PacBio RSII : first developments at GenoToul

❑ Céline Vandecasteele, Olivier Bouchez, Gérald Salin, Céline Jeziorski, Cécile Donnadieu, Denis Milan
INRA, GenoToul Genomics Platform (GeT-PlaGe), 31326 Castanet-Tolosan

❑ Baptiste Mayjonade, Jérôme Gouzy, Stéphane Muños
CNRS/INRA, UMR LIPM 441-2594, 31326 Castanet-Tolosan

❑ Cyrille Jarrin, David Vilanova, Patrick Robe
LIBRAGEN SAS, 31400 Toulouse

Single Molecule Real Time (SMRT) Sequencing – PacBio RSII P6C4

To resolve a complex genome assembly, PacBio technology could help to improve the whole genome sequencing thanks to its capacity to sequence long fragments. In the SUNRISE Project, the Sunflower reference genome is being improved thanks to this technology combined with additional developments for library preparation and data analysis.

For metagenomic analysis, high-throughput sequencing of the 16S rRNA gene has become a valuable tool for characterizing microbial communities. In the Meta-Pac project, we are evaluating the PacBio RSII for full-length 16S rRNA gene sequencing and community profiling.

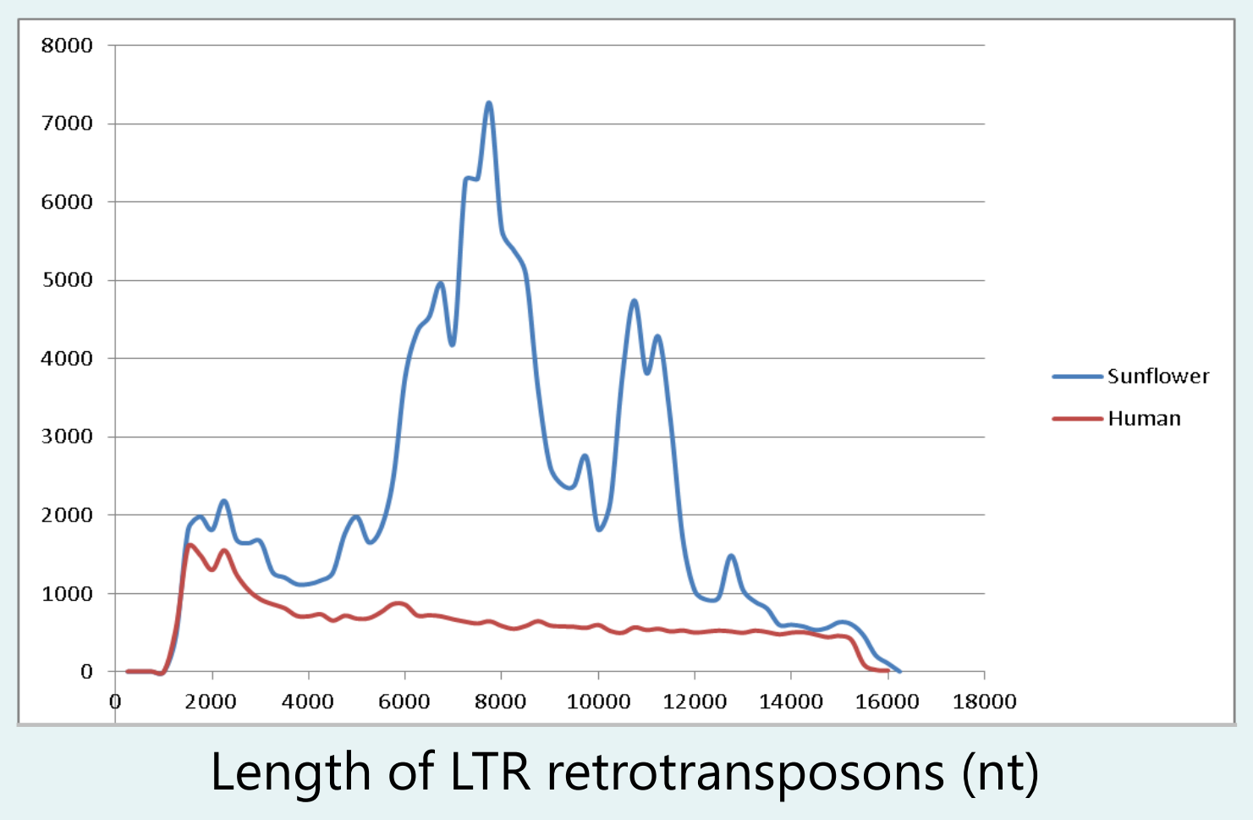
Towards a sequencing of longest fragments to resolve a complex genome assembly

