genotoul

GeT
Génome et
Transcriptome

Les techniques de séquençage
L'analyse qualité

http://get.genotoul.fr
http://bioinfo.genotoul.fr

IBiSA
INFRASTRUCTURES BIOLOGIE SANTÉ ET AGRONOMIE

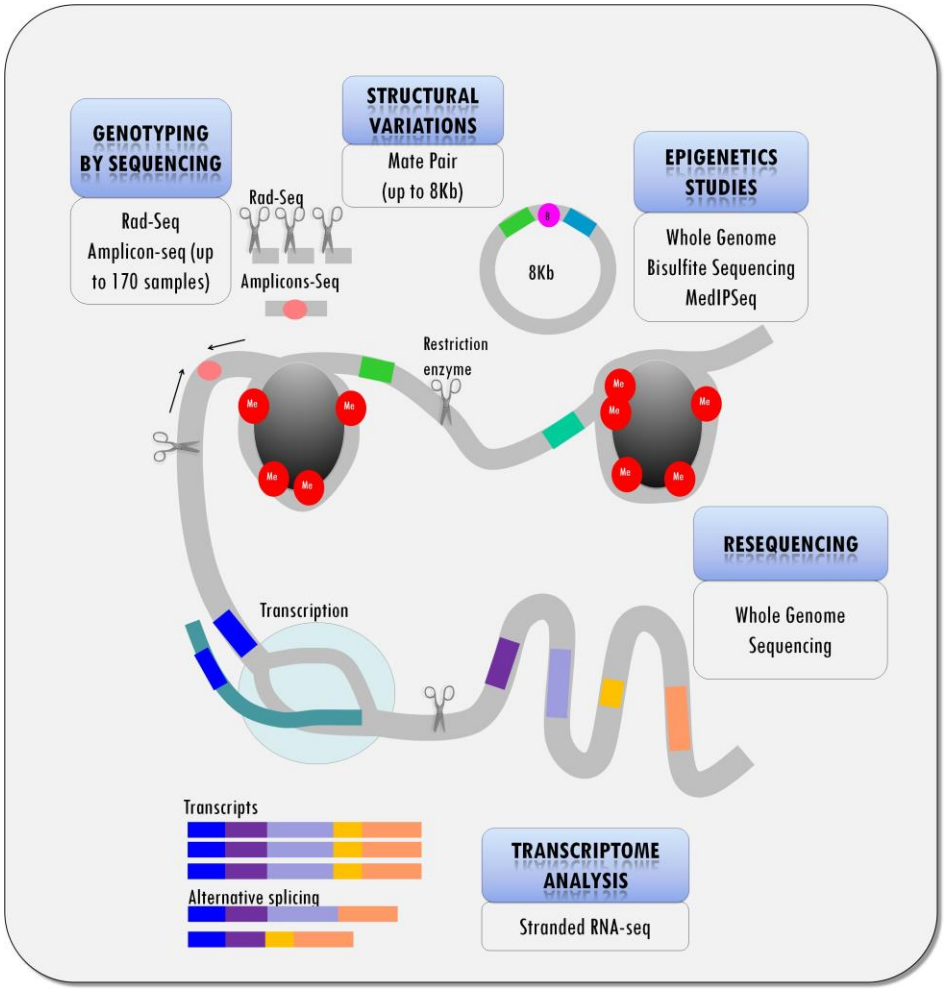
LLOYD'S REGISTER QUALITY ASSURANCE · ISO9001

- ✓ **Séquençage : déterminer la succession linéaire des bases A, C, G, T de l'ADN, la lecture de cette séquence permet d'étudier l'information biologique contenue par celle-ci**

- ✓ **Séquençage Nouvelle Génération (NGS, Next Generation Sequencing) : Séquençage à très haut débit, génération d'un très grand nombre de séquences simultanément**

A quoi sert le séquençage?

- ✓ RNA-seq : transcriptome sequencing. Informations sur les ARN via le séquençage de l'ADN complémentaire (cDNA)
- ✓ Re-séquencage : séquençage d'un fragment d'ADN et comparaison du résultat obtenu avec une séquence de référence connue - détection de SNP, haplotyping
- ✓ Séquencage *de novo* : séquençage d'un génome pour lequel il n'existe pas de séquence de référence, détermination d'une séquence inconnue
- ✓ Epigénétique : étude des changements d'activité des gènes qui sont transmis au fil des divisions cellulaires ou des générations sans faire appel à des mutations de l'ADN.
- ✓ Génotypage par séquençage : découverte et génotypage de plusieurs milliers de polymorphismes de type SNP/INDEL chez de nombreux individus
- ✓ Metagénomique : analyse du matériel génétique provenant directement des échantillons biologiques – étude de la diversité génétique d'un échantillon



