



Toulouse Genomic Core facility

Cécile Donnadieu

GeT Animator

GeT-PlaGe Manager

<http://get.genotoul.fr>

@GeT_Genotoul



The Core facility missions

- **To provide innovating technologies for genome analysis to the scientific community**
 - **identification of genes influencing traits of interests**
 - **study of genetic diversity in all reigns,**
 - **study of gene expression regulation**
 - **sequencing and genomic comparison.**
- **To Develop new protocols, new methodologies, acquire expertise and train in those technologies**
- **To animate workshop for user network**



Who are we?

- Genomics and transcriptomics core facility spread on 5 sites GeT in Toulouse
- National Infrastructure within the « France Génomique » program
- IBISA Label
- INRA strategic core-facility
- ISO9001 et NFX0800 Certification



Team and Expertise

- **A team of 30 people with:**
 - **Technical Specialty and scientific community by site**
 - **Experts in Agronomy, Environment, Microbiology, Health**
 - **Competence in biology, bioinformatics, biostatistics**



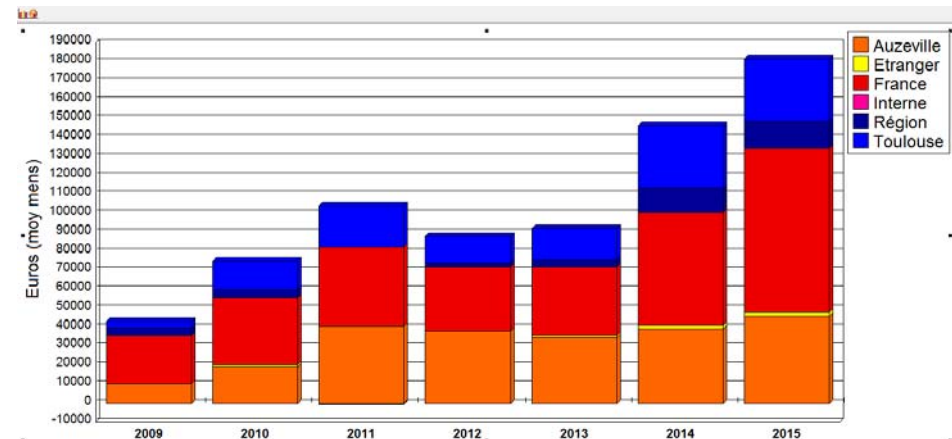
- **Partnership with Genotoul Bioinformatic core-facility (NG6) for:**
 - **Data storage and management**
 - **Data Quality analysis**

ACTIVITY

Activity per year

- **15 R&D projects**
 - Mate paired, chipseq, metylation, GBS, 3G NGS, capture, HIC...

- **More than 100 laboratories (INRA, CNRS, INSA, INSERM, CHU, CIRAD ...)**
 - More than 160 research teams
 - More than 250 projects
 - 2M€ of activity



Collaboration and partnership

- **17 projects in collaboration (France – Europe)**

- **CARTOSEQ** (2010-2014) INRA : Identification en masse des variants génétiques influençant les caractères d'élevage chez les trois principales races laitières françaises
- **FUNHYMAT** (2011-2015) Université Pau: Structure et fonctionnement de tapis microbiens contaminés avec des hydrocarbures
- **Domestichick** (2013-2016) INRA: De la génomique du genre Gallus à l'histoire de la domestication du poulet
- **EFFECTOORES** (2013-2016) : Exploitation des connaissances sur les effecteurs des Oomycetes pour la recherche de résistances durables aux maladies chez les plantes cultivées
- **IMPACT** (2014-2016) INRA: Identification of Matrix Proteins Affecting CalciteTexture in chicken and guinea fowls eggshells
- **PigLetBiota** (2014-2019): Une étude de biologie intégrative de l'influence du microbiote intestinal sur la robustesse des porcelets
- **AgENCODE** (2015-2016): a French pilot project to enrich the annotation of livestock genomes
- **Treasure** (2015-2019- H2020): Treatment and Sustainable Reuse of Effluents in semiarid climates
- **Bovano** (2015-2017): IDENTIFICATION AND FUNCTIONAL STUDY OF CATTLE DELETERIOUS MUTATIONS
- **Feed-A-Gene** (2015-2020 – H2020) : projet européen pour améliorer l'efficacité alimentaire des monogastriques
- **Vaisseaux et Cancer** (2013-2016 - INCA) : caractérisation moléculaire des vaisseaux qui contribuent à l'inhibition de la croissance tumorale
- **SELGEN - GenSSeq**: développement et la mise en œuvre des méthodes à haut débit permettant d'estimer la valeur génétique des animaux et végétaux
- **Agri-Métatranscriptomique** – diversité: Nouvelles perspectives dans l'étude des communautés microbiennes complexes
- **HeliOr** (2015-2018) : Séquençage du Génome de l'Orobanche et 2^{ème} génotype Tournesol dans le cadre de SUNRISE
- **Meta-Pac** (2015-2016): Mise au point des analyses de Metagénomique long read
- **B-TB** (2013 – 2016 ANR) Role of B cells tuberculosis immunity and inflammation.



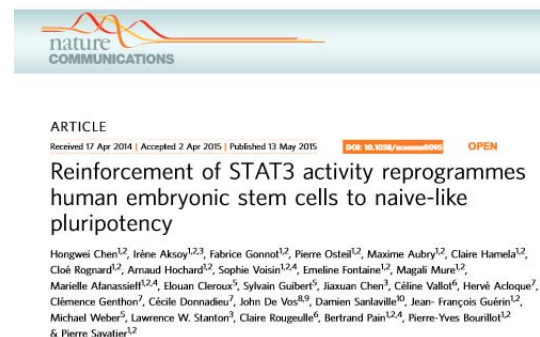
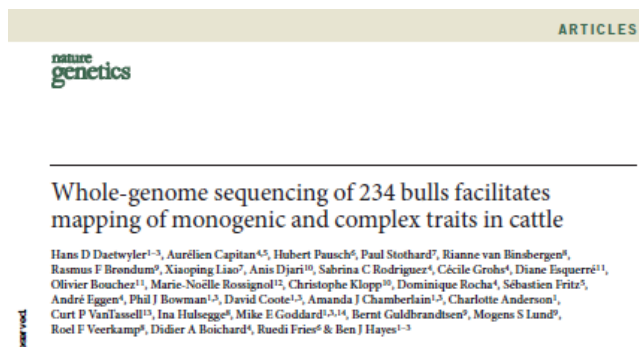
Collaboration and partnership

- **46 publications as co-author since 2012**

2012 (14) : PLoS ONE, NAR, Gene, Current Biology, Developmental and Comparative Immunology, Gene 500, Ecotoxicology, N, Biochimieew Phytologist, FEMS, Ecotoxicology, Nucleic Acids Research, PLOS Genetics, Hepatology

2013 (13) : Biotechnology for biofuel, BMC Genomics, Mol Ecol ressources, infection genetics and evolution, PLoS pathogene, BMC, molecular ecology, genome amouncements, molecular phyloenetics and evolution, FEMS, Gastroenterology & Hepatology, Biochem Biophys Res Commun, Front. Microbio

2014 (19): nature genetics, BMC Genomics, JAS, Exp and molecular pathology, Gene Nar, Molecular Biology ressources, EMBOJ, Emerging Infectious Diseases, molecular cell, PLoS ONE, PLoS Biol, Poultry science, FEMS, Biochemistry & Molecular Biology, Toxicology, Medecine, Research & Experimental, J Am Soc Nephrol, Pathologie Biologie, Journal of Biotechnology



TECHNOLOGIES

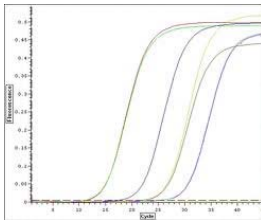
Tools to improve the activity

- **Sample and library quality controls**
- **Single cell capture (C1 Fluidigm)**
- **Pipetting platforms for sample preparation**
 - Partnership with Tecan (4 Evo + 1 genesis), Agilent Bravo
 - Acces array (fluidigm)



Tools to analyse gene expression and to genotype

PCR quantitative



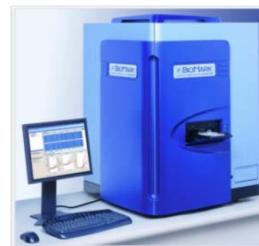
**Via7, QuantStudio,
ABI7900HT, ...**



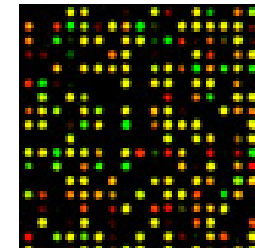
PCR quantitative microfluidic



BioMark (Fluidigm)