SUNRISE project : SUNflower Ressources to Improve yield Stability in a changing Environment

One of the objectives is to identify the genetic and molecular factors involved in mechanisms underlying oil yield stability of sunflower hybrids under water constraints. To achieve this objective, it's essential to have a robust reference genome and PacBio sequencing could help to improve the sunflower genome assembly. The XRQ line of sunflower (3,6 Gb) was sequenced at 100X depth with PacBio sequences only.

Why is it so difficult to assemble the sunflower genome ?

There is a lot of repeated sequences in the sunflower genome.
They are large (9-12kb) and highly conserved.



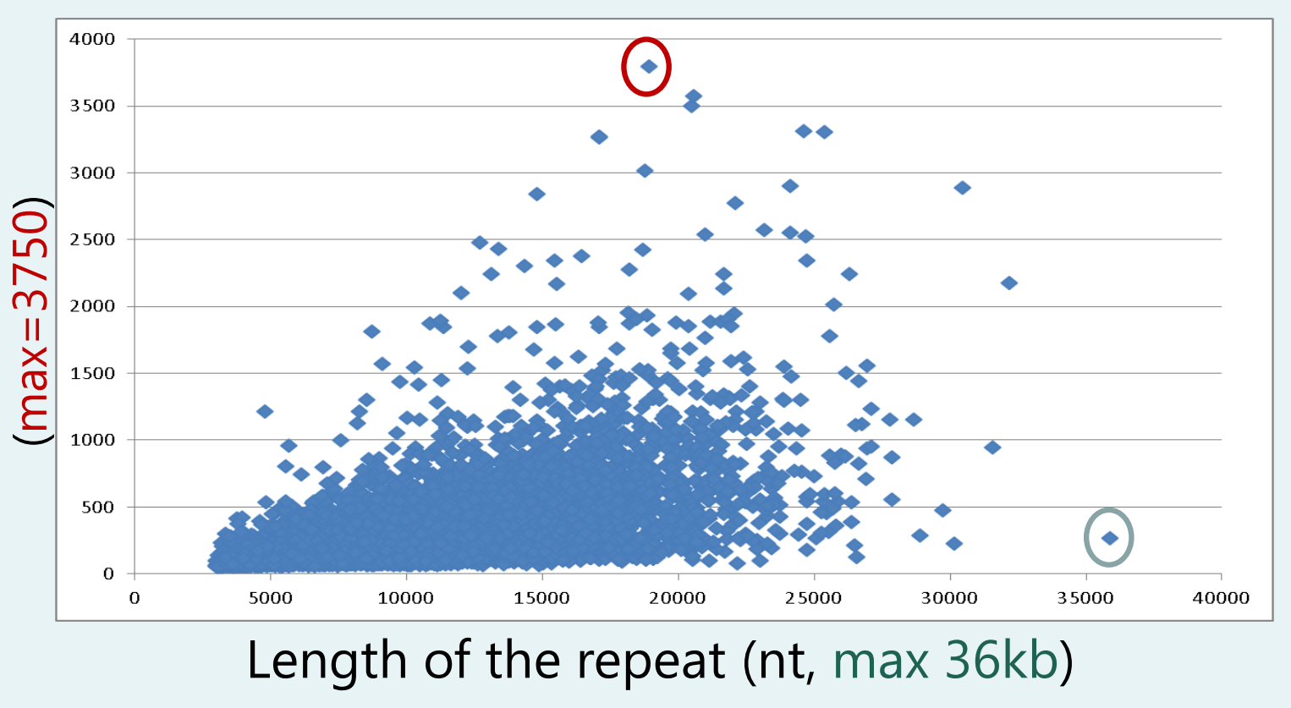
Length of LTR retrotransposons (nt)

Analysis of the composition of the LTR retrotransposons with LTRharvest in Human and Sunflower

(D. Ellinghaus *et al.* 2008, default parameters)

30% of the sunflower genome sequence is composed of LTR retrotransposons.

