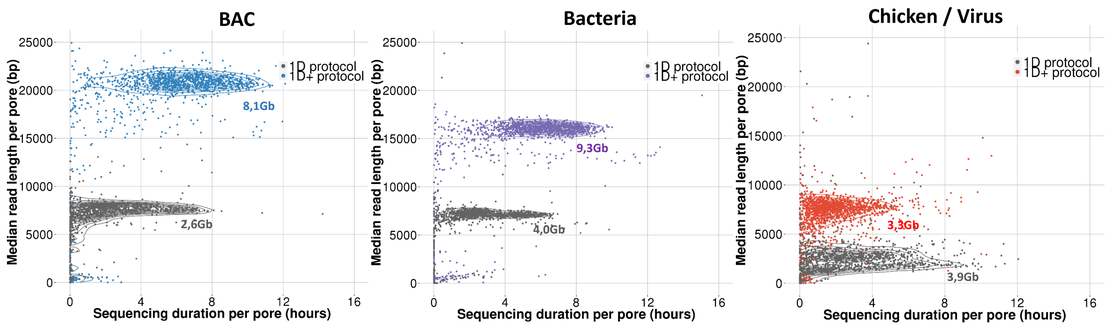
- *Xc* is a plant pathogenic bacteria
- Genome of 5 Mb
- Maximum of 10 bases homopolymers
- Total of 950 homopolymers between 6 and 10-mers

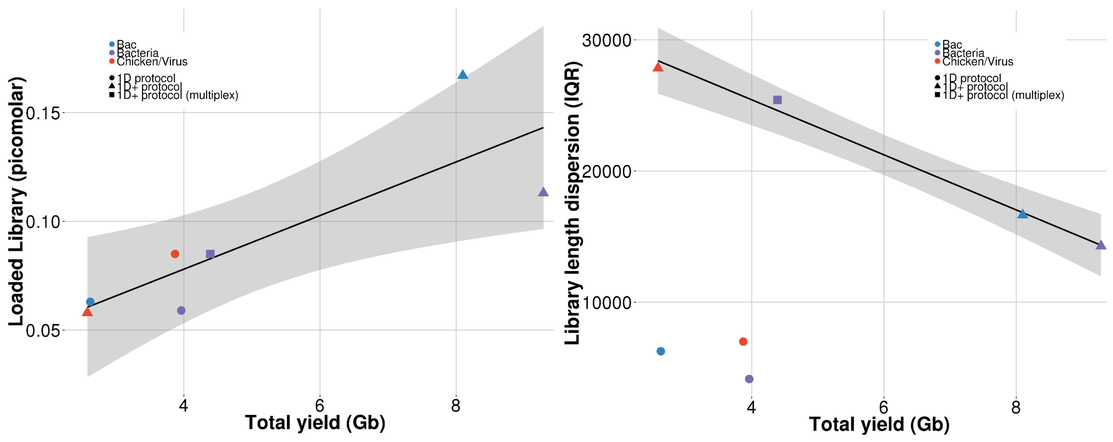
Assembly methods	% Complete genes	% Fragmented genes	% Missing genes
MinION	12.2	27	60.8
MinION-nanopolish	71.6	15.5	12.9
MinION-nanopolish-pilon	**89.2**	**5.4**	5.4
MinION-pilon	**95.3**	**0**	4.7
Pacbio-hgap3	**95.3**	**0**	4.7

BUSCO provides quantitative measures for the assessment of genome assembly. The comparison is based on *Escherichia coli* genome.

- The best results are given by Pacbio data and MinION+Illumina data (base calling with the Metrichor 1.121 version) : 95.3% of complete and 0% of fragmented genes on the *E. coli* genome.
- It seems that Nanopolish generates errors that then Pilon is not able to correct.
- With these versions of MinKNOW and Metrichor, ONT data have to be combined with Illumina data to get a good genome assembly.

2 / Impact of the library on the yield





- The size of fragments going through a pore seems to have an impact on the sequencing duration of this pore (and so on the yield).
- The amount of loaded molecules is positively correlated to the quantity of data. We haven't identify any profile of overloading.
- The ability to prepare library with fragments of similar size will increase the quantity of data (for the 1D+ protocol).