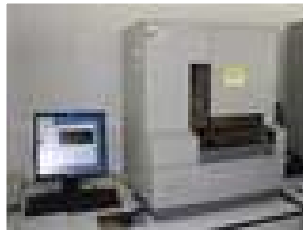
Microarray



**Affymetrix – Agilent
iScan**



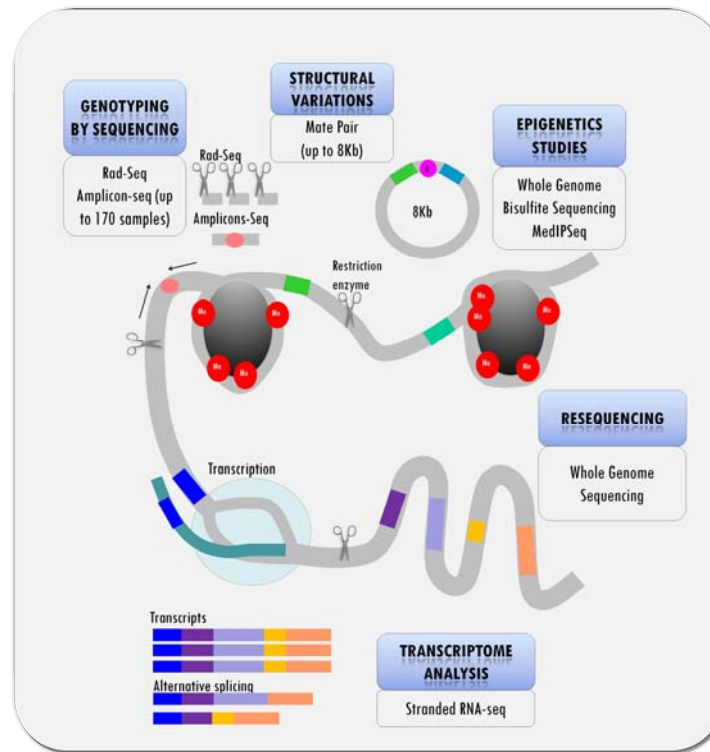
From Sanger to short read NGS revolution



400 pb
1Gb



200 pb
13 Gb



2x 300 pb
15 Gb



2x150 pb
700 Gb

From 1 human genome sequencing to....

... « 1000 genomes » projects for all species

From NGS short reads to long reads

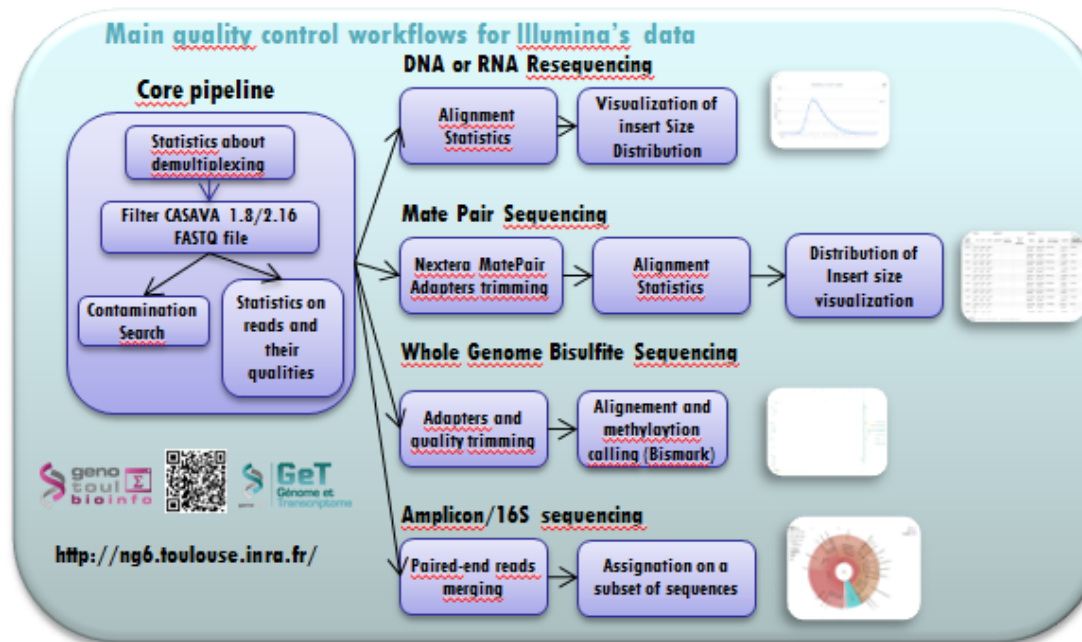
- **First PACBIO installed in France**
- **SUNRISE project to acquire expertise**
 - To validate quality of DNA
 - To improve librairie preparations
 - To increase the number of reads
 - To increase the length of reads
- **more projects to develop new applications :**
 - Whole genome sequencing on different species
 - Targeted sequencing
 - Complex population
 - RNA sequencing
 - Epigenetic



LIMS

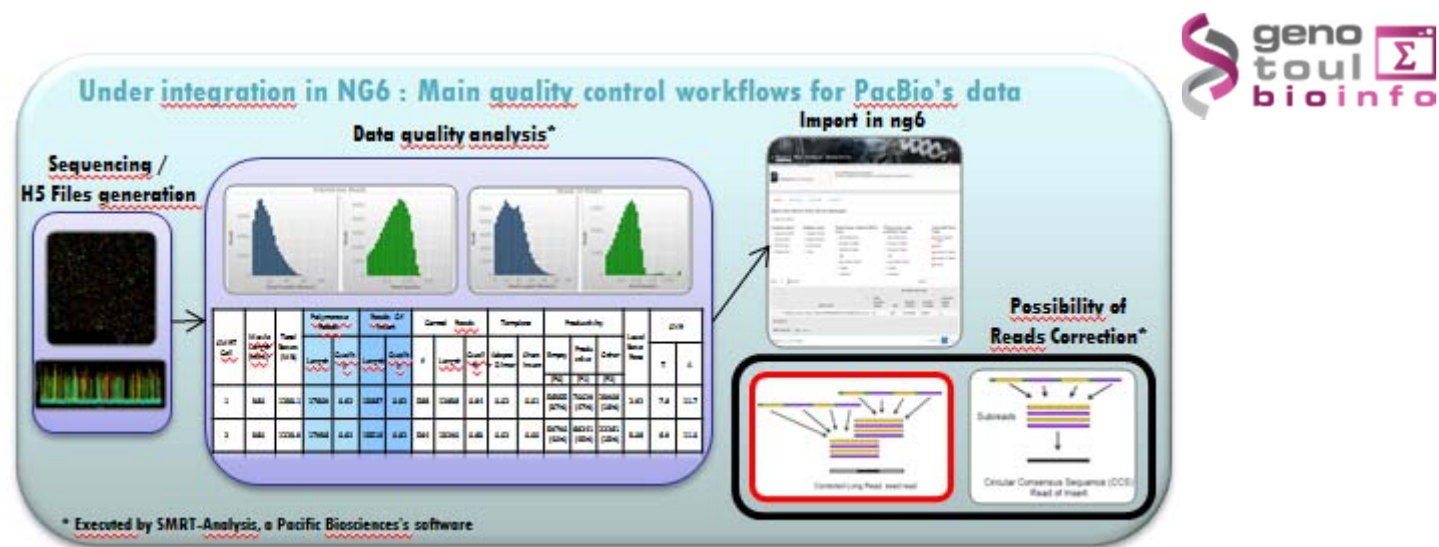
Storage of raw reads and quality control results

- Each step of the process is **tracked** by various modules of our in-house information system.
- Biological samples are uniquely identified using **barcodes**.
- NGS pipelines are provided by **NG6**, an information system built upon the jflow workflow management system



Storage of raw reads and quality control results

- **Current integration in NG6 : Main quality control workflows for PacBio's data**



- **Upcoming : A new LIMS for NGS samples, sequencing and analysis tracking**

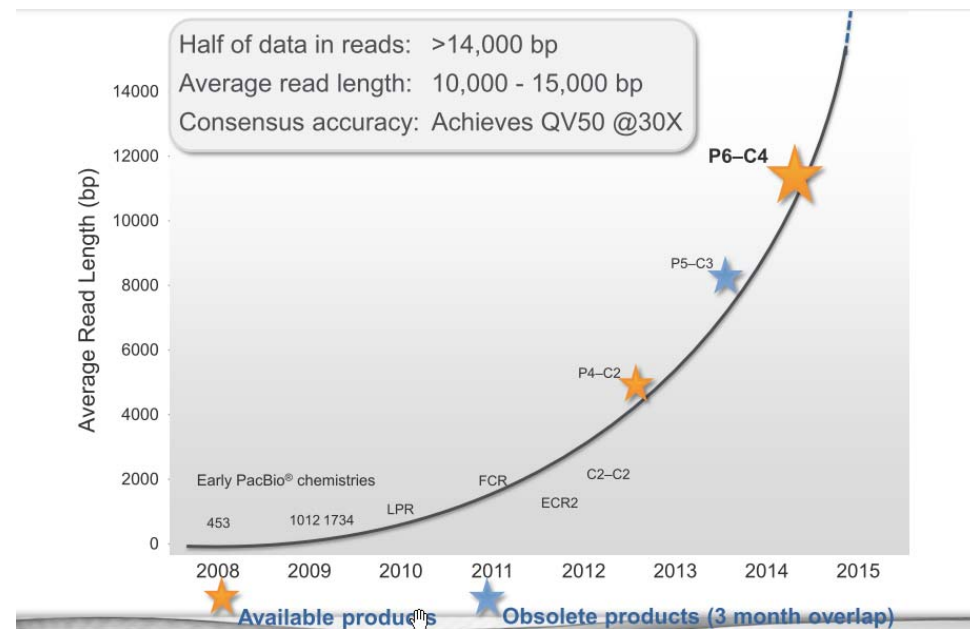


« LONG READ SEQUENCERS »

