genotoul

GeT
Génome et
Transcriptome

Toulouse Genomic Core facility

Olivier Bouchez

Next Generation Sequencing Lab Manager

olivier.bouchez@toulouse.inra.fr

http://get.genotoul.fr

@GeT_Genotoul

Who are we?

- **Genomics and transcriptomics core facility spreads on 5 sites GeT in Toulouse**
- **National Infrastructure within the « France Génomique » program**
- **IBISA Label**
- **INRA strategic core-facility**
- **ISO9001 et NFX50-900 Certification**

The Core facility missions

- To provide innovating technologies for genome analysis to the scientific community
  - Sequencing / Genotyping
  - Gene expression

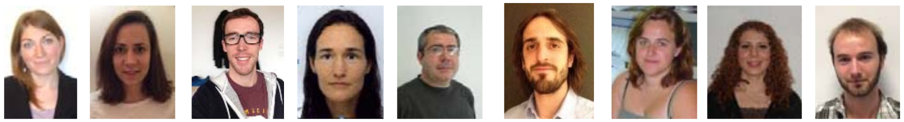
- **To Develop** new protocols, new methodologies, acquire expertise and train in those technologies

- To animate workshop for user network

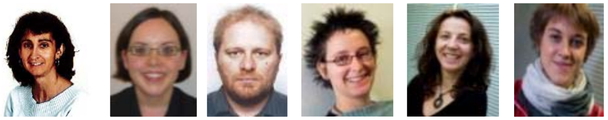
Team and Expertise

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

- GeT- PlaGe (INRA Auzeville):
  - NGS, 2nd et 3rd Generation
  - Bioinformatics
  - Agronomy, Environment, Ecology



- GeT-Biopuces (INSA Rangueil):
  - Microarray, NGS 2^{nd} Generation
  - Bioinformatics, Biostatistics
  - Microbiology



Team and Expertise

- **GeT-TRiX (INRA Saint Martin) :**
  - Microarray,
  - Bioinformatics, Biostatistics
  - Toxicology

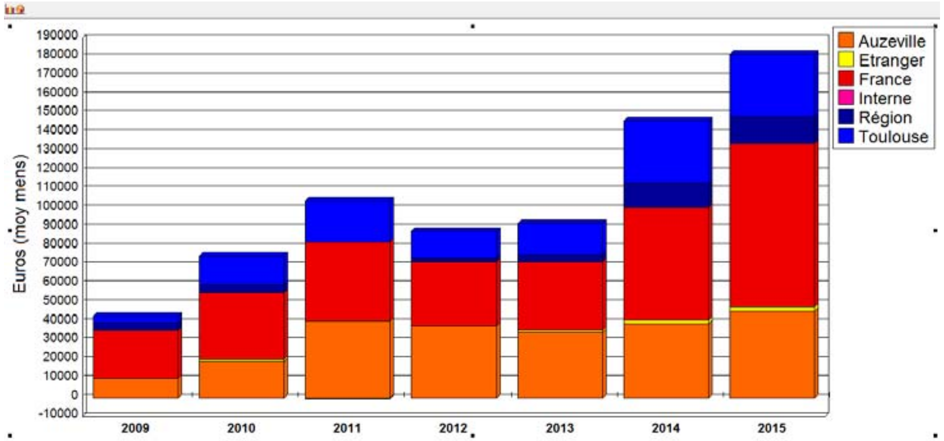
- **GeT-TQ (CHU Rangueil):**
  - QPCR, single cell,
  - Health

- **GeT-Purpan (CHU Purpan) :**
  - Sanger, NGS 2nd Generation
  - Health

ACTIVITY

Activity per year

- **15 R&D projects**
  - Mate paired, chipseq, methylation, 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 contriuent à l'inibition 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 2ème 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.

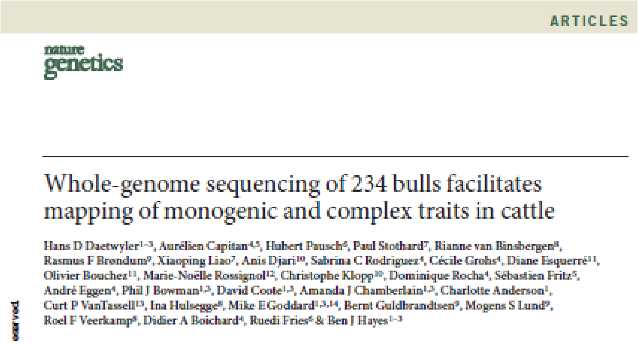
Collaboration and partnership

- **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 phylioenetics 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

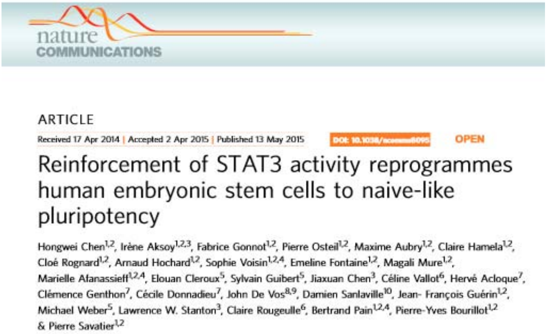


ARTICLES

nature genetics

Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle

Hans D Daetwyler[1–3], Aurélien Capitan[4,5], Hubert Pausch[6], Paul Stothard[7], Rianne van Binsbergen[8], Rasmus F Brøndum[9], Xiaoping Liao[7], Anis Djari[10], Sabrina C Rodriguez[4], Cécile Grohs[4], Diane Esquerré[11], Olivier Bouchez[11], Marie-Noëlle Rossignol[12], Christophe Klopp[10], Dominique Rocha[4], Sébastien Fritz[5], André Eggen[4], Phil J Bowman[1,2], David Coote[1,2], Amanda J Chamberlain[1,2], Charlotte Anderson[1], Curt P VanTassell[13], Ina Hulsegge[8], Mike E Goddard[1,2,14], Bernt Guldbrandtsen[9], Mogens S Lund[9], Roel F Veerkamp[8], Didier A Boichard[4], Ruedi Fries[6] & Ben J Hayes[1–3]



nature COMMUNICATIONS

ARTICLE

Received 17 Apr 2014 | Accepted 2 Apr 2015 | Published 13 May 2015 DOI: 10.1038/ncomms8095 OPEN

Reinforcement of STAT3 activity reprogrammes human embryonic stem cells to naive-like pluripotency

Hongwei Chen[1,2], Irène Aksoy[1,2,3], Fabrice Gonnot[1,2], Pierre Osteil[1,2], Maxime Aubry[1,2], Claire Hamela[1,2], Cloé Rognard[1,2], Arnaud Hochard[1,2], Sophie Voisin[1,2,4], Emeline Fontaine[1,2], Magali Mure[1,2], Marielle Afanassieff[1,2,4], Elouan Cleroux[5], Sylvain Guibert[5], Jiaxuan Chen[3], Céline Vallot[6], Hervé Acloque[7], Clémence Genthon[7], Cécile Donnadieu[7], John De Vos[8,9], Damien Sanlaville[10], Jean- François Guérin[1,2], Michael Weber[5], Lawrence W. Stanton[3], Claire Rougeulle[6], Bertrand Pain[1,2,4], Pierre-Yves Bourillot[1,2] & Pierre Savatier[1,2]

TECHNOLOGIES

Tools to improve the activity

- Sample and library quality controls

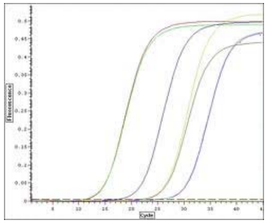


- Single cell capture (C1 Fluidigm)


- Pipetting platforms for sample preparation
  - Partnership with Tecan (4 Evo), Agilent Bravo
  - Access array (fluidigm)


Tools to analyse gene expression and to genotype

PCR quantitative

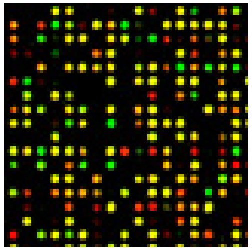


ViiA7, QuantStudio, ABI7900HT, …

PCR quantitative microfluidic

BioMark (Fluidigm)

Microarray

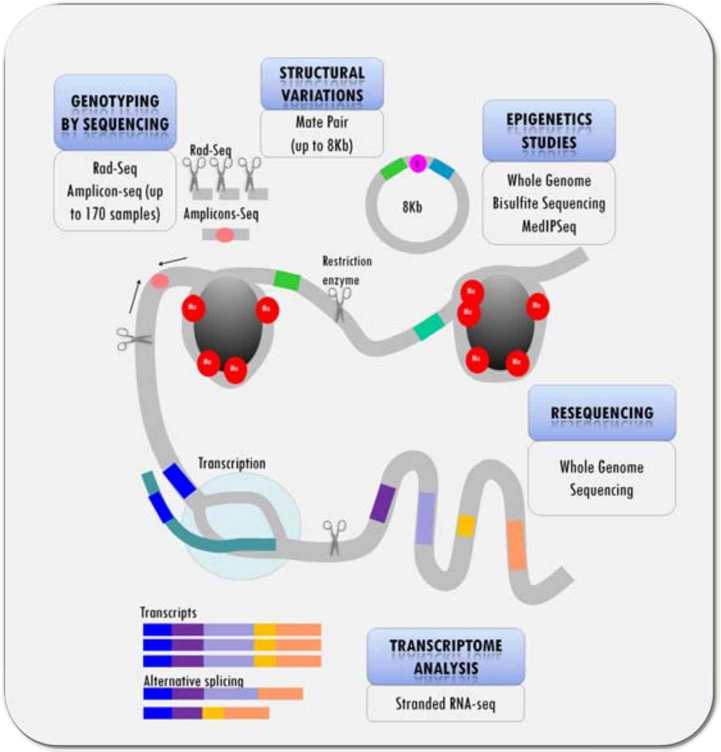


Affymetrix – Agilent iScan

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 library preparations
  - To increase the number of reads
  - To increase the length of reads