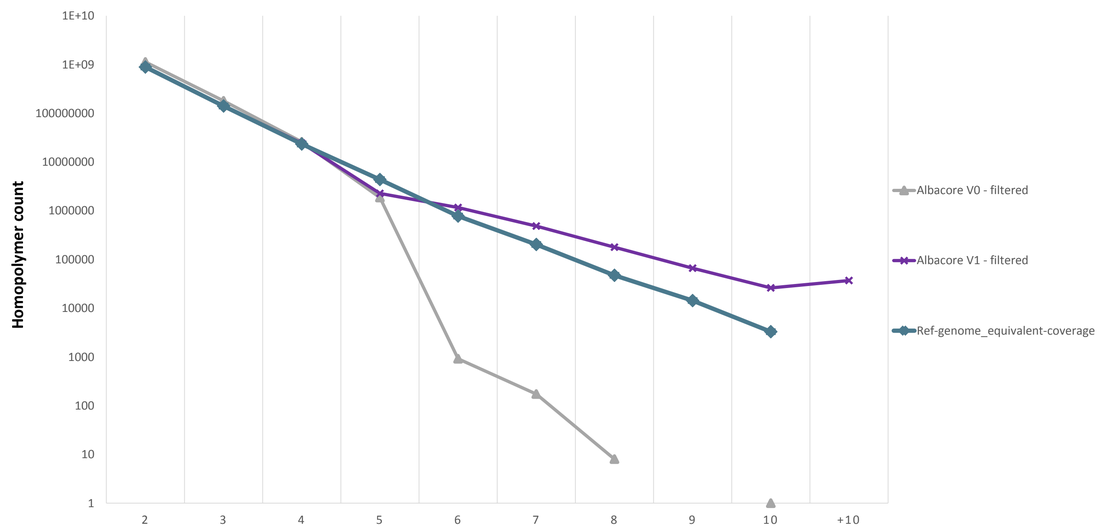
5 / Error rates on Illumina, Pacbio and ONT Raw data with Albacore

Technologie	Error rate
Illumina	0.35%
Pacbio	**18.72%***
MinION Albacore V0	**18.99%**
MinION Albacore V1	**13.68%**

The reference genome has been done using the Pacbio RSII data obtained with the same DNA.
* The Pacbio library wasn't optimal and generated 39,63% of clipping which dramatically increased the error rate.

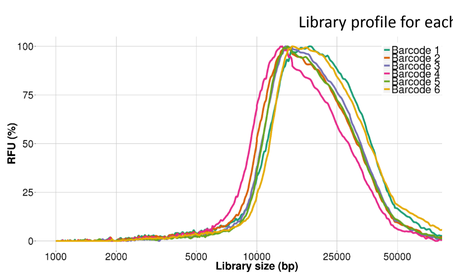
- On Raw data, the error rate is lower on ONT data (Albacore 1.0.1 – V1) than with Pacbio data.

6 / Albacore V0 and V1 comparison for homopolymer count on Raw data



- From 2 to 5-mers, the count is similar between base callers and correlated to the reference genome.
- 6-mers and more are partially or not identified by the V0.
- For 6-mers and more the V1 generates a high number of homopolymers (further analysis have to be performed to assess if they are « true » homopolymers).
- The V1 creates homopolymers of more than 10 bases (and up to 7000).

3 / 6-plex bacterial genome sequencing



	% of reads per barcode	N50 (Kb)
1	17	10
2	17	13
3	13	10
4	21	16
5	15	11
6	17	13

A 20kb Megaruptor shearing and a 1D library preparation have been performed. The quality were similar between barcodes.

- The total yield is 4.4 Gb and 297000 sequences. 76% of sequences have been assigned.
- By pooling 6 libraries with similar profiles, the pourcentage of assigned sequences per barcode is between 13% and 21% (expected 16.6%).

7 / Perspectives

- **To get a better yield, we have to assess the technology with very long fragments and a higher amount of loaded library as no overloading profile has been identified.**
- **For complex genome assemblies, we have to compare the combination of ONT data with Pacbio, 10X or BioNano data.**
- **Regarding Albacore improvement and the new 1D² protocol, we need to assess again MinION data corrected with Nanopolish .**