Principle of PACBIO RSII

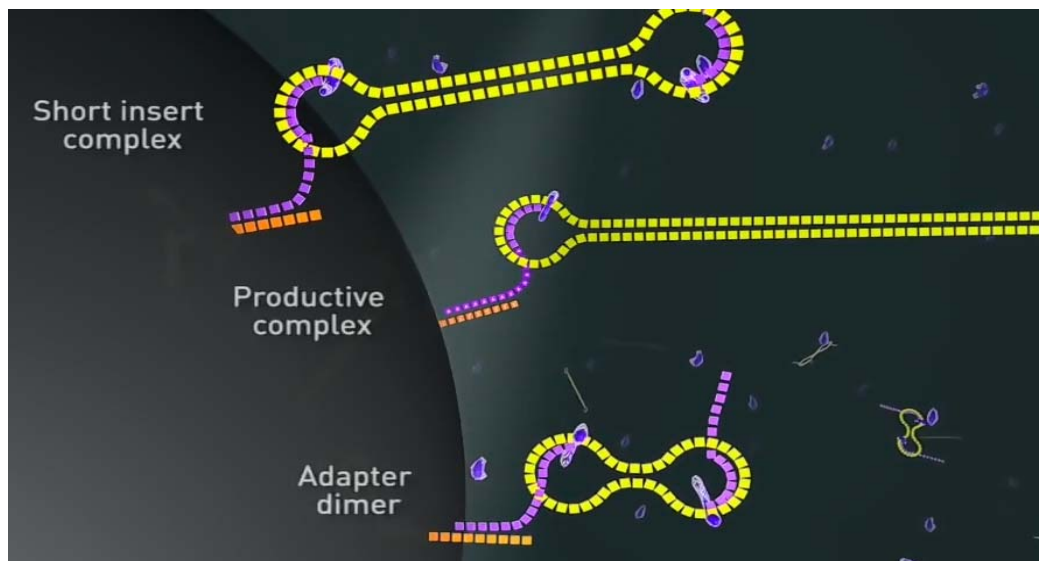
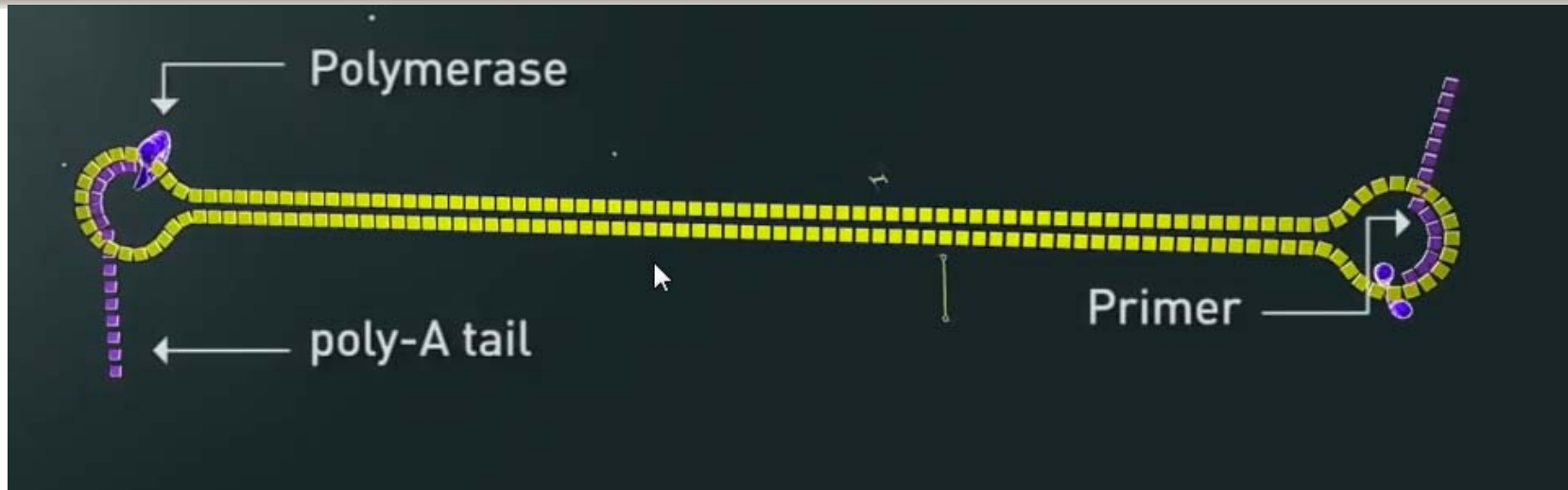
- **No PCR**
- **Native Enzyme**
- **Real-Time (SMRT®) technology**

- **Phospholinked nucleotides**
- **A technological revolution with P6-C4 chemistry**
 - **>10kb, 6h**
 - **Access to complex genomes**



First equipment in France installed in Toulouse in March 2015
(European and "Midi-Pyrénées" funding program)

Principle of PACBIO RSII

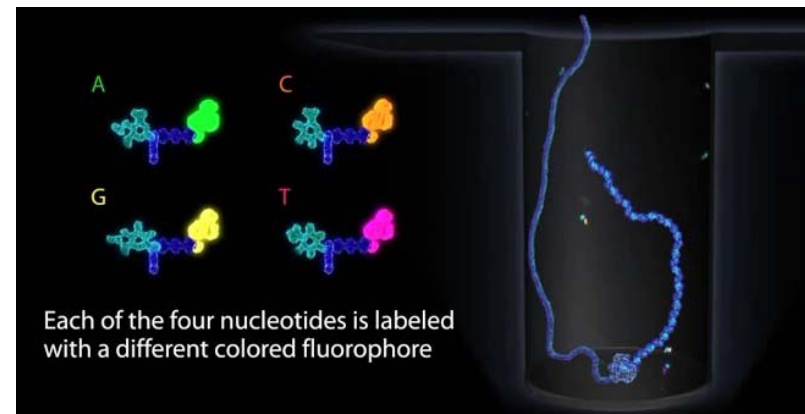
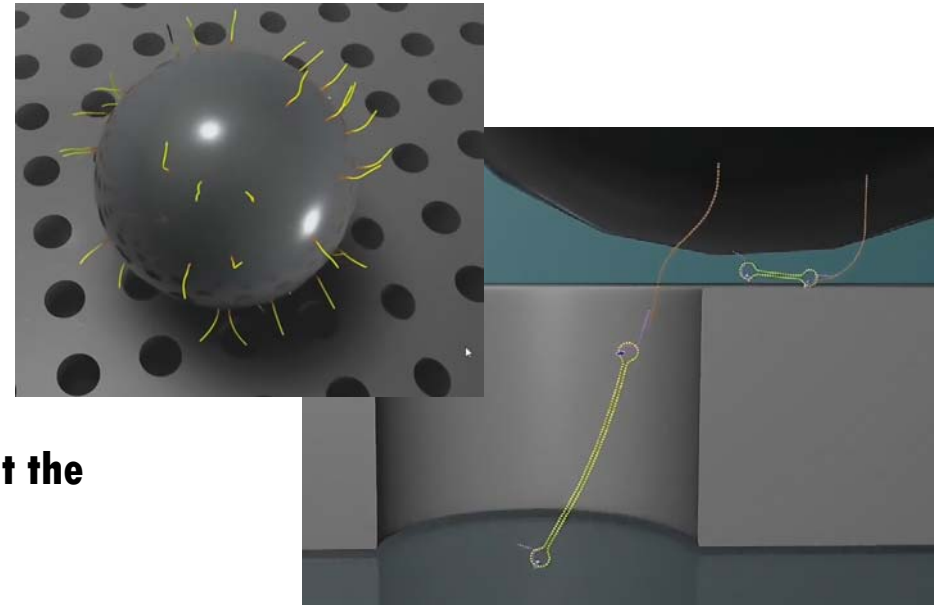


Principle of PACBIO RSII

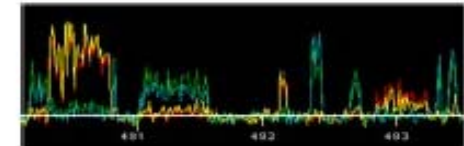
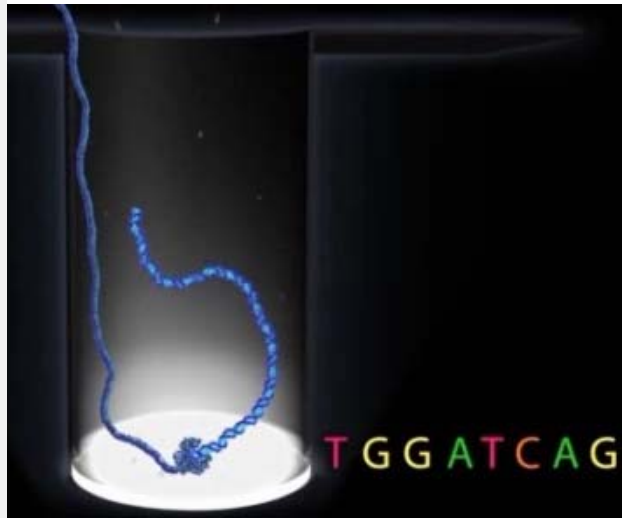
- **1 SMRTcell**
- **150 000 Zero-Mode Waveguides (ZMWs)**
- **50 000 to 70 000 usable ZMW**

- **DNA polymerase complex is immobilized at the bottom of the ZMW**

- **Nucleotides incorporation and release of the attached fluorophore**



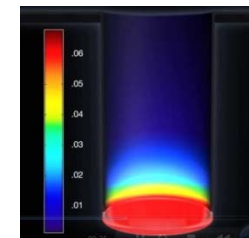
Principle of PACBIO RSII



GCAACGATCACCTAAA...GCAACGA
 TCACCTAAA...GCAACGATCACCTA
 AA...GCAACGATCACCTAAA...