GeT Génome et Transcriptome

From Sanger to 3rd generation

Unique sequence

Sanger

16 or 48 capillaries

Next Genration Sequencing

NGS - short reads

Illumina

2xHiSeq 3000
700 Gb
2 x 150 pb

3 x MiSeq
15 Gb
2 x 300 pb

Long Reads Sequencing

Chromium (10X genomics)

Oxford Nanopore
MinION

3G - long reads

PacBio RsII
70 000 reads - 20 kb

MiniSeq

MAX OUTPUT
8 Gb
MAX READ NUMBER
25 million
MAX READ LENGTH
2×150 bp

MiSeq

MAX OUTPUT
15 Gb
MAX READ NUMBER
25 million
MAX READ LENGTH
2×300 bp

NextSeq

MAX OUTPUT
120 Gb
MAX READ NUMBER
400 million
MAX READ LENGTH
2×150 bp

HiSeq 4000

MAX OUTPUT
1500 Gb
MAX READ NUMBER
5 billion
MAX READ LENGTH
2×150 bp

HiSeq X Ten

MAX OUTPUT
1800 Gb
MAX READ NUMBER
6 billion
MAX READ LENGTH
2×150 bp

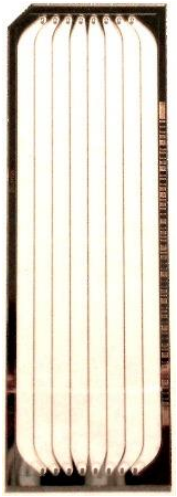
Présentation d'un séquenceur NGS Illumina HiSeq3000

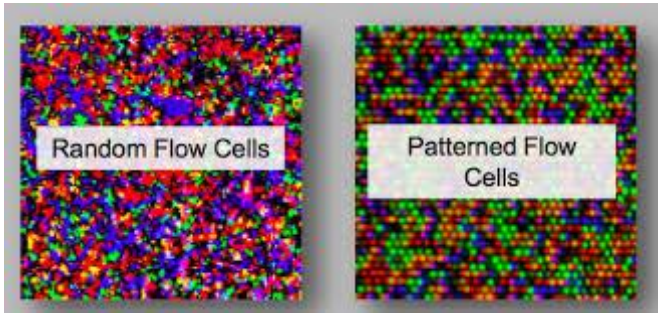
Spécifications :

- ✓ **700 Gb / 1 run (2,5 jours)**
- ✓ **Paired-end (2x150 bp) : ~600 millions de séquences/lane**
- ✓ **Séquencage par synthèse**

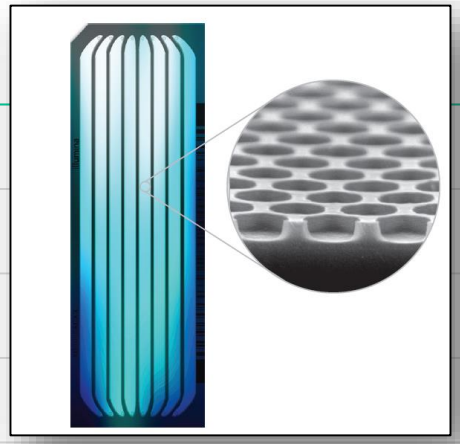
1 flowcell = une lame = 8 lignes

illumina





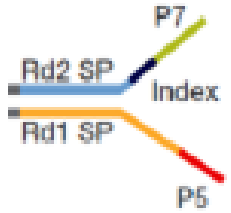
Productivité et coût



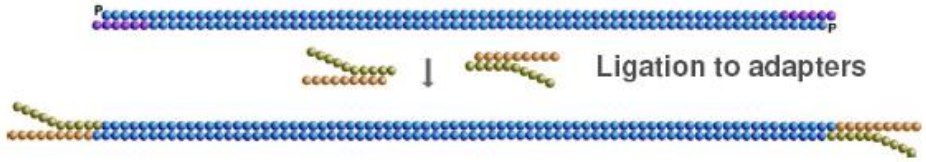
January 2015 : announcement HiSeq3000 & HiSeq4000
(same Flow Cells as Hiseq X)