8.8% of the human genome.



Length of the repeat (nt, max 36kb)

Construction of a database containing repeated sequences

Mapping of 1x of data on 2x of long reads (>= 20Kb)

Analysis of the coverage of the long reads (only hits > 3kb are analyzed)

Repeats pattern identification (MHAP/MinHash)

Improvements of the molecular biology steps have increased the length of the PacBio Sequences

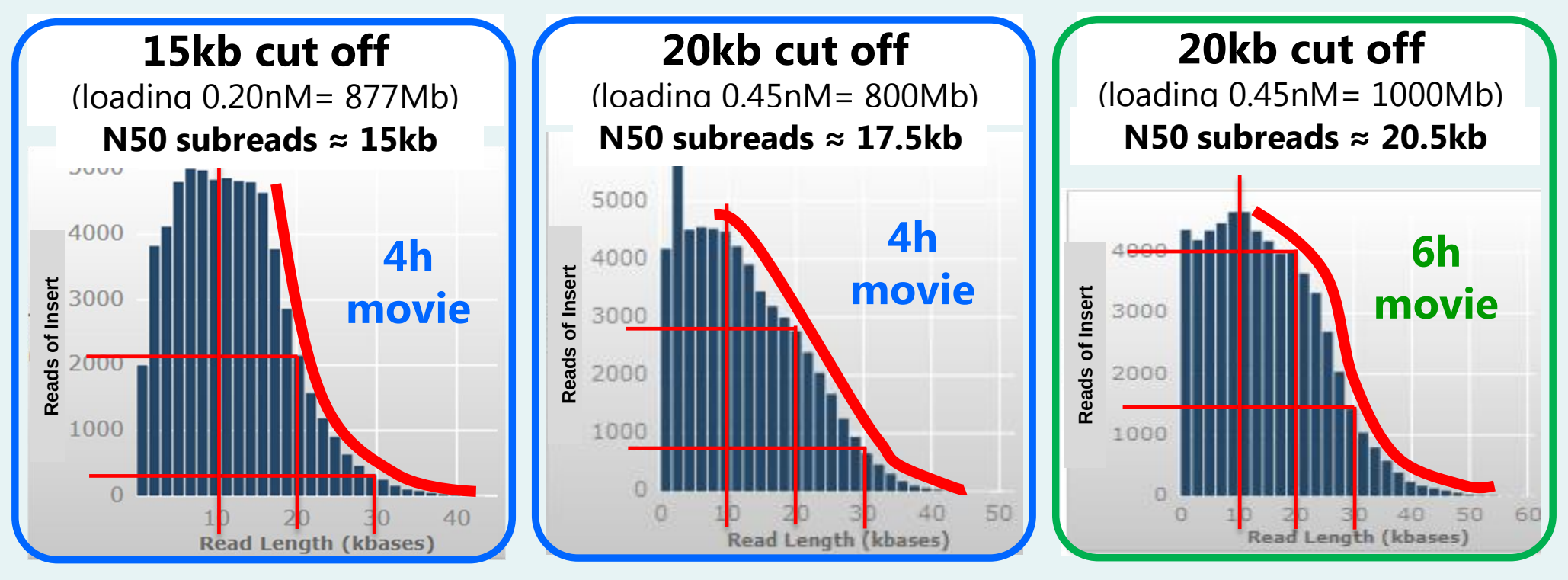
To fully cross the length of the repeats, very long reads were obtained by the improvement of these 3 steps :

High Molecular Weight gDNA Extraction — Long fragment library preparation (>15 kb) — Movie time (4H>6H)