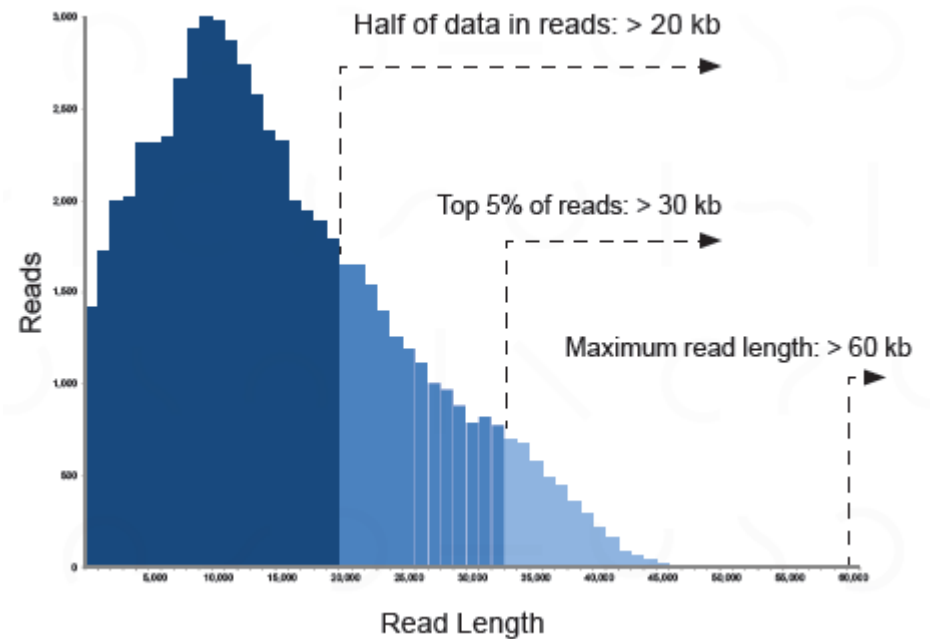
ACGATCACCTAAA...

- **Laser illuminates the ZMW from below and capture the fluorophore signal**
- **Read of the sequence in real-time**
- **Translation of the signal according to intensity and length of signal**
- **Format of files H5 and fastq**



Specification of PACBIO RSII

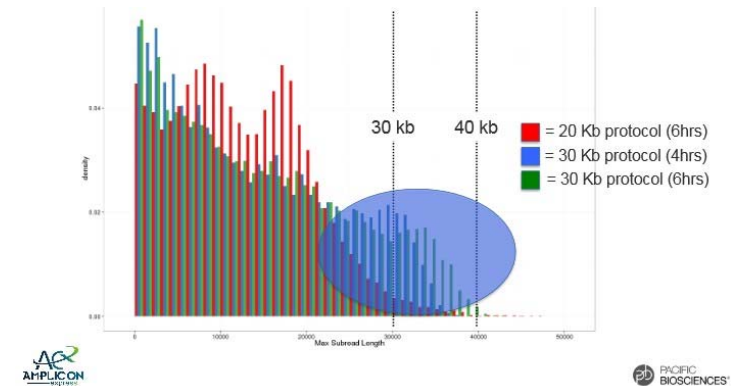
- **DNA : 10µg of high molecular weight level ***
- **Data per SMRT Cell: 500 Mb – 1,2 Gb**
- **Read length > 15 kb**
- **Error : 15% random (insertion – deletion)**
- **Run time : 4 to 6h**
- **Runs per week : 16 to 32 SMRTcell (mini 8)**
- **3,7pb/s**
- **Reagent Cost : 500€/SMRTcell**



Benefits and Disadvantages RSI I

- **Benefits**
 - **Good Results : to small genome (Mb) to complex genome >3Gb**
 - **Uniform coverage : high % of GC, repeated sequences, ...**
 - **Error : random**
 - **Accuracy : function of coverage**
 - **Variety of analysis methods available through SMRT Software**
 - **No need to combine with short read data**
 - **Simultaneous Epigenetic Characterization**
- **Disadvantages**
 - **High molecular weight level : difficult for some species**
 - **High % of Error : need a high coverage**
 - **Low production : 3Gb 100X = 20 week**
 - **High reagent price : 3Gb 100X = 150k€**

- **30% of reads >30kb : new protocol on PACBIO RSII**



- **Sequel system**

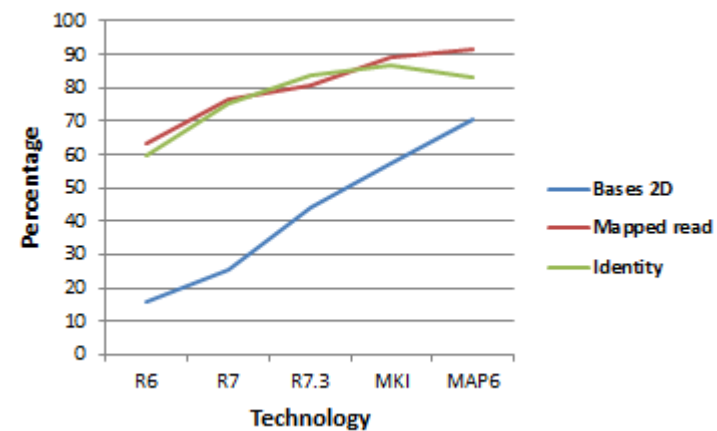
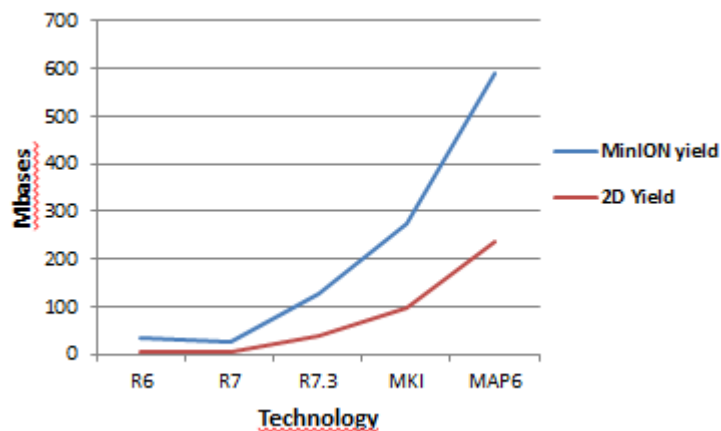
- System ~2X less : 350k€
- Generates ~7X more reads : 7Gb per SMRTcell
- Reagent cost ~2/4X less / Gb

- **In 2016, 5 Sequel systems will be installed in France Genomic core-facilities (France Genomique)**



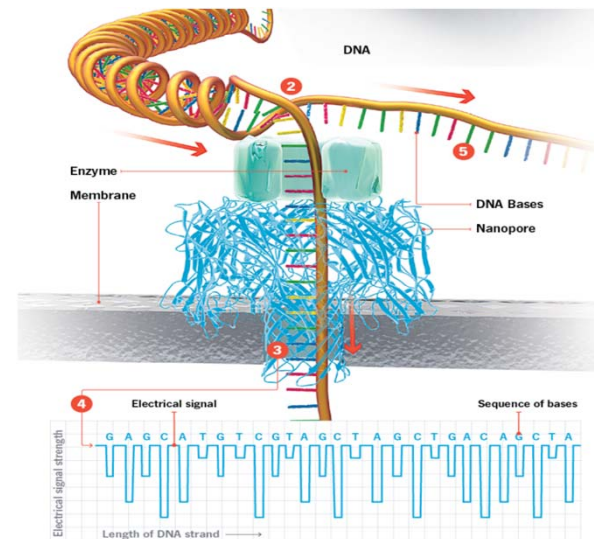
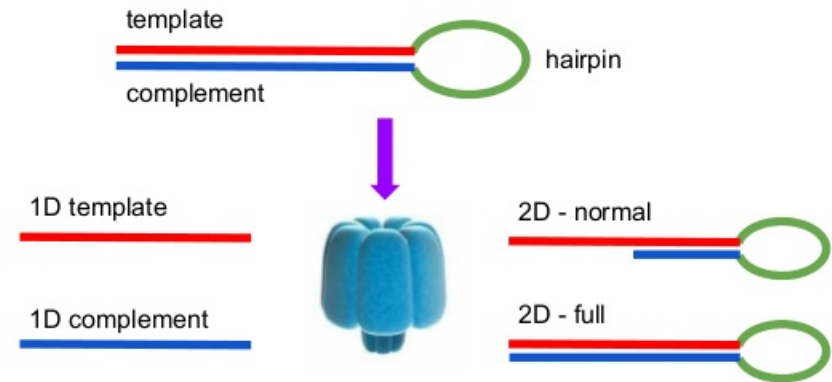
Principle of Oxford nanopore