IGM (San Diegao, USA)	202 SMR cells							
		NUM	MAX	N50 BP	N50 NUM	MEAN	MEDIAN	BP/SMRTcell
	moyenne	98666	45457	12211	28413	9176	9032	0,906 M
	max	146374	52725	12981	41602	9997	9809	1,36Gb
Lausanne University (Swiss)	59 SMRT cells							
	moyenne	106800	46800	15172	28371	10773	9821	1,15Gb
	max	144358	53253	16132	38325	11436	10568	1,6GB
Get-PlaGe (France)	146 SMRT cells							
	moyenne	77086,6301	52317,4932	15365,4795	19705,5822	10326,6773	9152,5137	800Mb
	max	126777	80974	20507	33133	13635	12295	1,3Gb

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

(B. Mayjonade)

From NGS short reads to long reads

- **Fifteen projects to develop new applications :**
  - Whole genome sequencing on different species
  - Targeted sequencing
  - Complex population & metagenomics
  - RNA sequencing
  - Epigenetic

Focus on Next Generation Sequencing

From Sanger to 3rd generation

Unique sequence

Next Genration Sequencing

Sanger

16 or 48 capillaries

NGS - short reads

Ion Torrent

Ion PGM
2 Gb
400 pb

Ion S5
15 Gb
200 pb

Illumina

2xHiSeq 3000
700 Gb
2 x 150 pb

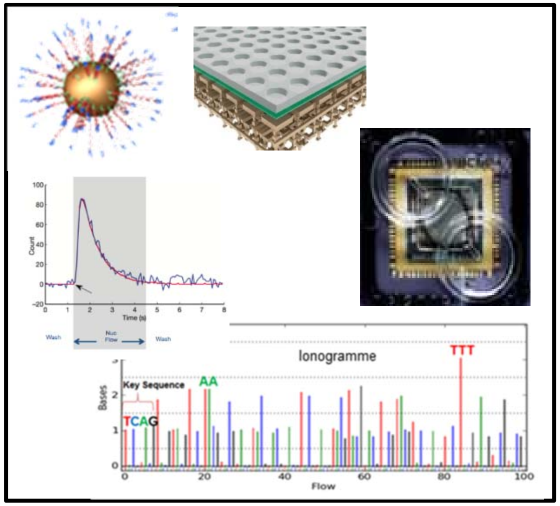
3 x MiSeq
15 Gb
2 x 300 pb

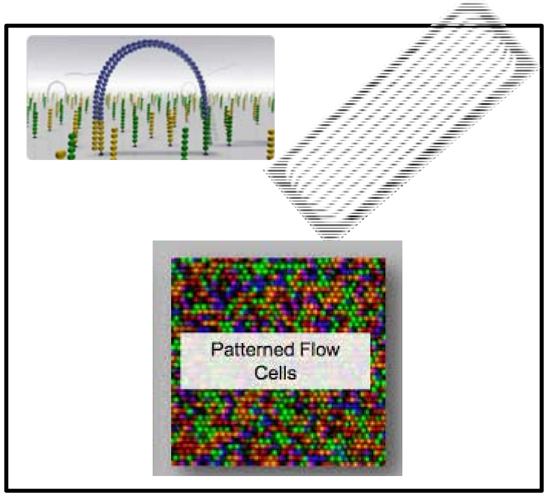
3G - long reads

PacBio RsII
70 000 reads - 20 kb

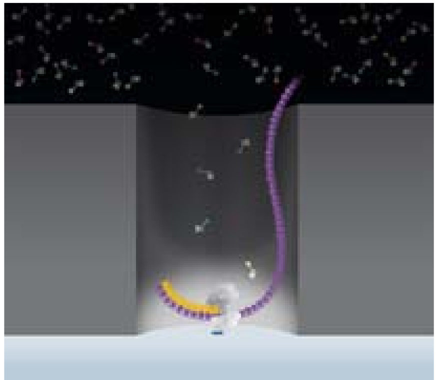
NGS : Ion Torrent et Illumina

Ion Torrent	Illumina
Scalable: different chips ↔ output: 50 Mb to 15 Gb	Output: 2 Gb to 90 Gb / lane
Reads : 1 x 200 pb or 1 x 400 pb	Reads: 1 x 50 pb to 2 x 300 pb
Bioinformatic tools pre-defined	Many developpements ongoing Data stored on the bioformatic core facility's clusters
New: S5 sequencer	New: 2nd HiSeq3000





3rd generation: PacBio RsII

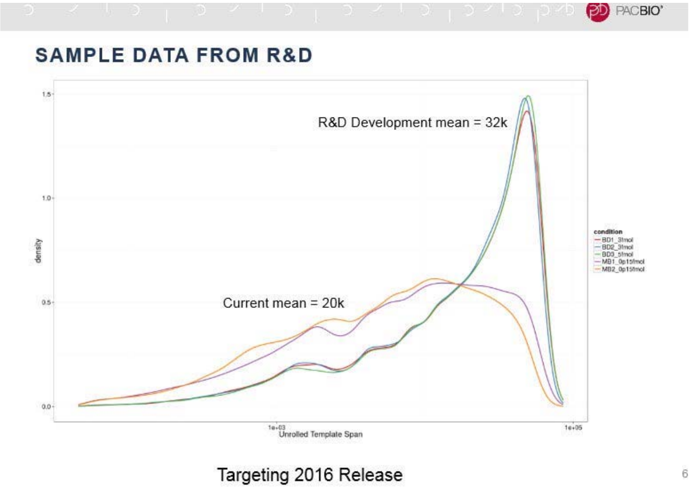


+ Single molecule real time sequencing
+ No PCR
+ Reads: 15-20 kb ⇨ 40 kb

- High error rate (>10%)

⇩

Random errors => corrected thanks to sequencing depth



SAMPLE DATA FROM R&D

Targeting 2016 Release

6

NGS: which applications ?

Genomics
Re-Sequencing

- Whole Genome
- Whole Exome
- Gene panels
- Hotspots

Transcritomics
Expression

- Whole transcriptome
- Full Lenght Transcipts
- Small RNA
- Targeted RNAseq

Epigenetics

- Whole Genome
- Targeted

Metagenomics

- 16S
- Amplicons

Genomics – Re-sequencing / *de novo* sequencing

➢ Whole Genome

Available:

- "PCR-free " librairies for Illumina
- 1 HiSeq3000 run = 8 genomes at 30X ≈ 2,5 days
- Ion S5 for small genomes
- GBS

⇨ mutations screening

Other applications :

- Sequening on PacBio RSII: 100 SMRT cells ⬄ 3 Gb genome at 30X

- Sequencing on PacBio RSII: 1 SMRT cell ⬄ 1 bacterial genome at 100X

⇨ *De novo* sequencing, screening for chromosome modifications, haplotype detection, SNP identification…

Illumina HiSeq 3000

PacBio RsII

Ion S5

Genomics – Re-sequencing

➢ Whole Exome (human)

⇨ Mutation/Biomarker identification

- Under development : AmpliSeq Exome Kit (Life Tech)
  + : easy and fast, cheaper than capture,
  Dedicated bioinformatic pipelines,
  CNV detection
  - : only CDS, less uniforme than capture kits,
  small to medium groups

- Possible developments: Nextera Rapid Capture exome
  + : "expanded" version with UTRs and miRNA,
  only 50 ng input
  - : 8 Gb in 2x75 pb needed,
  Dedicated to large groups

Ion S5

Illumina HiSeq

Genomics – Re-sequencing

➢ Targeted re-sequencing (panel, hotspots)

⇨ Mutation / biomarker screening on regions of interest

Already available:

- AmpliSeq (Life Tech) Custom Panel
- AmpliSeq (Life Tech) Ready-to-use Panel
  Ex : Comprehensive Cancer Panel
- TruSight (Illumina) kits
  Ex : TruSight Amplicon Cancer Panel
- TruSeq Custom Amplicon Kit

Ion PGM

Ion S5

Illumina MiSeq

Transcriptomics - Expression

- **Whole transcriptome sequencing**

⇨ Identification / quantification

Available:
Stranded RNAseq (HiSeq3000, 2 x 150 pb)

- **Full lenght transcripts sequencing**

Under dvelopment on PacBio RsII
⇨ Fusion transcripts, alternative transcripts

- **small RNA sequencing**

Available on Ion S5 sequencing
< 100 pb – stranded sequencing

Illumina HiSeq

PacBio RsII

Ion S5

Transcriptomics - Expression

➢ Targeted RNAseq ⇨ Expression

Design possible for custom panels:
On Ion Torrent and Illumina

➢ Targeted RNAseq - gene fusion

On Ion Torrent and Illumina :

- **RNA apoptose / RNA cancer panel**
- **TruSight RNA Pan-Cancer…**
- **custom**

Ion PGM

Ion S5

Illumina MiSeq

Epigenetics

➢ Bisulfite conversion

Available for whole genome on HiSeq
RRBS (Reduced Representation Bisulfite Sequencing)

➢ lncRNA/small RNA

RNAseq protocols modifications (ribodepletion vs poly-A selection)

➢ ChIP-seq/MedIP-seq

Illumina compatible

➢ "direct« methylome

Available on PacBio RsII for small genomes

Illumina HiSeq

Ion S5

PacBio RsII

Metagenomics - Metabarcoding

➢ 16S – commercial kit

- 7 variable regions on 9
- Analysis on Ion Reporter
- Low throughput

➢ 16S full lenght

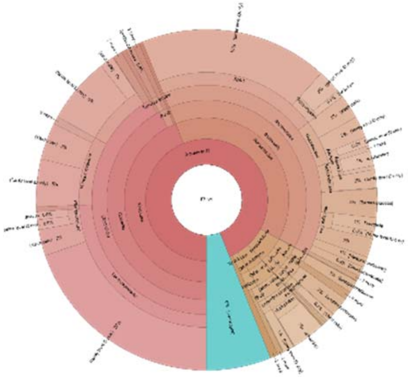
- Available PacBio RsII

➢ Custom Amplicons

- Available on Illumina – fully automated
- > 300 barcodes available
- 1 region 300-400 pb