	MAX	N50 BP	MEAN	BP/SMRTCell
IGM (San Diego, USA) 202 SMRTCells – 12 kb library – 4H movie				
MEAN	45457	12211	9176	906 Mb
MAX	52725	12981	9997	1,36 Gb
Lausanne University (Swiss) 59 SMRTCells – 15 kb library – 4H movie				
MEAN	46800	15172	10773	1,15 Gb
MAX	53253	16132	11436	1,6 Gb
INRA, GeT-PlaGe Platform (France) 146 SMRTCells – 15 to 20 kb library – 4 to 6H movie				
MEAN	52317	15365	10327	800 Mb
MAX	80974	20507	13635	1,3 Gb

Subread statistics on PacBio Data obtained for the XRQ sunflower line (3 months - 407 SMRTCells)

Impact of the size selection and movie time on subread length distribution



The Whole Genome Sequencing by PacBio technology has improved the *de novo* genome assembly in Sunflower

#ctg	MAX	N50 BP	#>N50	MEDIAN	Gb
102 X depth PacBio P6C4 sequences assembly					
13124	4.4 Mb	498 kb	1700	118 kb	3.03
127 X depth HiSeq sequences assembly					
1007165	237.4 kb	9.4 kb	34006	392 bp	1,56

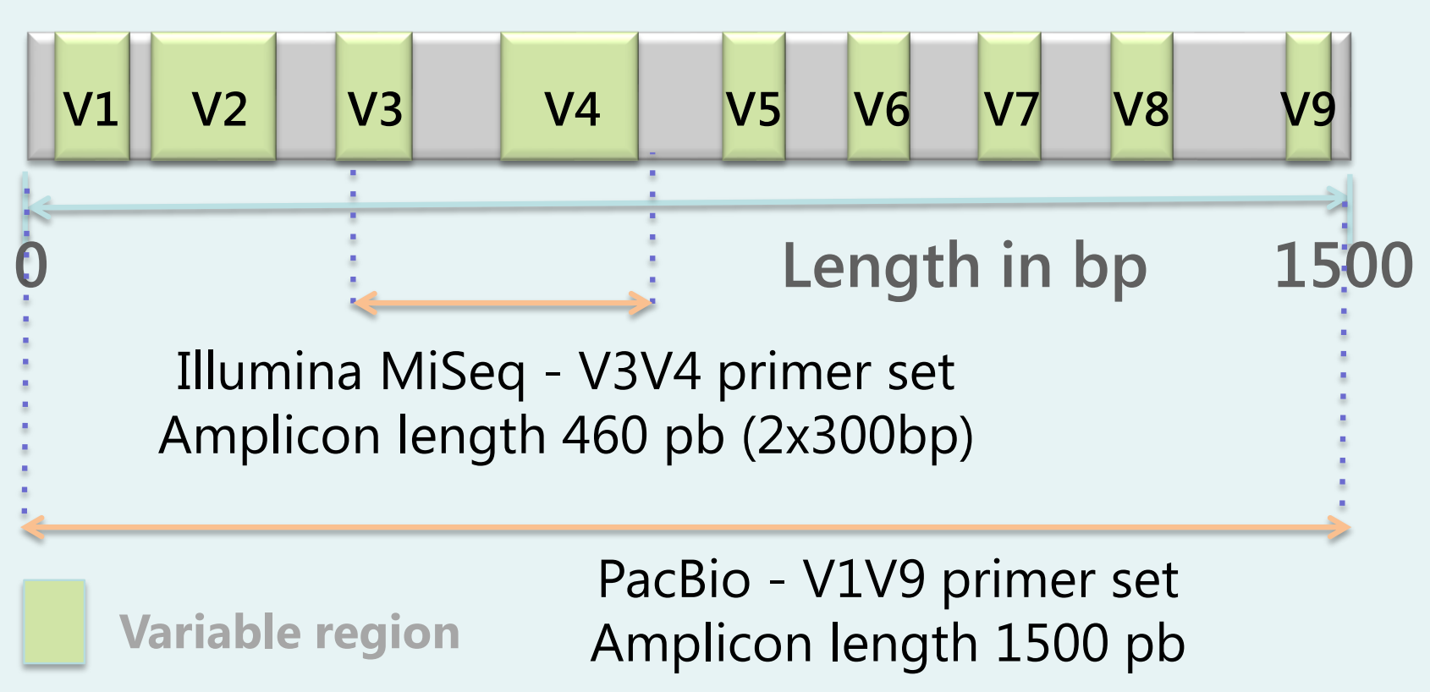
Using PacBio sequences only, the coverage of the sunflower genome was improved from 43% (127X HiSeq) to 84% (102X PacBio) and the size of the contigs have been highly increased. With only 18X depth of PacBio sequences and 2 days of computation (PBcR 8.3rc1), we obtained an assembly with metrics similar to the previous assembly obtained with 127X of HiSeq data (SOAPdenovo).