- **A pocket sequencer**
- **Single strand DNA is passed through a nanopore**
- **Motrice Enzyme and nanopore reader (transposase)**
- **Detection of variation of current when DNA moves through the pore**
- **Fast evolution of the technology**



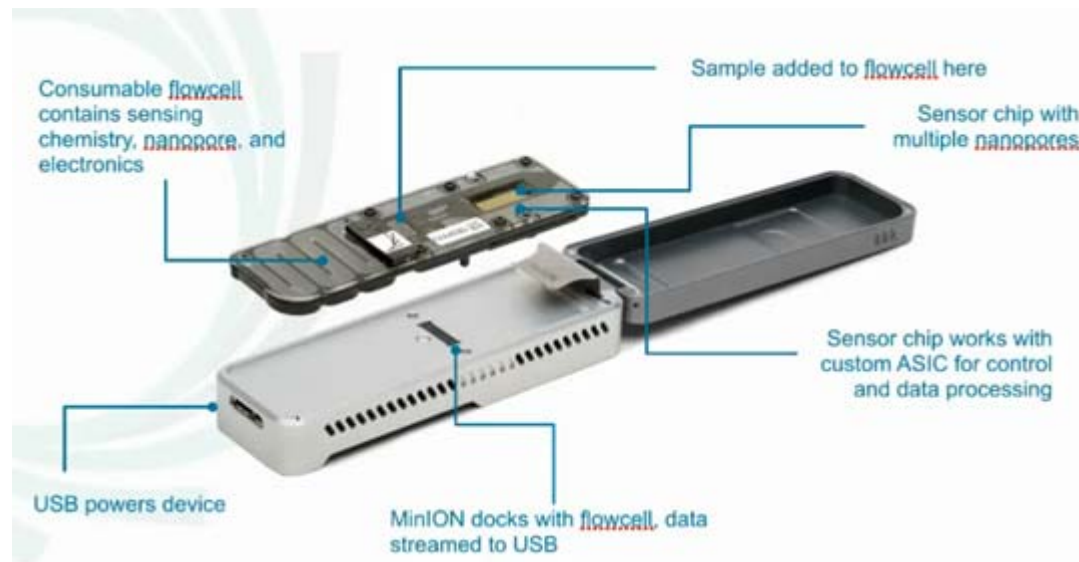
Principle of Oxford nanopore

- **1 Minion**
- **512 circuits (4 pores/circuits)**
- **DNA preparation : 1D or 2D**
- **DNA capture : by motor enzyme**
- **Run time: 48h it's possible to sequence several samples in the same run**
- **Signal : measures 6 bases simultaneously**
- **R7: 70 b/s**
- **Basecalling : on Oxford cloud**
- **Files format: H5 and fastq**



Specification of Oxford nanopore

- **DNA : of high molecular weight level ***
- **Data per Minion R7: 600Mb 1D – 250Mb 2D**
- **Read length R7: 8kb**
- **Error R7: 15% D2 not random**
- **Run time : 48h**
- **Run per week : 3**
- **Reagent Cost : 500€/Minion**



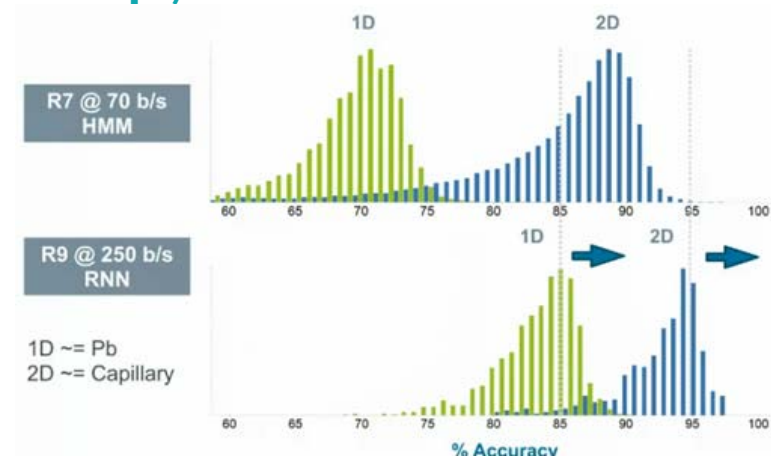
Benefits and Disadvantages

- **Benefits**

- Promising technology... constantly changing
- Interesting for small genome : 12 – 25Mb (Genoscope)
- Easy to install
- Cheap technology

- **Disadvantages**

- No random error
- Basecalling on Oxford cloud
- Unsuitable for complex genome
- Need to be combined with short read data



Promising technology

	Mk 1 MinION R9	Single PromethION Flow Cell	PromethION (48 Flow Cells)
Number of channels available for sequencing	Up to 512	Up to 3,000	Up to 144,000
Run time ⁴	1 minute - 48 hours	1 minute - 48 hours	1 minute - 48 hours
Flow cell lifetime ⁴	~72hrs	>= 72hrs	>= 72hrs
Number of reads at 10Kb at standard speed (280bps) ⁴	Up to 2.5M	Up to 14.5M	Up to 700M
Number of reads at 10kb in Fast Mode (500bps) ⁴	Up to 4.4M	Up to 26M	Up to 1250M
Read Length	8kb	no limit	no limit
1D Yield ⁵ at 280 bps in 48 hours	2Gb	Up to 145 Gb	Up to 7 Tb
1D Yield ⁵ at 500 bps in 48 hours	4Gb	Up to 256 Gb	Up to 12 Tb
Base calling accuracy ⁶	85% 2D	Up to 96%	Up to 96%
Flow Cell Cost (depending on order type and volume) ⁸	\$500- \$900	POA	POA
Instrument Access Fee	Starter kit \$1000	\$75K	\$75K





Sample preparation

Blocking Agents



- **Polysaccharides**
- **Lypopolysaccharides**
- **Growth media residuals**



- **Chitin**
- **Fats**
- **Proteins**
- **Pigments**



- **Chitin**
- **Protein**
- **Secondary metabolites**
- **Pigments**
- **Growth media residuals**

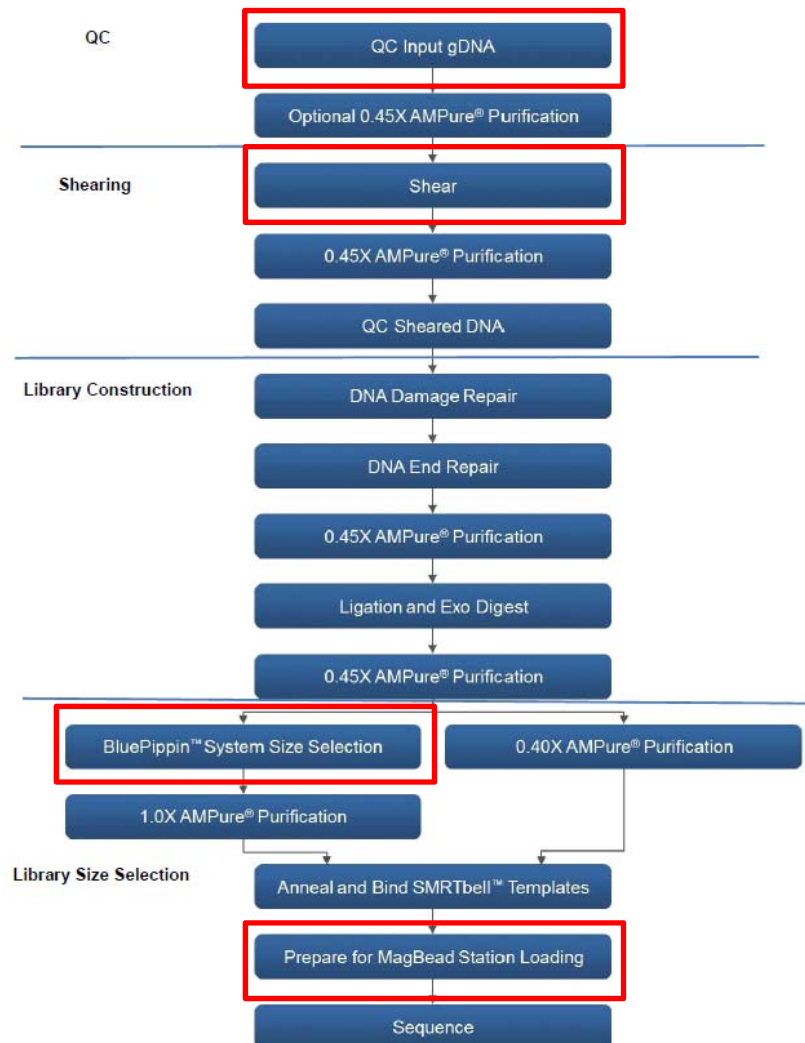


- **Polyphenols**
- **Polysaccharides**
- **Secondary metabolites**
- **Pigments**