➢ Whole metagenome / transcriptome

- Available on HiSeq



Ion PGM

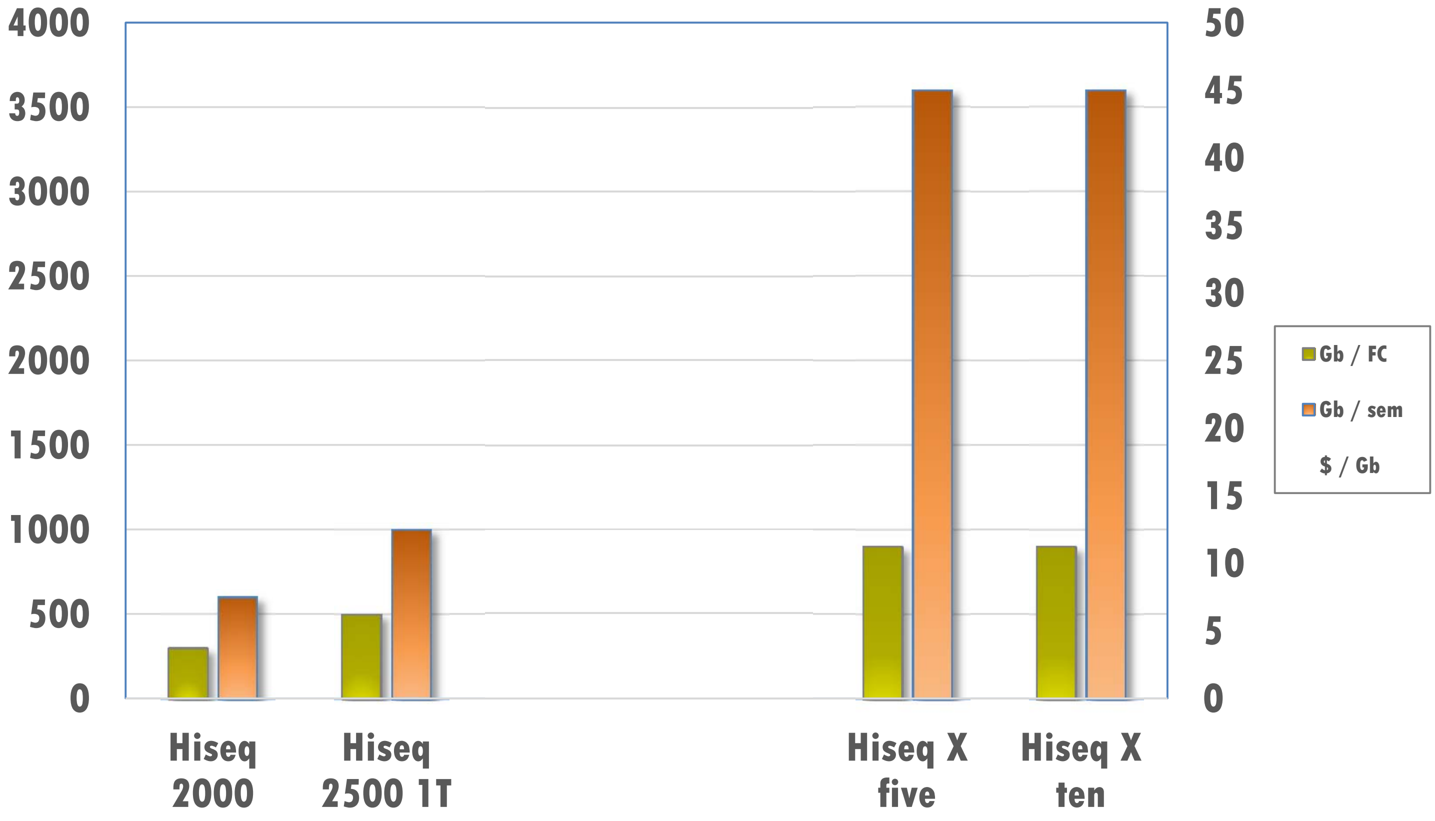
PacBio RsII

Ion S5

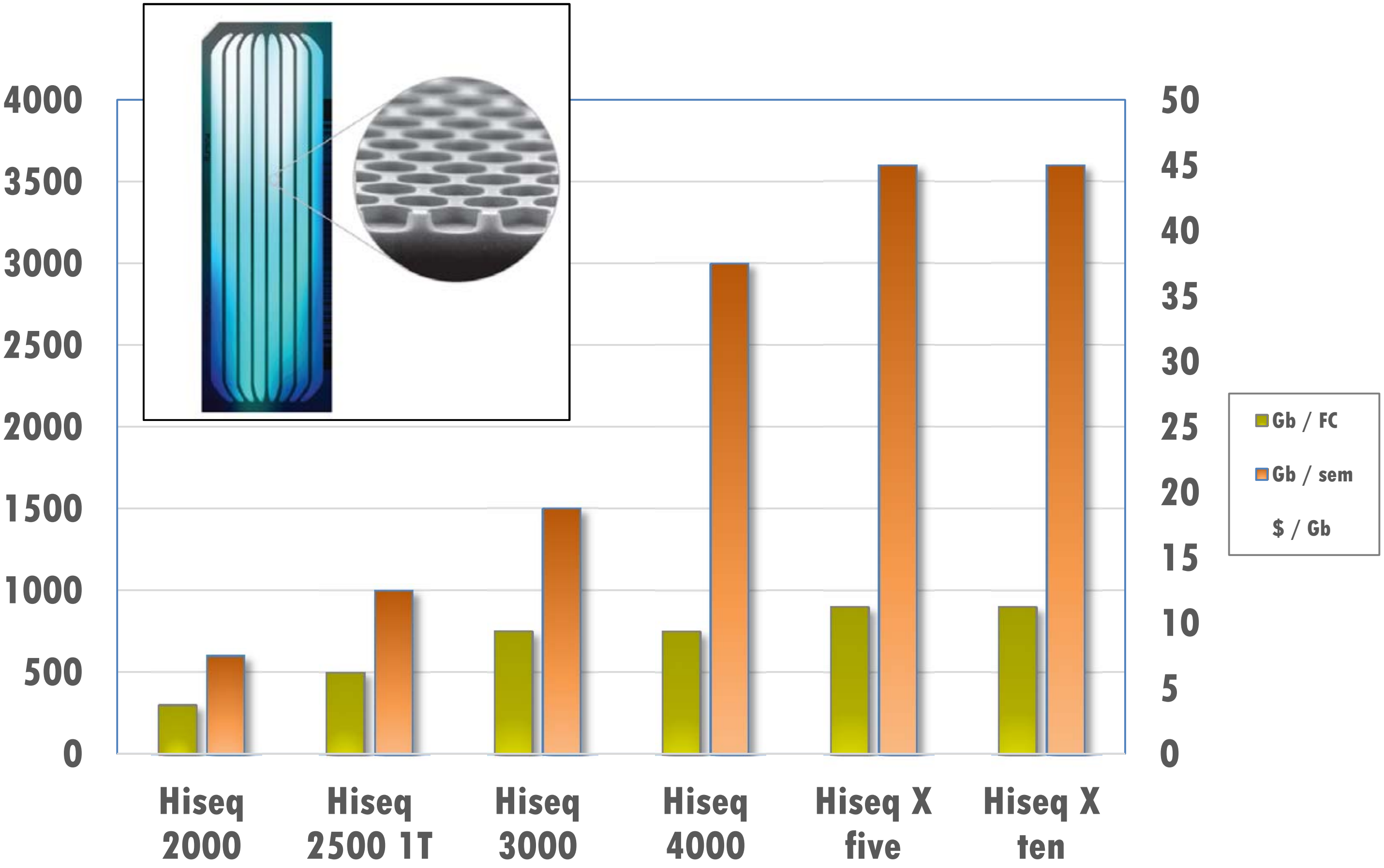
Illumina MiSeq

Illumina HiSeq

HiSeq productivity and costs

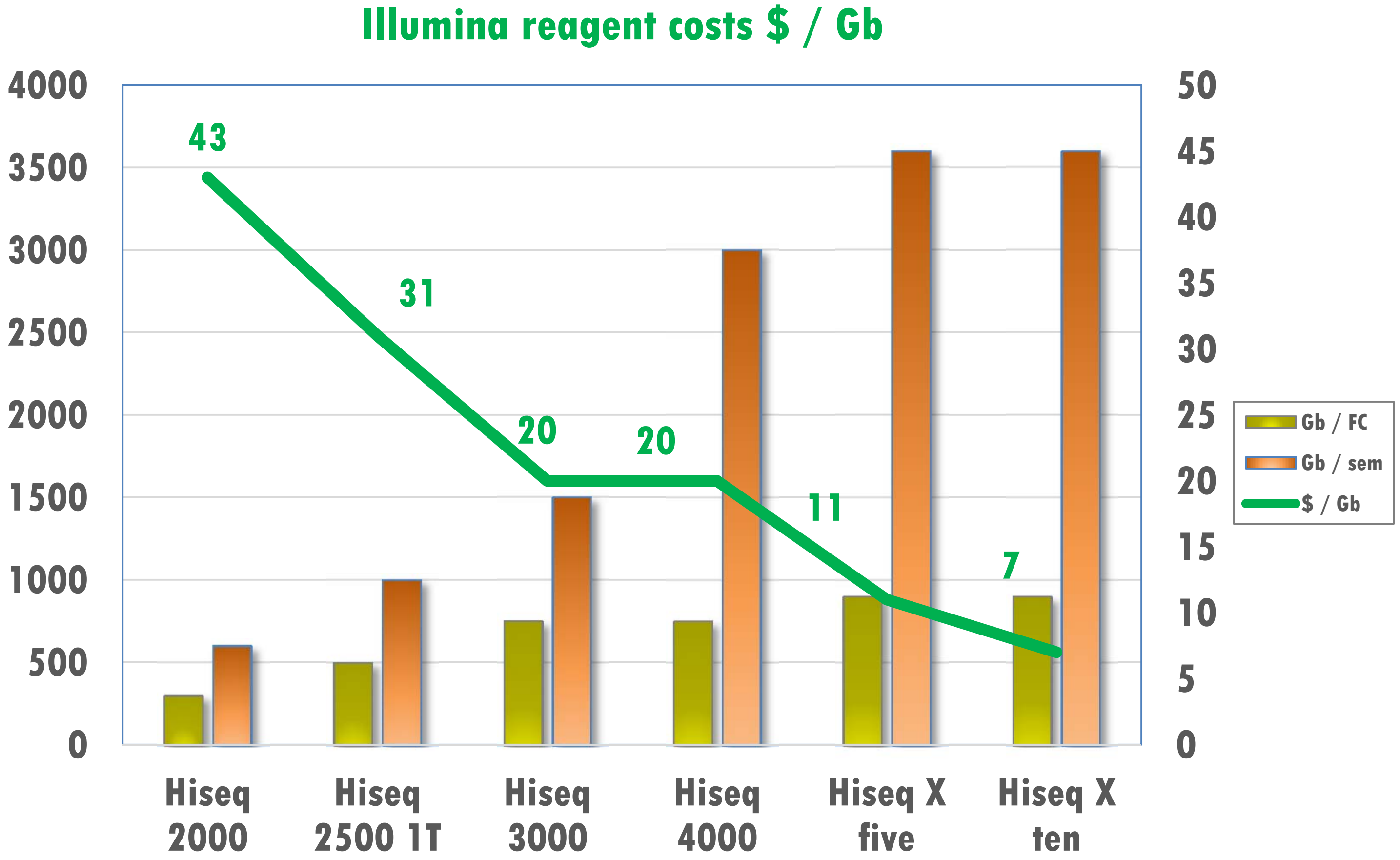


HiSeq productivity and costs



January 2015 : announcement of HiSeq3000 & HiSeq4000
(same Flow Cells as Hiseq X)

HiSeq productivity and costs



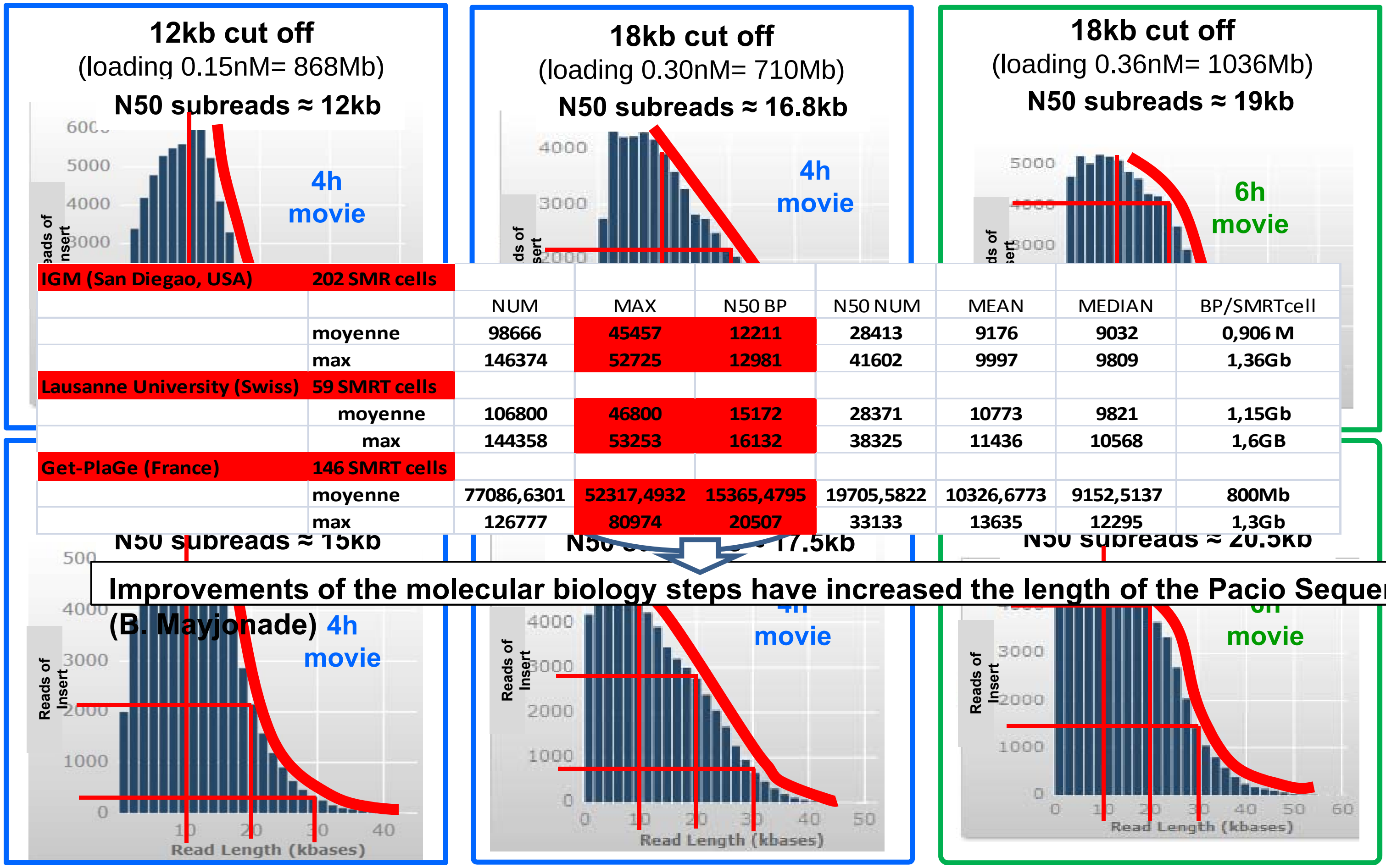
October 2015 : HiSeqX open to non humain (30x coverage) but 5 HiSeqX at least…

genotoul

GeT
Génome et
Transcriptome

Example of PacBio sequencing projects

Read lenght increase



12kb cut off
(loading 0.15nM= 868Mb)
N50 subreads ≈ 12kb

18kb cut off
(loading 0.30nM= 710Mb)
N50 subreads ≈ 16.8kb

18kb cut off
(loading 0.36nM= 1036Mb)
N50 subreads ≈ 19kb

N50 subreads ≈ 15kb

N50 … 17.5kb

N50 subreads ≈ 20.5kb

IGM (San Diegao, USA)	202 SMR cells							
		NUM	MAX	N50 BP	N50 NUM	MEAN	MEDIAN	BP/SMRTcell