Construction des librairies Illumina

- ✓ **Ligation des adaptateurs (contiennent l'index)**

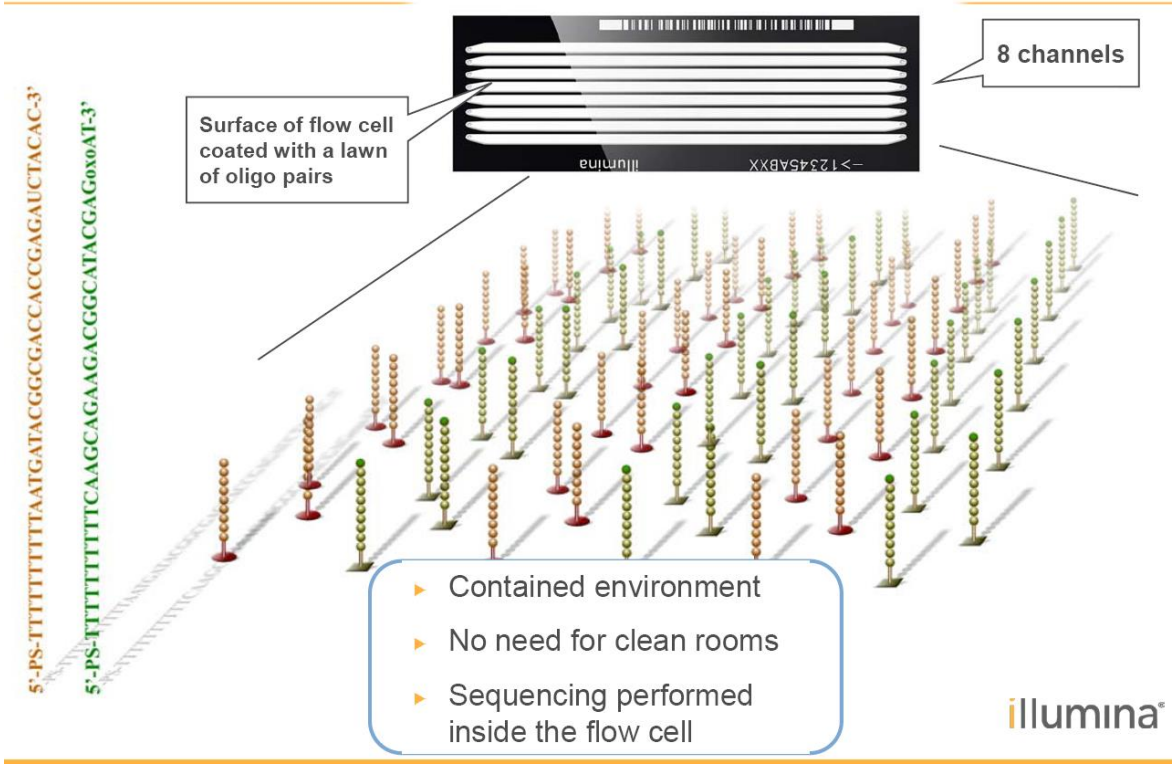


  - **P5 et P7: Fixation sur la flowcell**
  - **SP 1 et 2: primers de séquençage**
  - **Tag: index (6 bases): multiplexage par 24**