Prerequisite for PACBIO RSII

- **Extraction kit advise by Pacific Bioscience**
 - **Qiagen MagAttract HMW kit - Qiagen Genomic-tip kit - Qiagen Genra Puregene**
- **Quantity (Picogreen/QuBit) : >10 µg DNA for 1 library**
- **Purity (Nanodrop) :**
 - **260/280 : 1.8-2**
 - **260/230 = 2-2,2**
- **Size >50kb perfect**
- **Advice:**
 - **No Freezing – Defreezing**
 - **No temperature >65°C**
 - **No extreme pH (<6 ou>9)**
 - **No vortex**
 - **No fluorescent or ultraviolet**

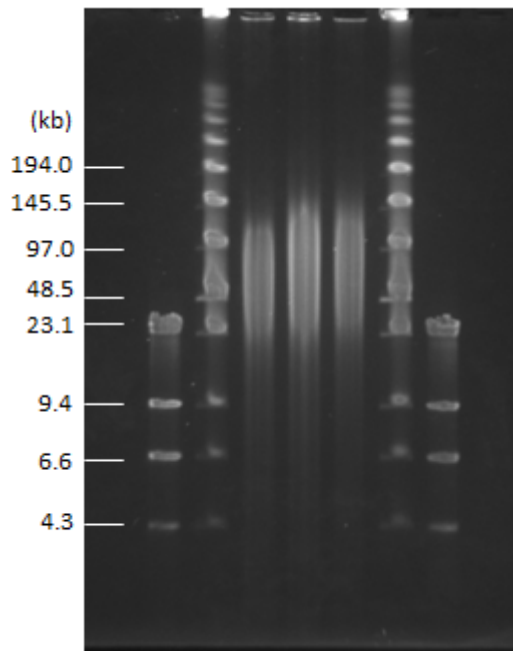
Library process on GeT core facility



- **4 systematic different QC "GeT" core-facility**
 - Capillary electrophoresis – Fragment Analyser (50kb)
 - Pulsed Field – Pippin Pulse
 - Spectrometrics based – nanodrop
 - Intercalent based – Qubit
- **Shearing with Megaruptor**
- **Size selection with Blue pippin**
- **DNA is a living element, it will evolve during the process**

Example for Fish – 900Mb – 54X – 48 SMRT

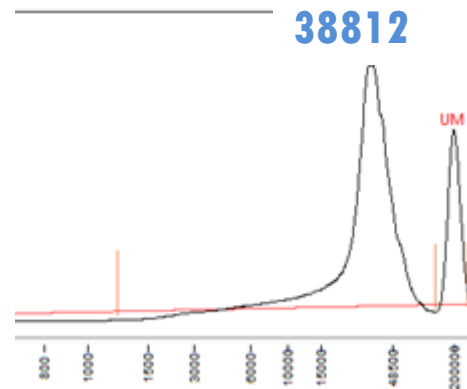
Pulsed Field



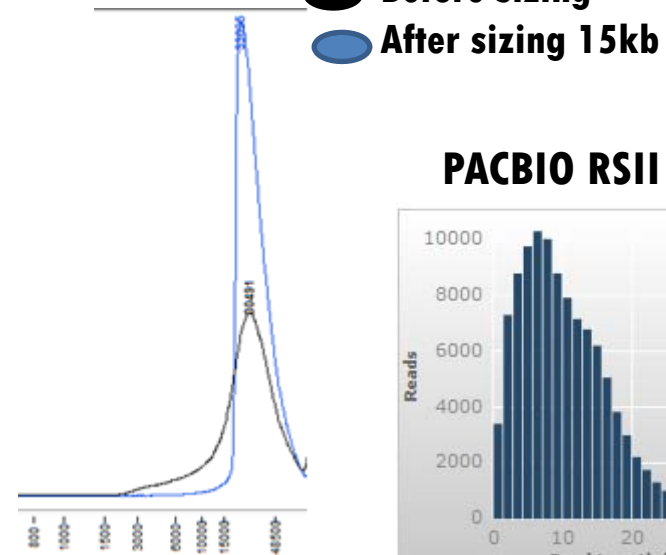
Spectrometrics - Intercalent

Sample ID	Conc.	Units	260/280	260/230	Cursor abs.	Qbit 1/10	Valeur
javf3	317,7	ng/ul	1,84	1,95	6,354	47,4 ng/ul	474 ng/ul

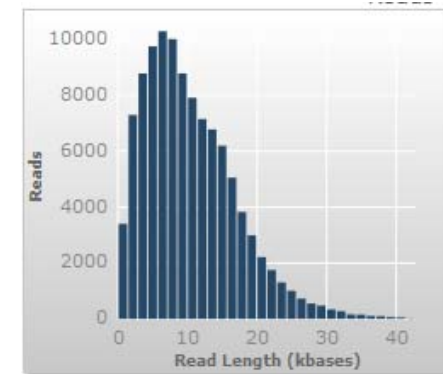
Fragment Analyser



Sizing



PACBIO RSII Reads

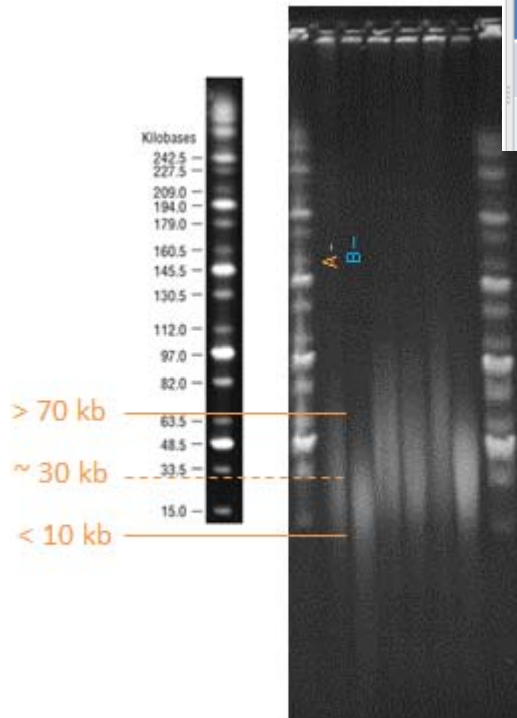


PACBIO RSII Run

Total Bases (MB)	Polymerase Reads		Reads Of Insert		Control Reads			Template		Productivity		
	Length	Quality	Length	Quality	#	Length	Quality	Adapter Dimer	Short Insert	Empty (P0)	Productive (P1)	Other (P2)
1398.82	12990	0.84	10816	0.85	161	21196	0.86	0.03	0.01	12239 (8%)	107685 (72%)	30368 (20%)

Exemple for Phytoplankton – 13Mb – 120X – 3SMRT

Pulsed Field

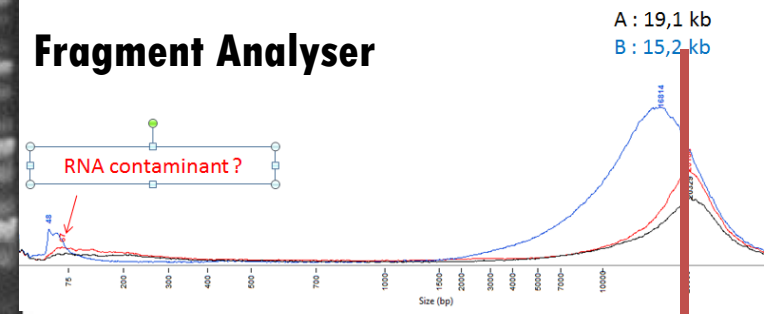


Spectrometrics - Intercalant

Sample name	Qubit Concentration (ng/μL)	NanoDrop Concentration (ng/μL)	Volume (μL)	Quantity (μg) (Qubit)	A260/280	A260/230	Length (pb) (Fragment analysis)
A	65	116	80	5,2	2,01	2,2	18068
B	100	167	50	5	2,02	1,97	15270

RNA contaminant ?