A full length 16S sequencing for a complete and accurate Metagenomic analysis

Meta-Pac Project : Analysis of Full-Length Metagenomic 16S Genes by SMRT® Sequencing

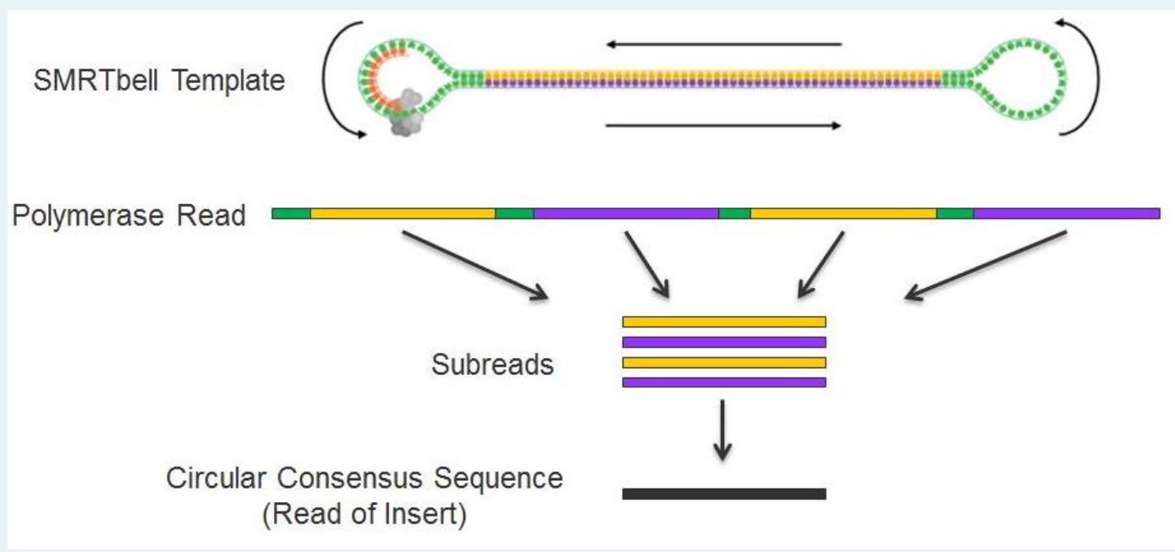
The capacity of sequencing and assembly by PacBio to obtain a reliable full-length 16S Genes (1,5 kb) is evaluating in order to improve the metagenomic analysis.

Full-length 16S sequencing achievable with PacBio technology



Comparison of two sequencing strategies for metagenomic analysis using variable regions within 16S rRNA gene from synthetic MOCK community and natural samples.

Profile metagenomic communities with single-molecule reads using circular consensus sequencing



Circular Consensus Sequencing (CCS)

The circular nature of the SMRTbell DNA template allows polymerase to sequence the same DNA molecule multiple times with multiple passes. This produces high intra-molecular consensus accuracy.

At least 8 full-pass subreads from an insert allow to reduce error rate and obtain a reliable full-length 16S gene.

The Amplicon Sequencing by PacBio technology seems to improve the Metagenomics analysis

The first data analysis seems to show that the Pacbio technology generates an increase error rate compared to the Miseq platform (mainly highly increased Indel errors). However, thanks to its capacity to generate longer reads, the Pacbio technology seems to offer better resolution for 16S analysis and an increase of species richness and number.

Acknowledgments : Thanks to GeT-PlaGe team and INRA administration for the quickness and effectiveness of PacBio installation, Gerrit Kuhn (PacificBiosciences) for his precious advices and GenoToul Bioinformatics platform for data analysis.

http://get.genotoul.fr/