	moyenne	98666	45457	12211	28413	9176	9032	0,906 M
	max	146374	52725	12981	41602	9997	9809	1,36Gb
Lausanne University (Swiss)	59 SMRT cells							
	moyenne	106800	46800	15172	28371	10773	9821	1,15Gb
	max	144358	53253	16132	38325	11436	10568	1,6GB
Get-PlaGe (France)	146 SMRT cells							
	moyenne	77086,6301	52317,4932	15365,4795	19705,5822	10326,6773	9152,5137	800Mb
	max	126777	80974	20507	33133	13635	12295	1,3Gb

Improvements of the molecular biology steps have increased the length of the Pacio Sequences (B. Mayjonade)

Read lenght increase

147 SMRT on 1st sunflower genome:

- **N50 15365**
- **800 Mb / SMRT cell**

103 SMRT on 2nd sunflower genome:

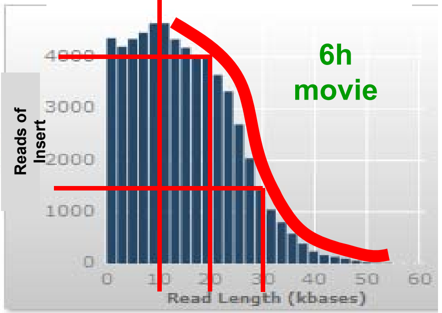
- **N50 18510**
- **1 041 Mb / SMRT cell (max 1 445 Mb)**

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


SMRT® Cells

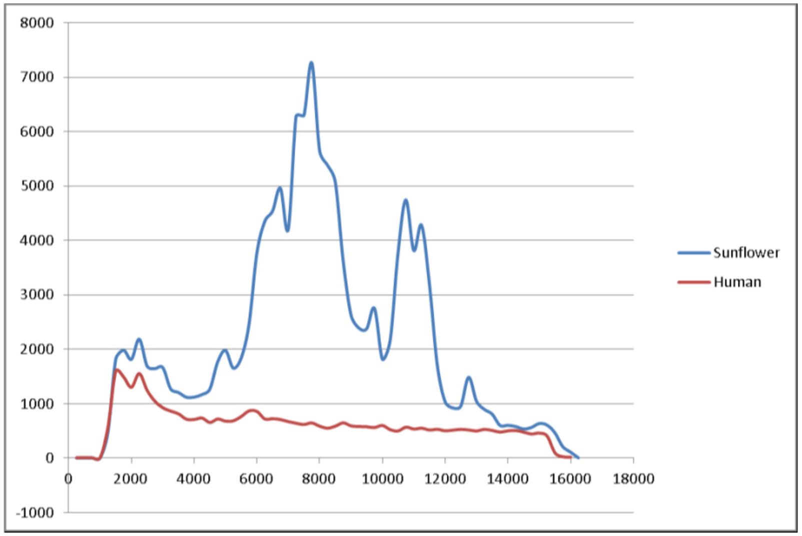
18kb cut off
(loading 0.36nM= 1036Mb)
N50 subreads ≈ 19kb