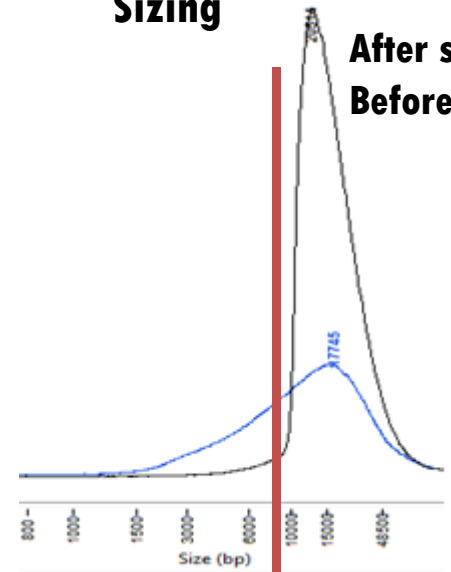
Fragment Analyser



A : 19,1 kb
B : 15,2 kb

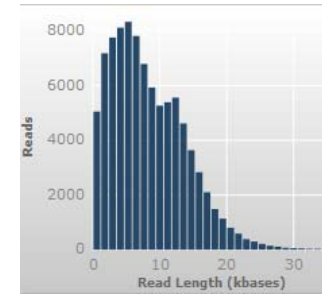
20kb

Sizing



9kb

After sizing 9kb
Before sizing



PACBIO RSII Run

Total Bases (MB)	Polymerase Reads		Reads Of Insert		Control Reads			Template		Productivity		
	Length	Quality	Length	Quality	#	Length	Quality	Adapter Dimer	Short Insert	Empty (P0)	Productive (P1)	Other (P2)
1028.44	11188	0.85	8640	0.86	419	17781	0.85	0.03	0.01	42865 (29%)	91924 (61%)	15503 (10%)

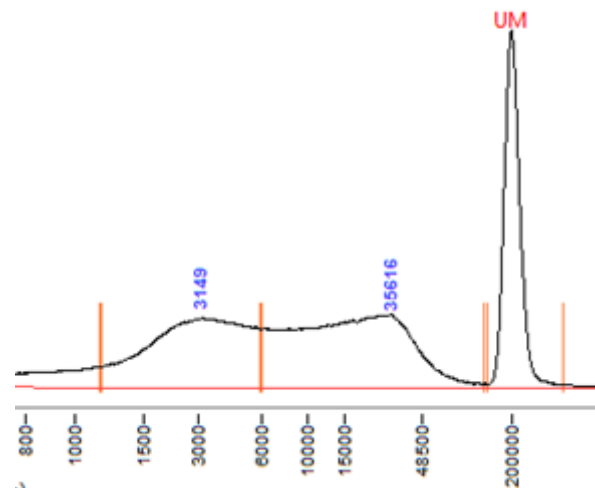
PACBIO RSII Reads

Bouillon of bacterial culture – 5-9Mb - ?- 5SMRT

Spectrometrics - Intercalent

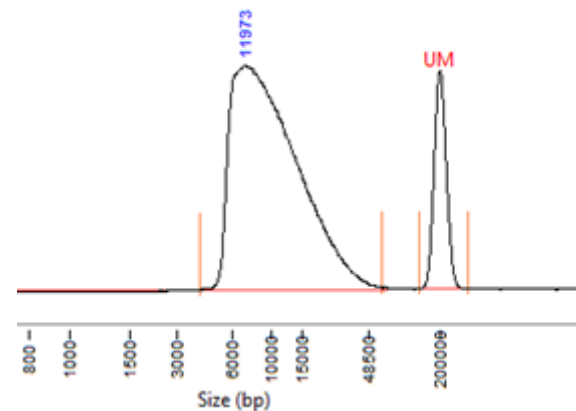
Sample ID	Conc.	Units	Nanodrop				Qbit (ng/μl)	
			A260	A280	260/280	260/230	dil 1/5	Echantillon
..... XI-2013	43,3	ng/ul	0,866	0,416	2,08	0,93	2,42	12,1

Fragment Analyser

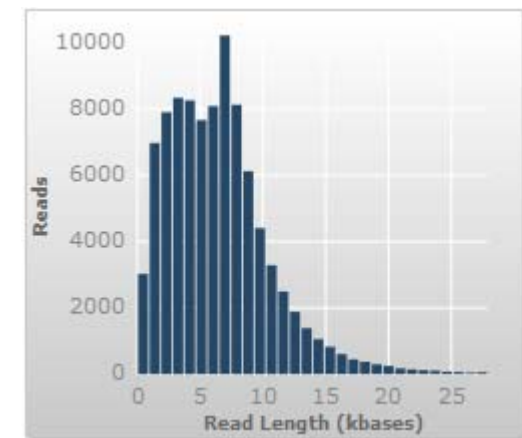


Sizing

After sizing 6kb



PACBIO RSII Reads



PACBIO RSII Run

Total Bases (MB)	Polymerase Reads		Reads Of Insert		Control Reads			Template		Productivity		
	Length	Quality	Length	Quality	#	Length	Quality	Adapter Dimer	Short Insert	Empty (P0)	Productive (P1)	Other (P2)
1135.39	12184	0.84	6775	0.86	180	18681	0.83	0.01	0.00	12961 (9%)	93189 (62%)	44142 (29%)



More results on sunflower

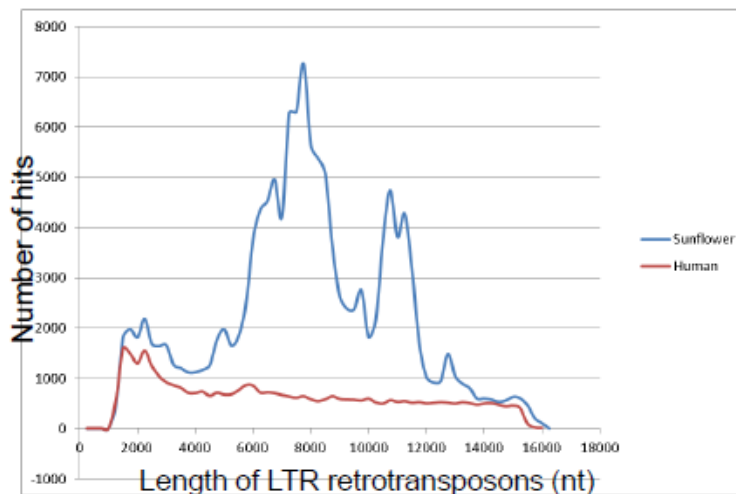
Nicolas Langlade, Stephane Munoz, Jérôme Gouzy,
Baptiste Mayjonade,



Goals of the project

- **Improve the sunflower genome assembly of the XRQ line by sequencing :100X depth with PacBio sequences only.**
- **Diploid with $2n = 17$ chromosome pairs,**
- **3.6Gb,**
- **Lot of repeated sequences : large 9-12kb and highly conserved.**

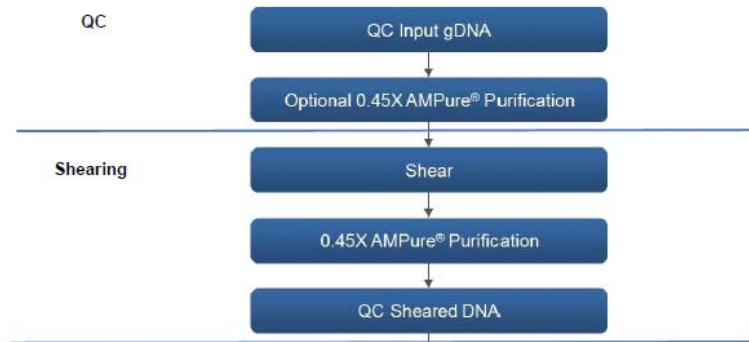
Analysis of the composition of the LTR retrotransposons with LTRharvest (D. Ellinghaus *et al.* 2008, default parameters)



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

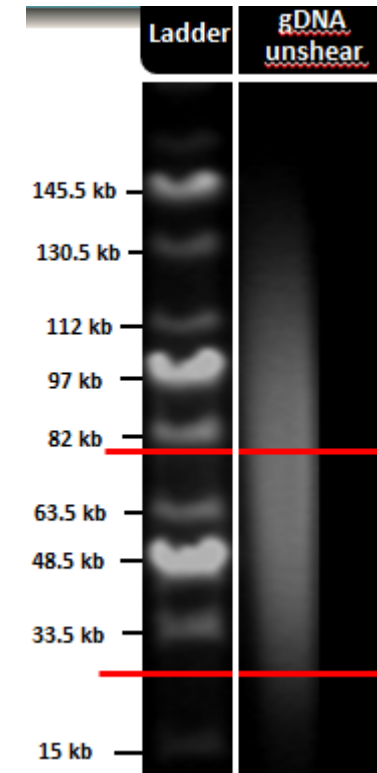
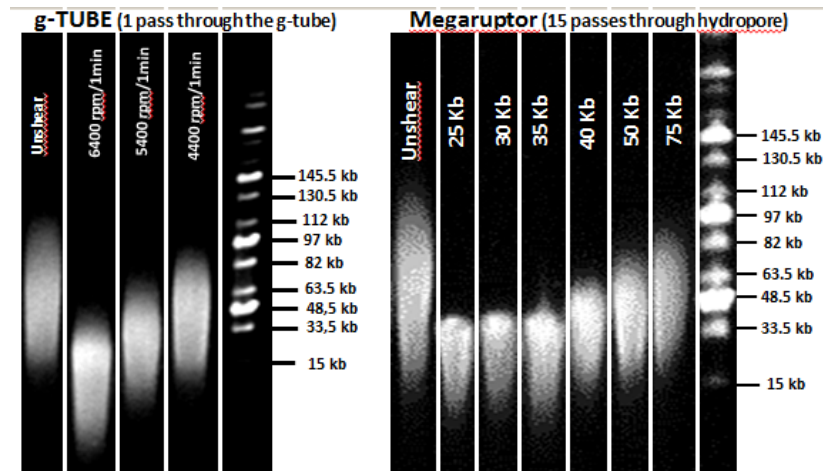
8.8% of the human genome.