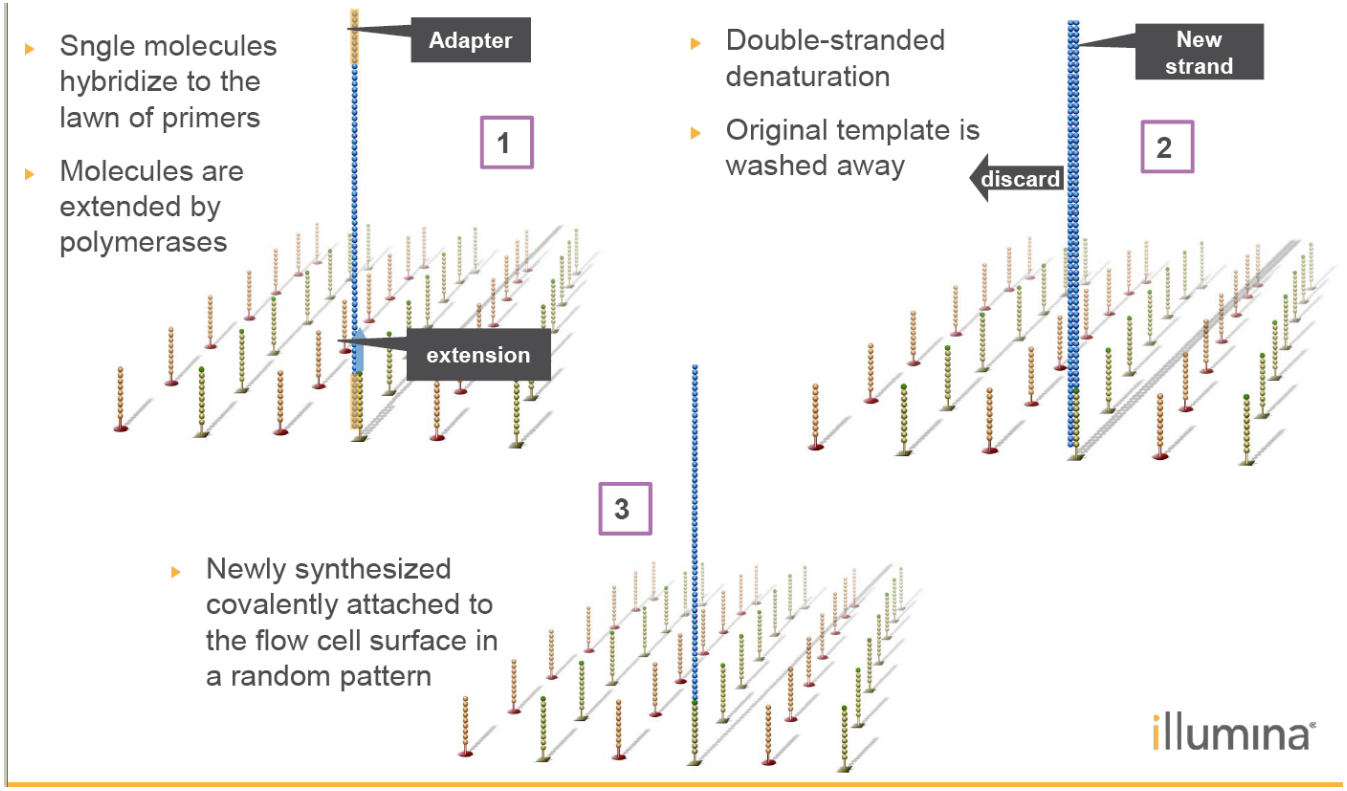
Génération des clusters

Design de la flowcell



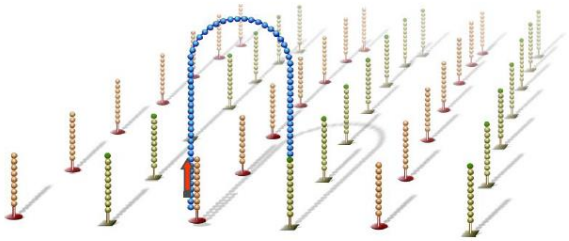
Génération des clusters

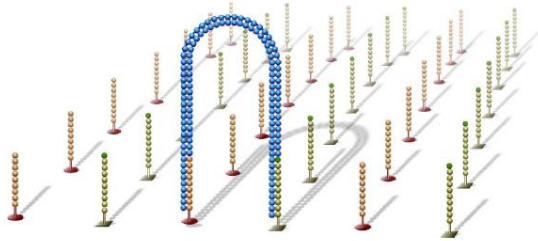
- ✓ **Hybridation des librairies grâce aux adaptateurs à l'intérieur de la flowcell**
- ✓ **Synthèse du brin complémentaire**
- ✓ **Dénaturation**



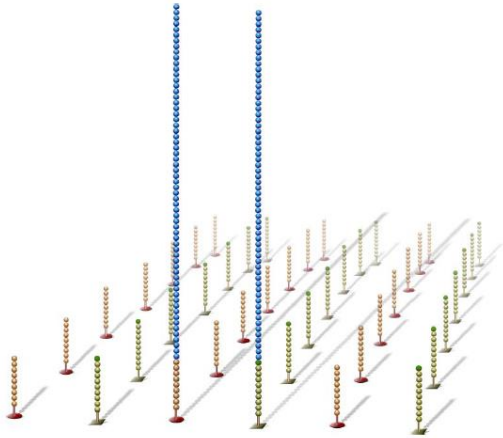
Génération des clusters

- ✓ **Formation d'un pont**
- ✓ **Synthèse du brin complémentaire**
- ✓ **Dénaturation**





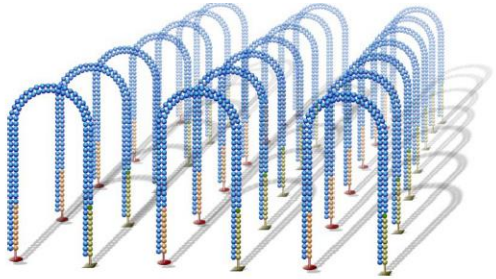
- Single-strand flips over to hybridize to adjacent primers to form a bridge
- Hybridized primer is extended by polymerases
- Bridge is denatured