20kb cut off
(loading 0.45nM= 1000Mb)
N50 subreads ≈ 20.5kb



Sunflower genome sequencing

Sunflower : 30 % repeated sequences(LTR)
Homme : 8.8 % repeated sequences



The Challenge: obtention of sequences longer than the repeats

- **INRA (Sunflower team):**

  Hiseq, 127 X

  ➔ 43 % genome covered

- **International consortium:**

  454, Hiseq, Genetic & Physic maps (BAC)

  ➔ 63 % genome covered

- **INRA (Sunflower team & GeT-PlaGe) :**

  PacBio, 107 X (407 SMRT)

  ➔ 84 % genome covered

#ctg	MAX	N50 BP	# > N50	MEDIAN	Gb
13 124	4.4M	498 kb	1700	118 kb	3.03

Direct methylation analysis on PacBio

Goal: identification of different methylation patterns on evolved strains of the bacteria Ralstonia solanacearum form PacBio sequencing

2 strains: ancestral strain & evolved strain
3 SMRTcells/strain

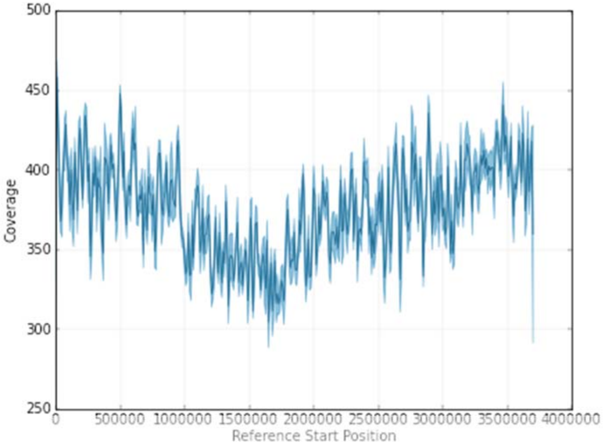
Direct methylation identified from PacBio sequencing:
- ✓ **6mA @50X**
- ✓ **4mC @50X**
- ✓ **5mC @500X**

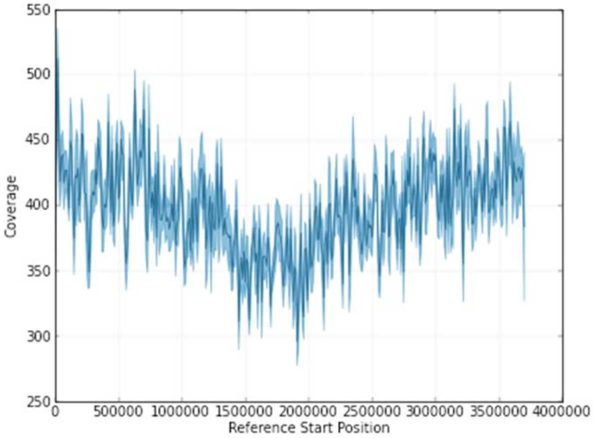
Coverage

Ancestral strain

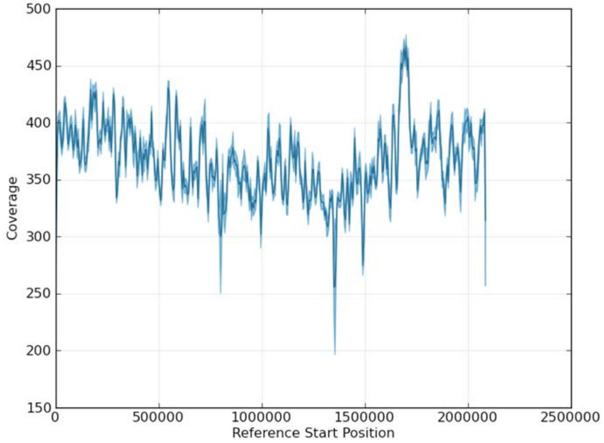
Evolved strain

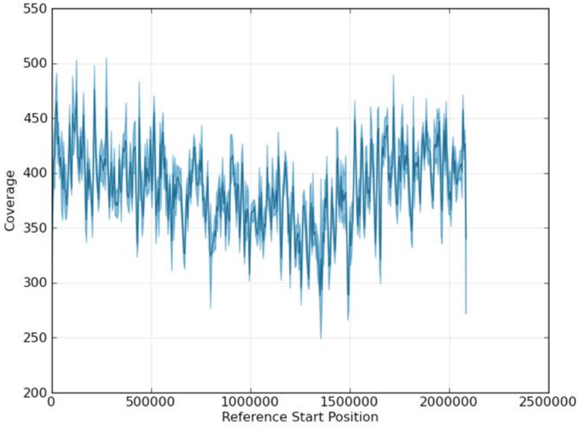
Chromosome





Plasmid



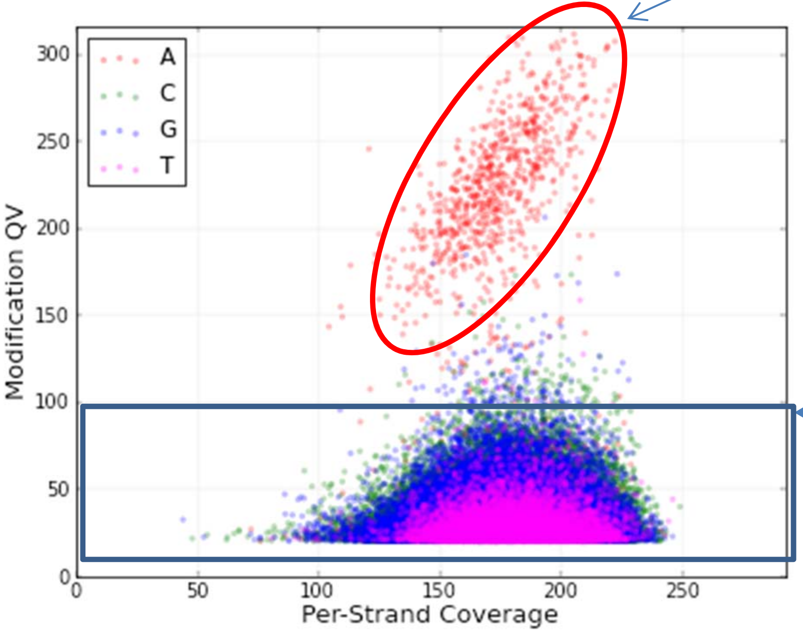


Modifications identified

Ancestral strain

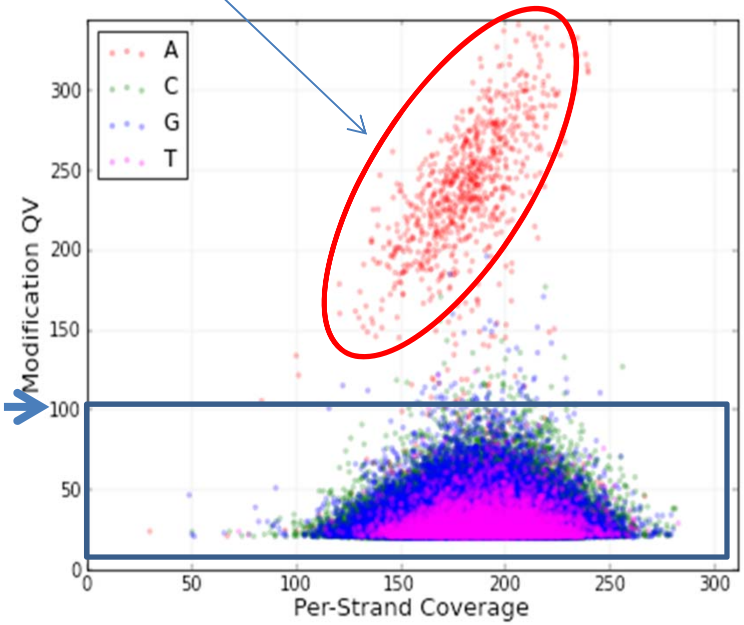
Evolved strain

6mA methylation signal



Threshold QV 100

Noise



Preliminary results

6mA pattern: GTWWAC
784 GTWWAC patterns in the genome
>98% of 6mA with a méthylation at a QV >100
    >98% 6mA -> typical pattern for bacteria

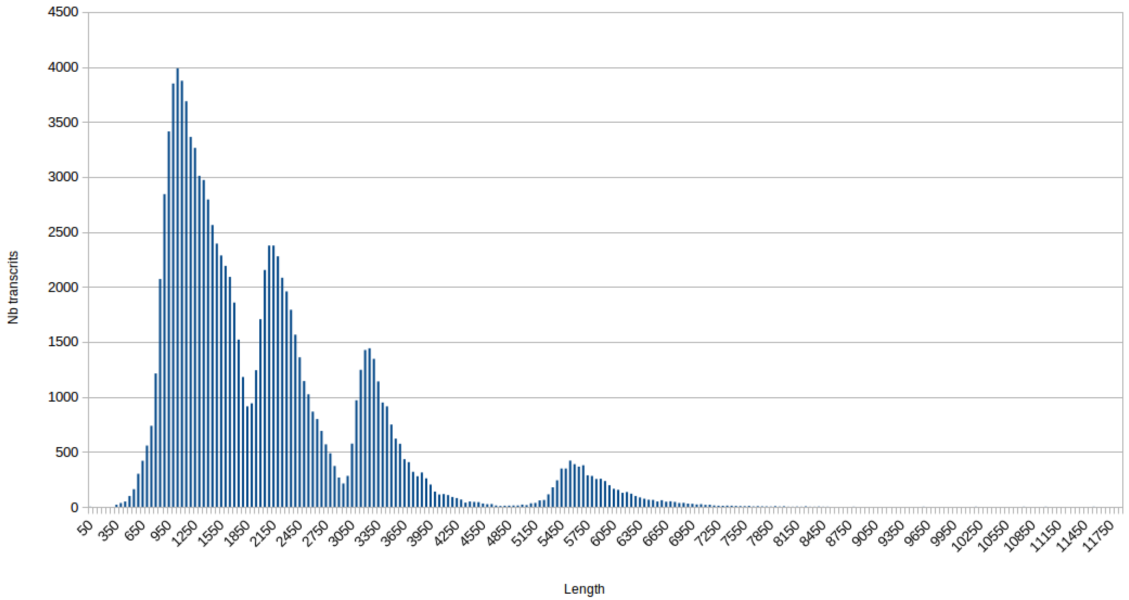
Preliminary analysis:

- ✓ GTWWAC pattern associated with 99,2% of the 6mA
- ✓ GCCGGC pattern associated with 33.2% of the 4mC
- ✓ 5mC difficile à voir en PacBio

~100 different patterns identified between ancestral and evolved strains => genes or intergenic regions, including many regulatory genes

Iso-seq

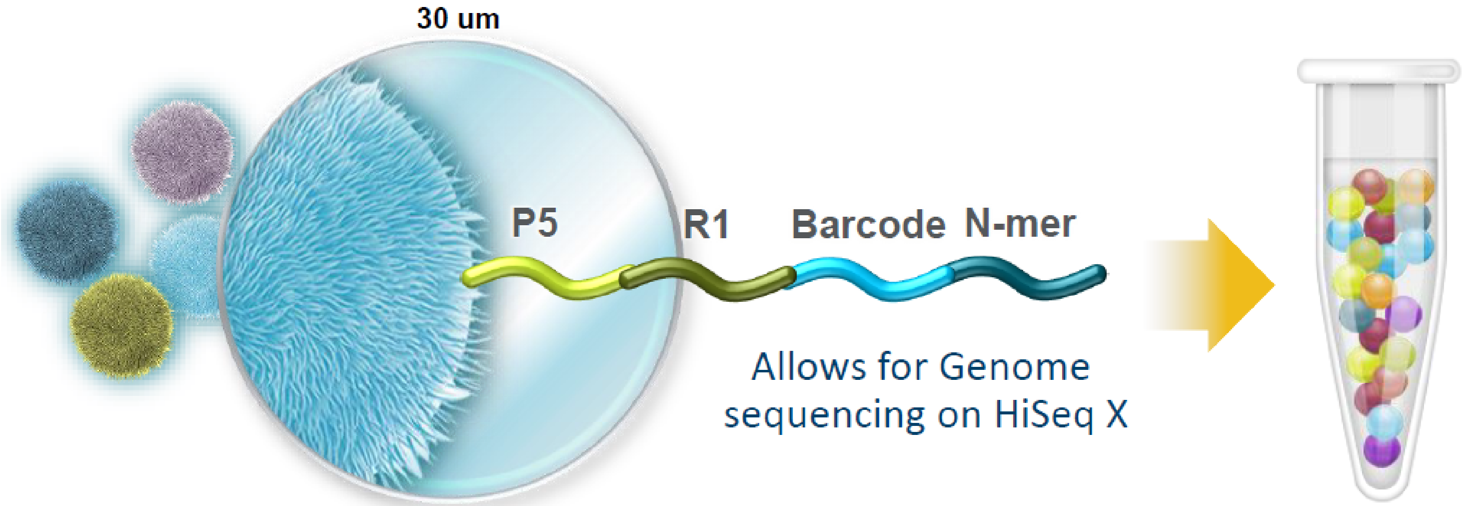
Iso-seq = full lenght transcripts sequencing on PacBio RSII



Preliminary results : ~110000 different transcripts identified

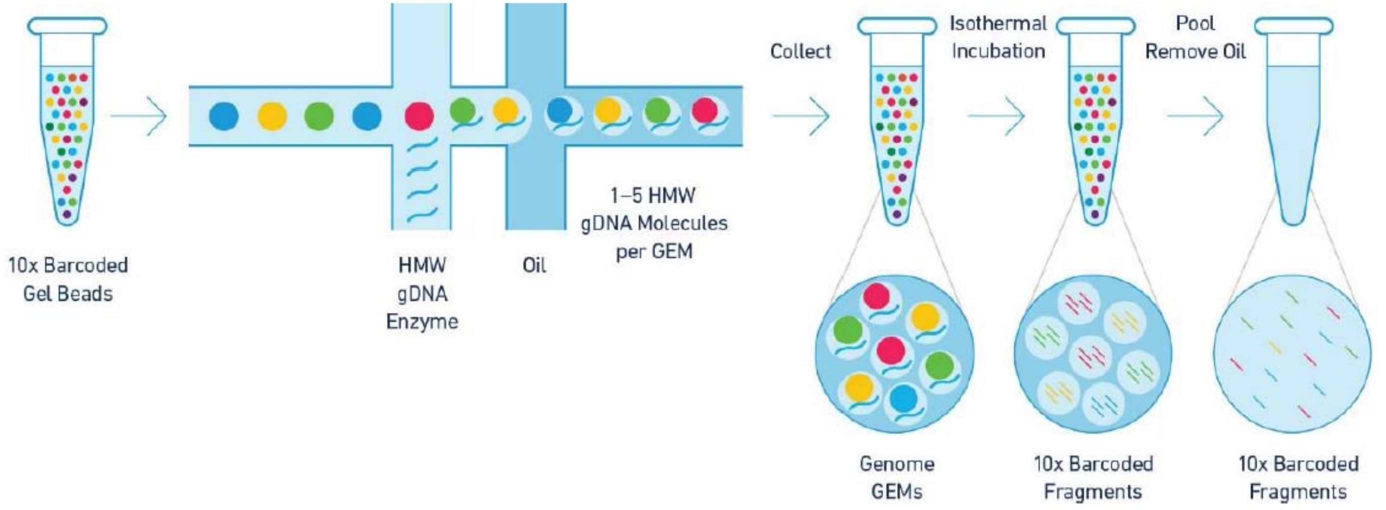
New techlonogy available

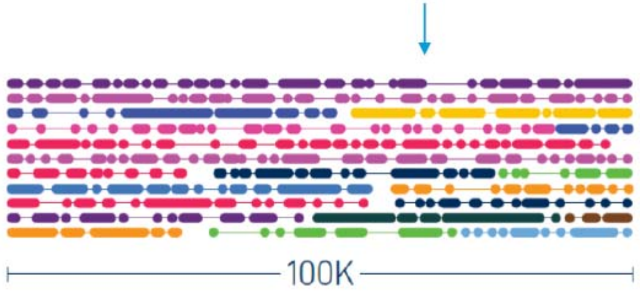
Chromium (10X genomics)
- Library prep for Illumina sequencing
- Long range genomic (~100 kb), haplotyping
- Single cell analysis
- Exome sequencing



Allows for Genome sequencing on HiSeq X

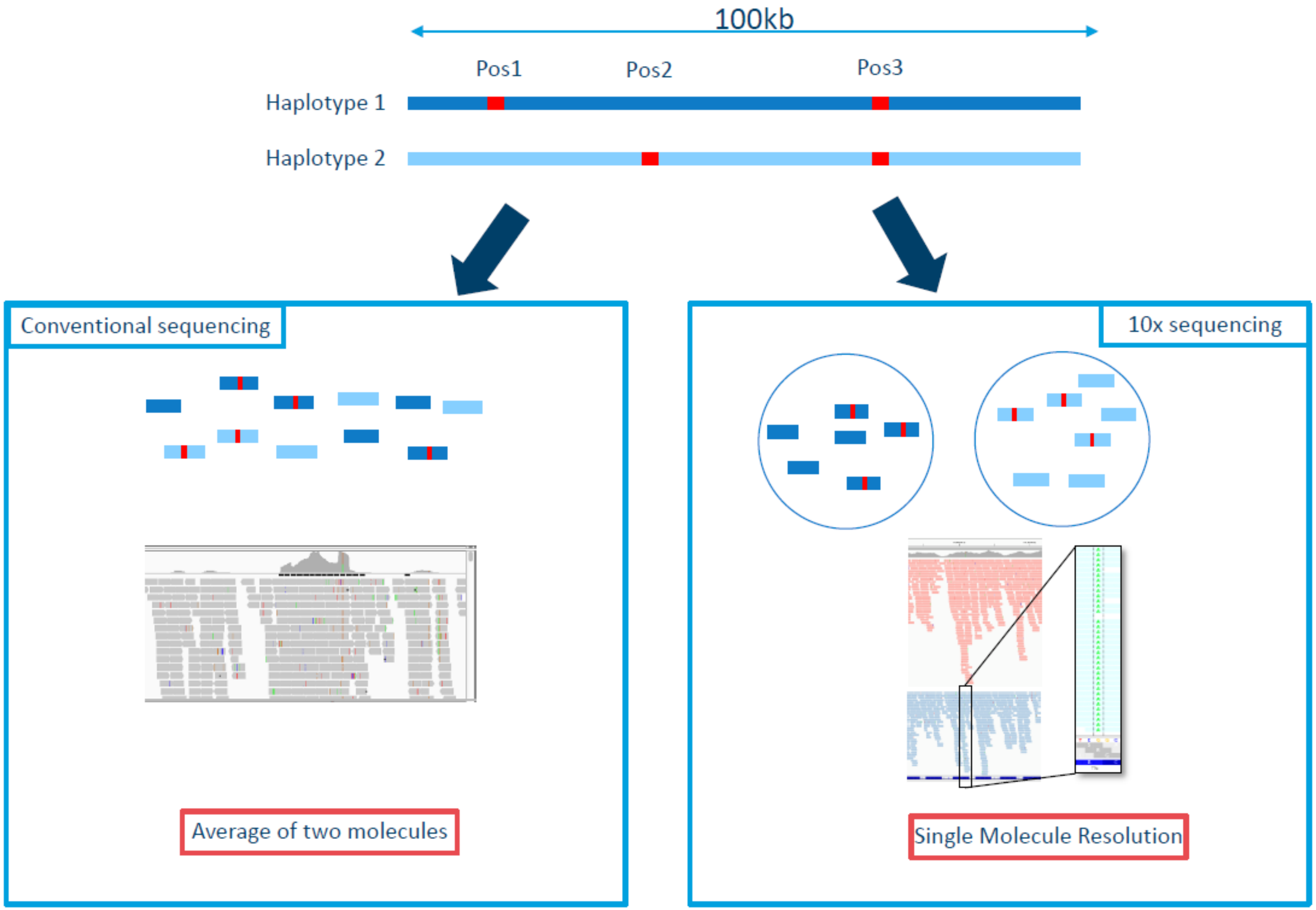
10XGENOMICS





Linked-Reads

10XGENOMICS



100kb

Pos1 Pos2 Pos3

Haplotype 1

Haplotype 2

Conventional sequencing

Average of two molecules

10x sequencing

Single Molecule Resolution

The future?