Improvement of large insert library prep

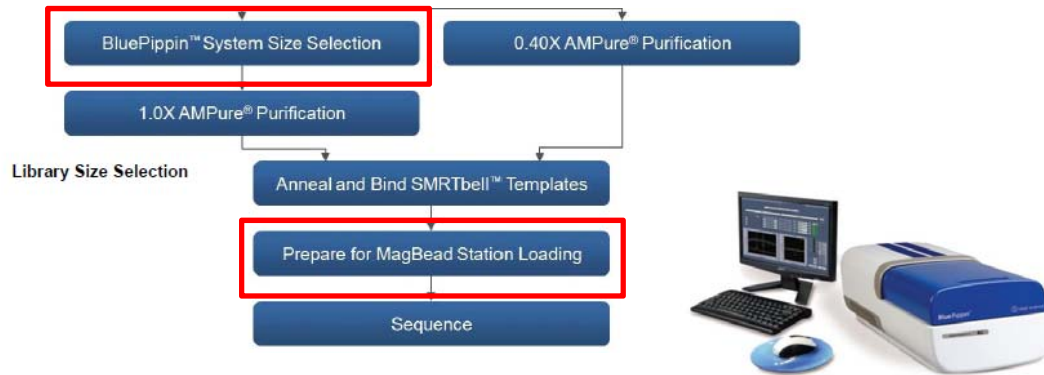


- **High molecular weight DNA (50 to 100kb)**

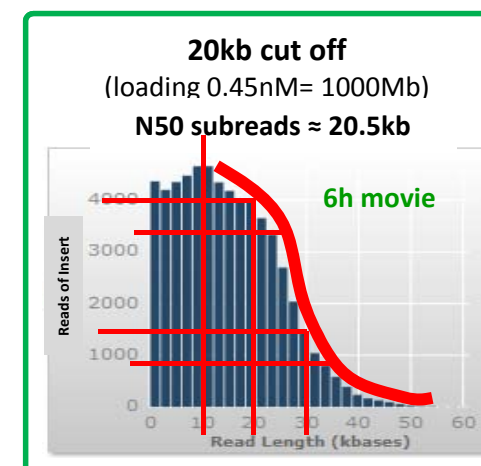
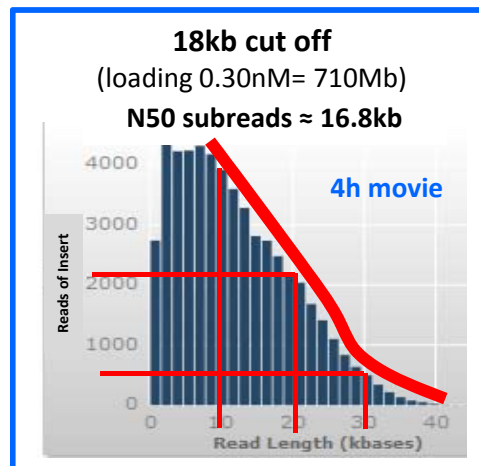
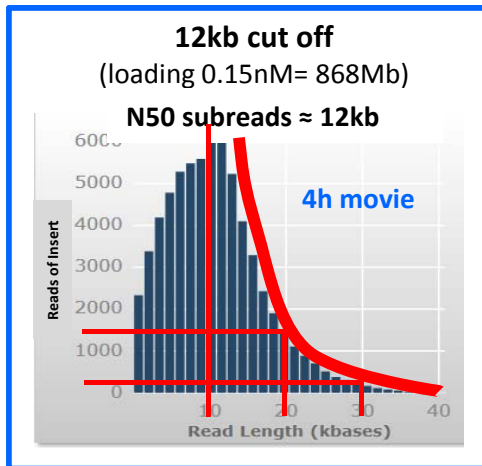
- **Shearing with megaruptor**



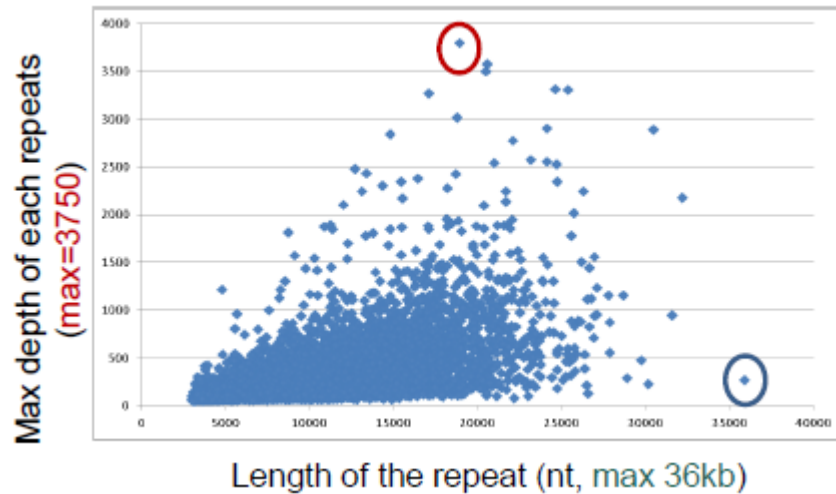
Improvement of large insert library prep



Size selection cutoff	% Recovery (bluepippin input ≈ 4μg)	Migration time
10kb (High pass 6-10kb)	55-65%	2h
15kb (High pass 15-20kb)	40-48%	3-4h
18kb (High pass 15-20kb)	30-35%	4-5h
20kb (High pass 15-20kb)	18-20%	4-5h
30-40kb High pass 30-40kb	????	????



Final Results



Top 10 of our longest subreads

80974 bp
 79860 bp
 79834 bp
 78105 bp
 77481 bp
 76881 bp
 76558 bp
 76355 bp
 75569 bp
 75559 bp

- INRA (Team Sunflower) :**
HiSeq 127 X → 43 % coverage
- International project : 454, HiSeq, Genetic and physical map (finger printing of the BAC clones) → 63 % coverage**
- INRA (Team Sunflower)) :**
PacBio 107 X (407 SMRT)
→ 84 % coverage
3,03Gb
13124 contigs N50 = 498 kb
+90 % anchored

THANKS

GeT-PlaGe Team

Denis Milan
Gerald Salin
Olivier Bouchez
Diane Esquerre
Anne Fleurbe
Sandra Fourre
Isabelle Hochu
Céline Vendecastel
Dounia Ben Salah
Claire Kuchly
Gaelle Vilchez
Catherine Zanchetta
Marie Vidal
Pauline Heuillard
Céline Jeziorski
Clémence Genthon
Maarten Pirson
Alain Roulet
Céline Roques
Adeline Chaubet
Frédéric Toppan

LIPM Team

Jérôme Gouzy
Nicolas Langlade
Munos Stéphane
Chris Grassa
Sébastien Carrere
Erika Sallet
Ludovic Legrand
Marie-Claude Boniface
Nicolas Pouilly

GeT-BioPuce

Marie-Ange Teste
Lidwine Trouilh
Nathalie Marsaud
Delphine Labourdette
Sophie Lamarre
Matthieu Guionnet

GeT-TRIX Team

Yannick Lippi
Claire Naylies

GeT-TQ Team

Jean-José Maoret
Frédéric Martins

GeT-Purpan Team

Nicolas Borot
Nathalie Jonca
Emeline Lhuillier
Marion Roy



Programme de la journée du 27 mai 2016

<https://seminaire.inra.fr/pacbio-get-plage>

9h Accueil

- 09h20 – 09h30 Evolutions de la plateforme GeT en 2016 - Denis Milan (INRA Directeur Scientifique GeT)
- 09h30 – 09h40 Qualité des matrices et prérequis - Alain Roulet (INRA GeT-PlaGe)

Session De novo

- 09h45 – 10h15 Les génomes complexes du tournesol (3.6Gb) et de sa plante parasite *Orobanche cumana* (2Gb), assemblés grâce à la technologie PacBio - Jérôme Gouzy (INRA Toulouse LIPM)
- 10h15 – 10h45 Chromosome plasticity in the smallest Photosynthetic Eukaryotes *Ostreococcus* (Chlorophyta) - Gwenaél Piganeau (Observatoire Océanologique de Banyuls sur mer)
- 10h45 - 11h15 Pause
- 11h30 - 12h00 Yann Guigen (INRA Rennes, Fish Physiology and Genomics)
- 12h00 - 12h30 Retour d'expérience sur le séquençage de Bac en PacBio chez la poule et le porc - Valérie Fillon & Yvette Lahbib-Mansais (INRA Toulouse GenPhySE)
- 12h30 - 13h00 Oxford Nanopore Technology : Données et applications - Jean-Marc Aury (Institut de Génétique - Genoscope, Evry)

13h00 – 14h00 Repas sur GeT-PlaGe

Session Meta-génomique

- 14h00 – 14h30 En attente confirmation
- 14h30 - 15h00 De novo assembly of individual bacterial genomes from the seed microbiome - Mathieu Barret (INRA Angers IRHS)

Session Autres applications PACBIO

- 15h00 - 15h30 IsoSeq : transcriptomic analysis using long reads - Christophe Klopp (INRA Toulouse Plateforme Bioinformatique)
- 15h30 – 16h00 En attente confirmation
- 16h00 - 16h15 Pause
- 16h15 - 16h45 En attente confirmation

Ouverture ...

- 16h45 - 17h15 Mise en place d'un nouveau protocole sur la plateforme: le séquençage 3D avec l'Hi-C - Sylvain Foissac (INRA Toulouse GenPhySE)
- 17h15 - 17h45 En attente confirmation

Principle of oxford nanopore