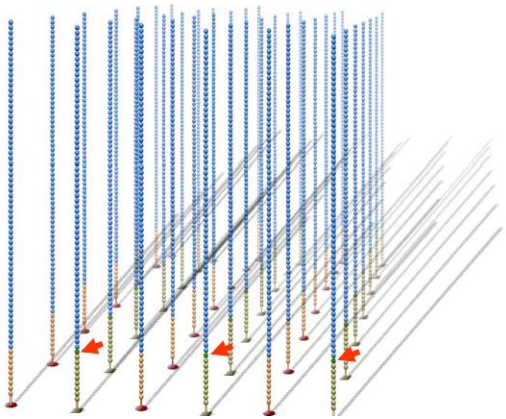


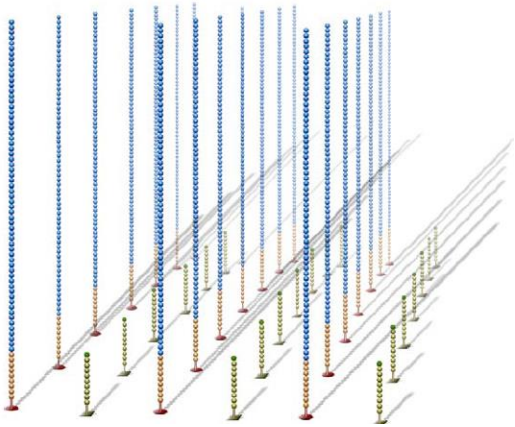
illumina®

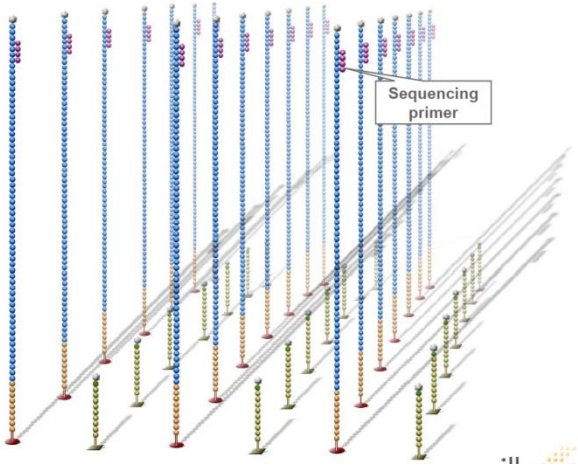
Génération des clusters

- Bridge amplification cycle repeated until multiple bridges are formed
- Bridges denaturation
- Reverse strands cleaved and washed away