Sequel (PacBio)
- Same chemistry as the PacBio RSII (SMRT),
- Throughput x 7
- Similar error rate
- Price/Gb /4

PromethION (Oxford Nanopore)
- Nanopore technology
- "MinION multiple"

The future?

22 juin 2016 - Communiqué

Remise du rapport "France Médecine Génomique 2025" : faire entrer la France dans l'ère de la médecine génomique

Home » Business, Policy & Funding » Research Funding » France Plans to Invest €670M in Genomics, Personalized Medicine

France Plans to Invest €670M in Genomics, Personalized Medicine

Jun 23, 2016 | a GenomeWeb staff reporter

NEW YORK (GenomeWeb) – The French government announced that it plans to invest €670 million ($760.8 million) in a genomics and personalized medicine program meant to improve the diagnosis and prevention of disease in the country.

Illumina HiSeq X ten : 18000 genomes/year, 1000$ genome

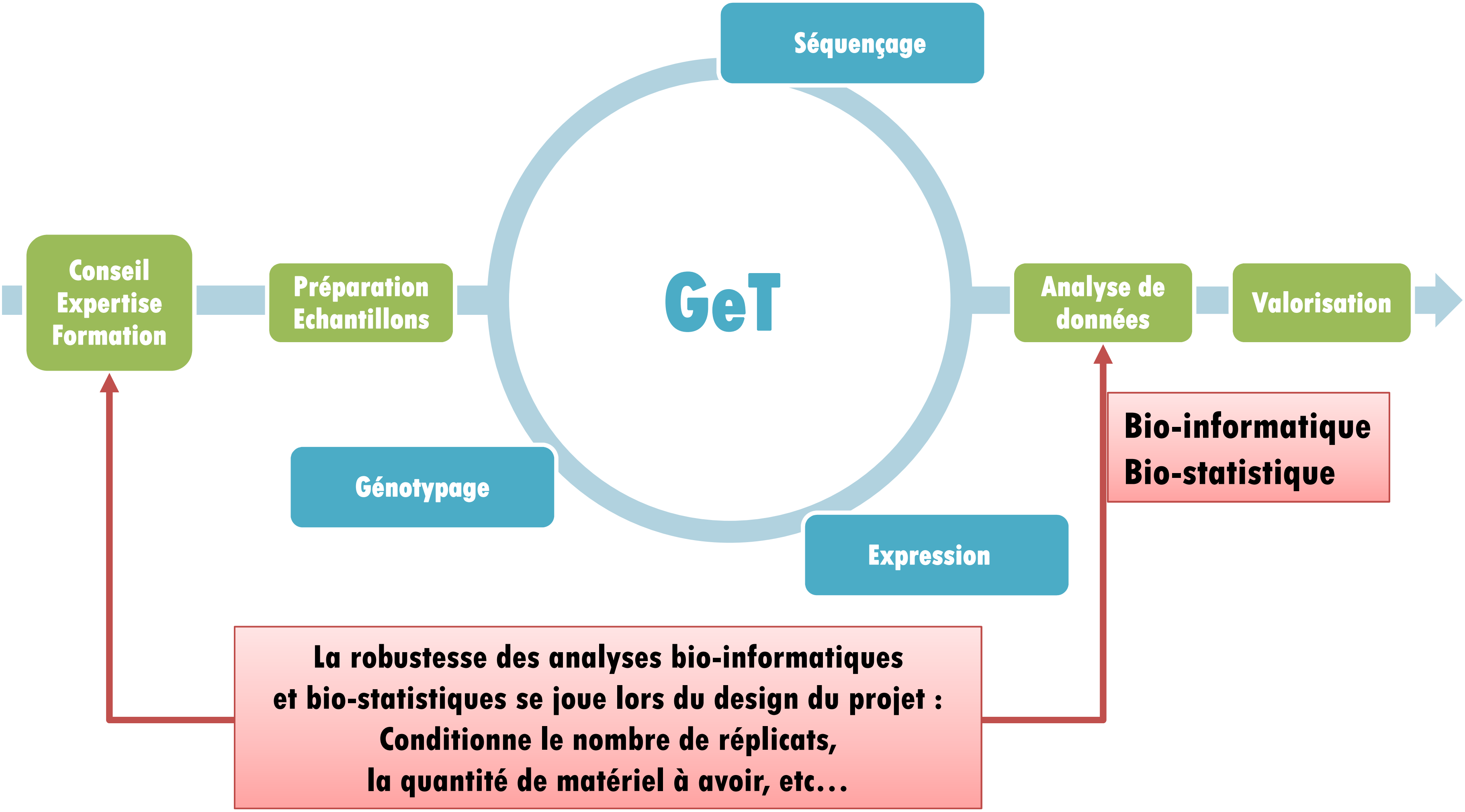
Target: 230000 genomes/year (2020)
12 sites in France

genotoul

GeT
Génome et
Transcriptome

Thanks!

Démarche construction projets



Conseil Expertise Formation

Préparation Echantillons

Séquençage

GeT

Génotypage

Expression

Analyse de données

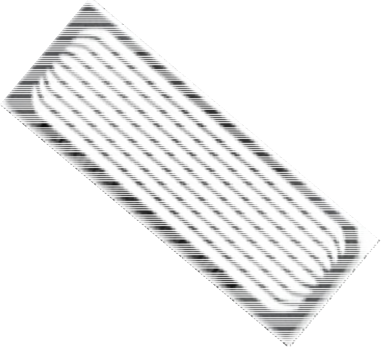
Valorisation

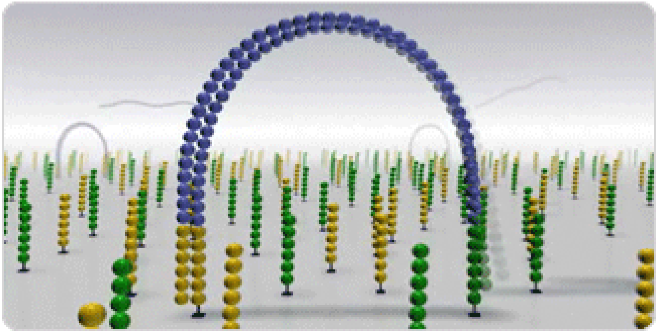
Bio-informatique
Bio-statistique

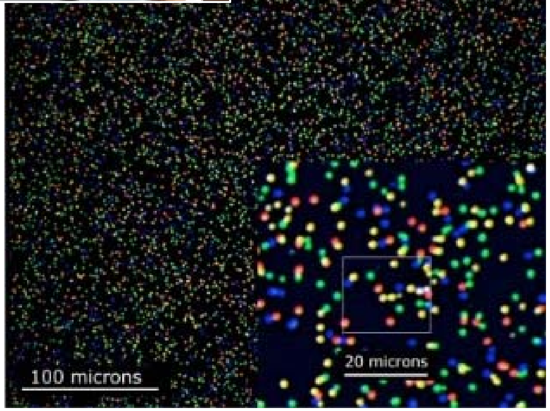
La robustesse des analyses bio-informatiques et bio-statistiques se joue lors du design du projet :
Conditionne le nombre de réplicats,
la quantité de matériel à avoir, etc…

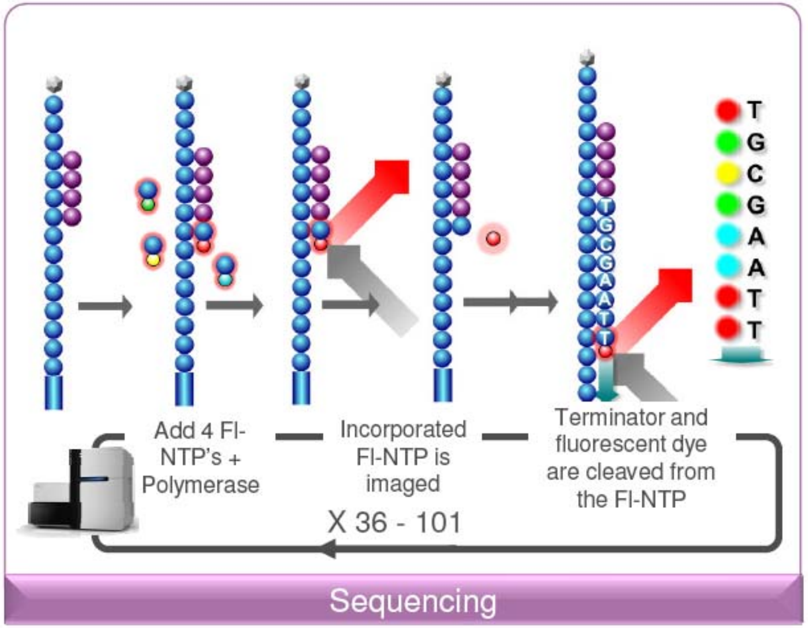
Chimie Illumina en détails

- **Séquençage par Synthèse**
- **Détection par fluorescence**



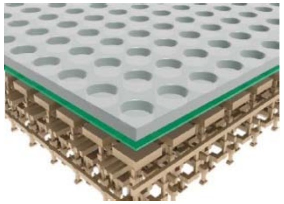


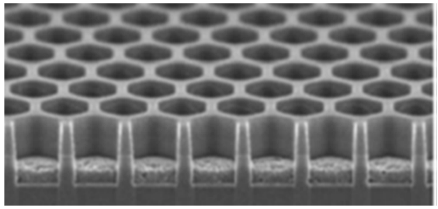


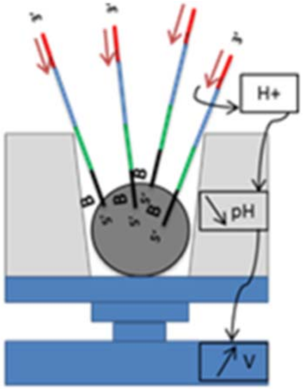


Chimie Ion Torrent en détails

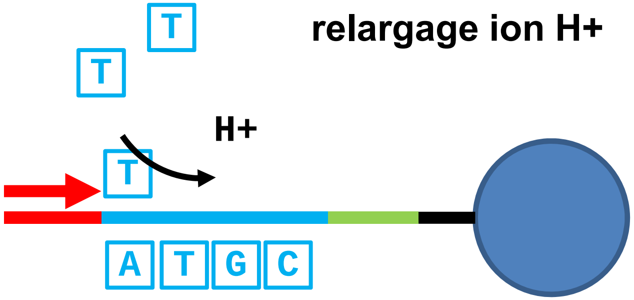
- ➢ **Séquençage par Synthèse**
- ➢ **Par semi-conduction**



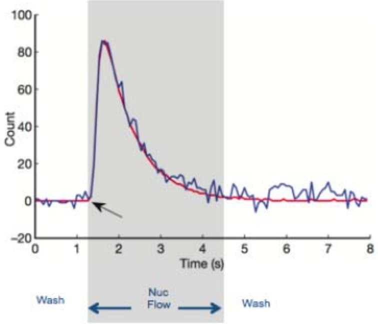


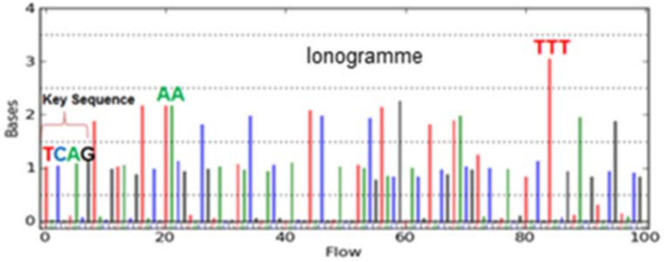


Flow de dNTP
relargage ion H+



différence de potentiel
détectée par la machine

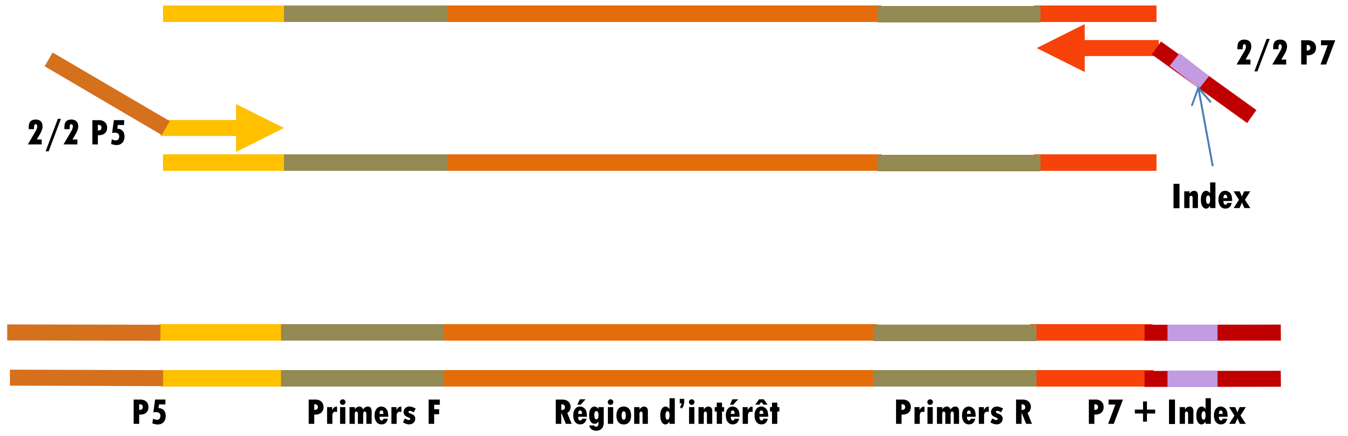




Séquençage amplicons sur Illumina

PCR1 (classique, 30 à 35 cycles) : amplification région cible

PCR2 (réduite, 10 à 12 cycles) : ajout adaptateurs + index



1 Librairie = 2 PCRs, Réalisable en plaque 96 puits, Automatisée

Analyses qualitatives des données Illumina

Images

↓ Base calling HCS/RTA

Fichiers Bcl

Séquences/qualité

CASAVA ↓

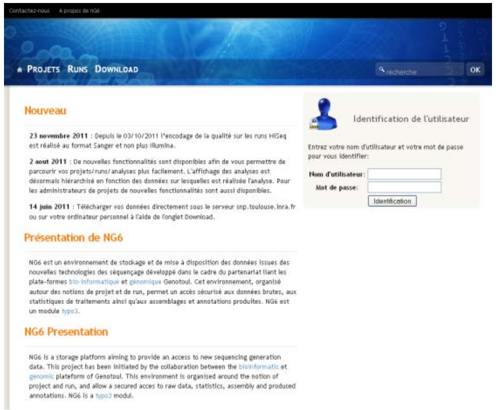
Démultiplexage

↓ FastQ Illumina filter

Séquences/qualité

↓ FastQC

Affichage NG6



Nouveau

23 novembre 2011 : Depuis le 03/10/2011 l'encodage de la qualité sur les runs HiSeq est réalisé au format Sanger et non plus Illumina.

2 aout 2011 : De nouvelles fonctionnalités sont disponibles afin de vous permettre de parcourir vos projets/runs/analyses plus facilement. L'affichage des analyses est désormais hiérarchisé en fonction des données sur lesquelles est réalisée l'analyse. Pour les administrateurs de projets de nouvelles fonctionnalités sont aussi disponibles.

14 juin 2011 : Télécharger vos données directement sous le serveur snp.toulouse.inra.fr ou sur votre ordinateur personnel à l'aide de l'onglet Download.

Présentation de NG6

NG6 est un environnement de stockage et de mise à disposition des données issues des nouvelles technologies des séquençage développé dans le cadre du partenariat liant les plate-formes bio-informatique et génomique Genotoul. Cet environnement, organisé autour des notions de projet et de run, permet un accès sécurisé aux données brutes, aux statistiques de traitements ainsi qu'aux assemblages et annotations produites. NG6 est un module typo3.

NG6 Presentation

NG6 is a storage platform aiming to provide an access to new sequencing generation data. This project has been initiated by the collaboration between the bioinformatic et genomic plateform of Genotoul. This environment is organised around the notion of project and run, and allow a secured acces to raw data, statistics, assembly and produced annotations. NG6 is a typo3 modul.

Identification de l'utilisateur

Entrez votre nom d'utilisateur et votre mot de passe pour vous identifier:

Nom d'utilisateur:

Mot de passe:

Identification

Génomique - Reséquençage

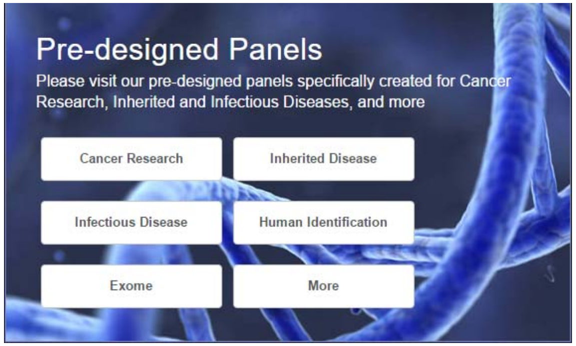
➢ Séquençage ciblé (panel, hotspots)

⇨ Recherche / typage mutations sur régions d'intérêt
⇨ Possibilité de recherche en profondeur

Peut être réalisé :

- **Sur Ion Torrent : tout type de panels AmpliSeq**
  - Custom - humain et souris - 200 ou 400 pb
  - Communautaires - 200 pb
  - Ready-to-use : panels dédiés cancérologie - 200 pb
    - Cancer Hotspot Panel v2 *FFPE
    - Comprehensive Cancer Panel *FFPE

⇨ www.ampliseq.com



Pre-designed Panels

Please visit our pre-designed panels specifically created for Cancer Research, Inherited and Infectious Diseases, and more

Technologies compatibles

Ion Torrent

Génomique - Reséquençage

➢ Séquençage ciblé (panel, hotspots)

⇨ Recherche / typage mutations sur régions d'intérêt
⇨ Possibilité de recherche en profondeur

Des développements peuvent être envisagés:

- **Sur Illumina** 2x75 ou 2x150pb
  - TruSight Tumor 15
  - TruSight Myeloid
  - TruSight Amplicon Cancer Panel

  - TruSeq Custom Amplicon Kit
    existe une version low input - FFPE
    ⇨ DesignStudio Illumina

Technologies compatibles

Illumina MiSeq

Illumina MiniSeq ?