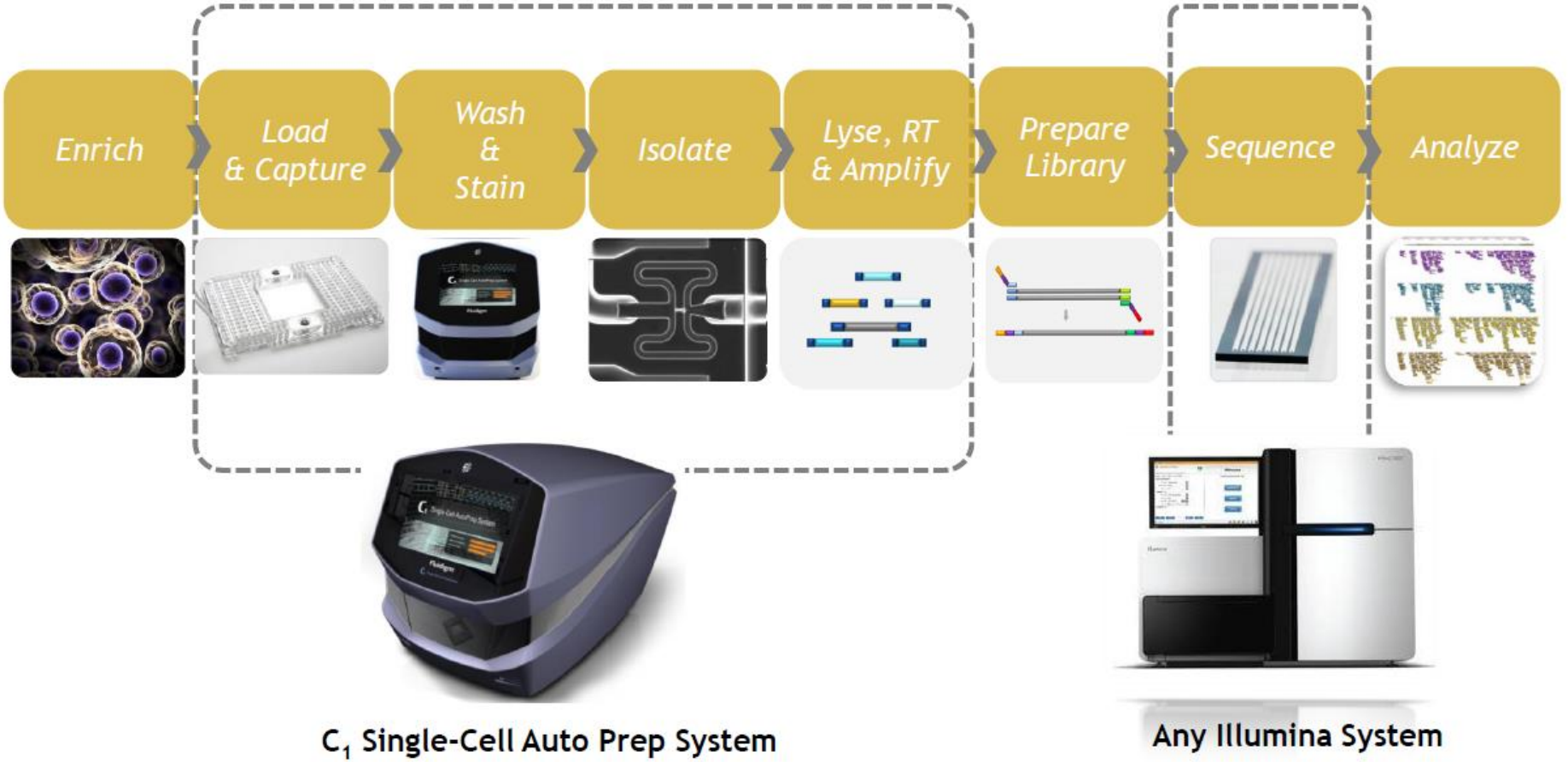






➢ **Formation de 750 000-850 000 clusters / mm²**

Fluidigm C1 Single Cell RNA seq



GeT Génome et Transcriptome

From Sanger to 3rd generation

Unique sequence

Sanger

16 or 48 capillaries

Next Genration Sequencing

NGS - short reads

Illumina

2xHiSeq 3000
700 Gb
2 x 150 pb

3 x MiSeq
15 Gb
2 x 300 pb

Long Reads Sequencing

Chromium (10X genomics)

Oxford Nanopore
MinION

3G - long reads

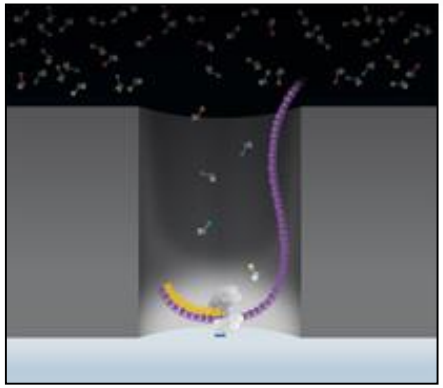
PacBio RsII
70 000 reads - 20 kb

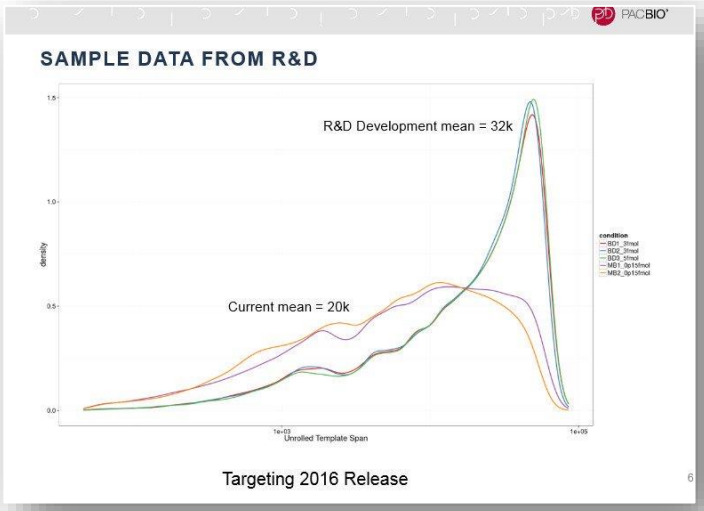
3ième génération de séquenceurs

+ Séquençage molécule unique
+ Pas d'étape de PCR
+ Lectures longues

- Beaucoup d'erreur pour l'instant (>10%)

Ce sont des erreurs aléatoires qui se corrigent bien avec une couverture importante





Read length increase

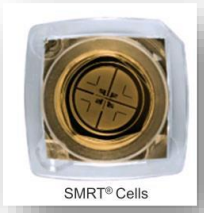
147 SMRT on 1st sunflower genome:

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

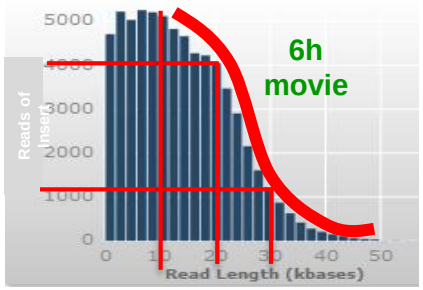
103 SMRT on 2nd sunflower genome:

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

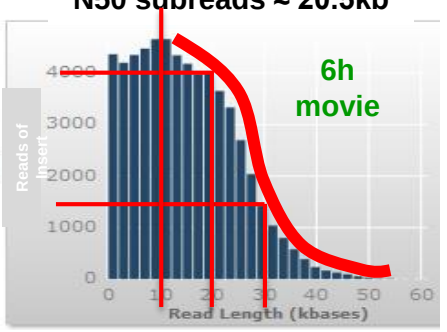
Top 10 of our longest subreads
80974 bp
79860 bp
79834 bp
78105 bp
77481 bp
76881 bp
76558 bp
76355 bp
75569 bp
75559 bp



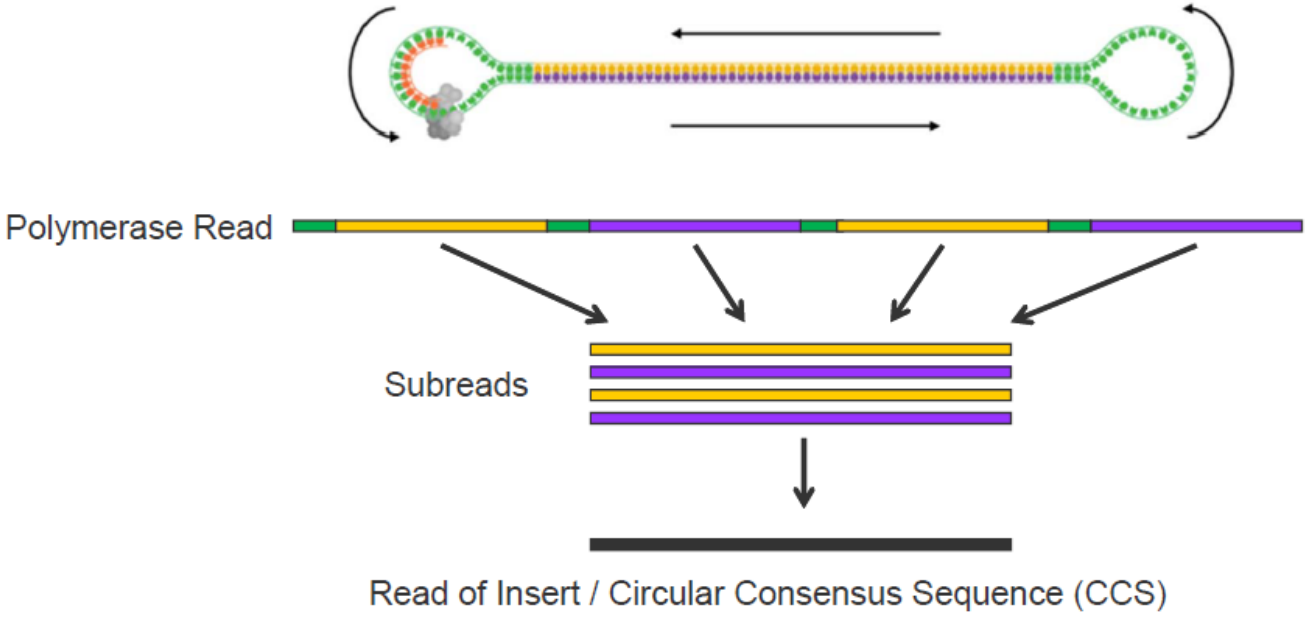
18kb cut off
(loading 0.36nM= 1036Mb
N50 subreads ≈ 19kb



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

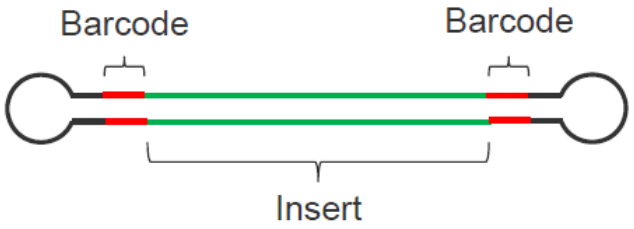


From Polymerase Reads to Read of Insert / CCS



- Subreads (purple and gold) are separated by adapter sequences (green)
- Read of Insert represents the highest quality single-sequence for an insert, regardless of number of passes
- ≥ 2 full polymerase passes required for CCS

Barcoding Background



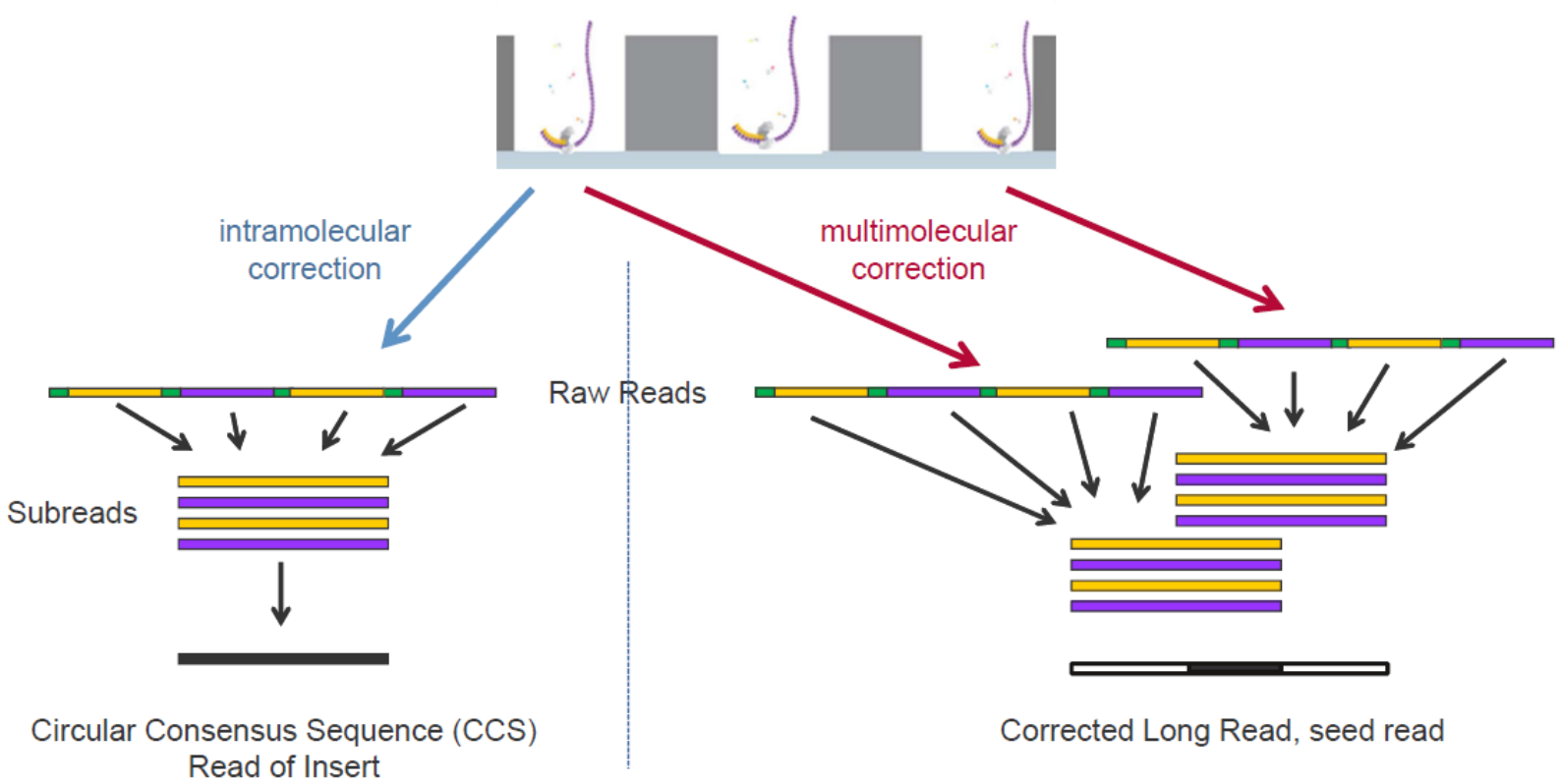
Short Insert

Polymerase will go around multiple times; multiple opportunities to view barcode

Long Insert

Few polymerases may make >1 pass; many polymerases may not see first barcode (or second one)

Correction: Intra- versus Intermolecular



GeT Génome et Transcriptome

From Sanger to 3rd generation

Unique sequence

Sanger

16 or 48 capillaries

Next Genration Sequencing

NGS - short reads

Illumina

2xHiSeq 3000
700 Gb
2 x 150 pb

3 x MiSeq
15 Gb
2 x 300 pb

Long Reads Sequencing

Chromium (10X genomics)

Oxford Nanopore
MinION

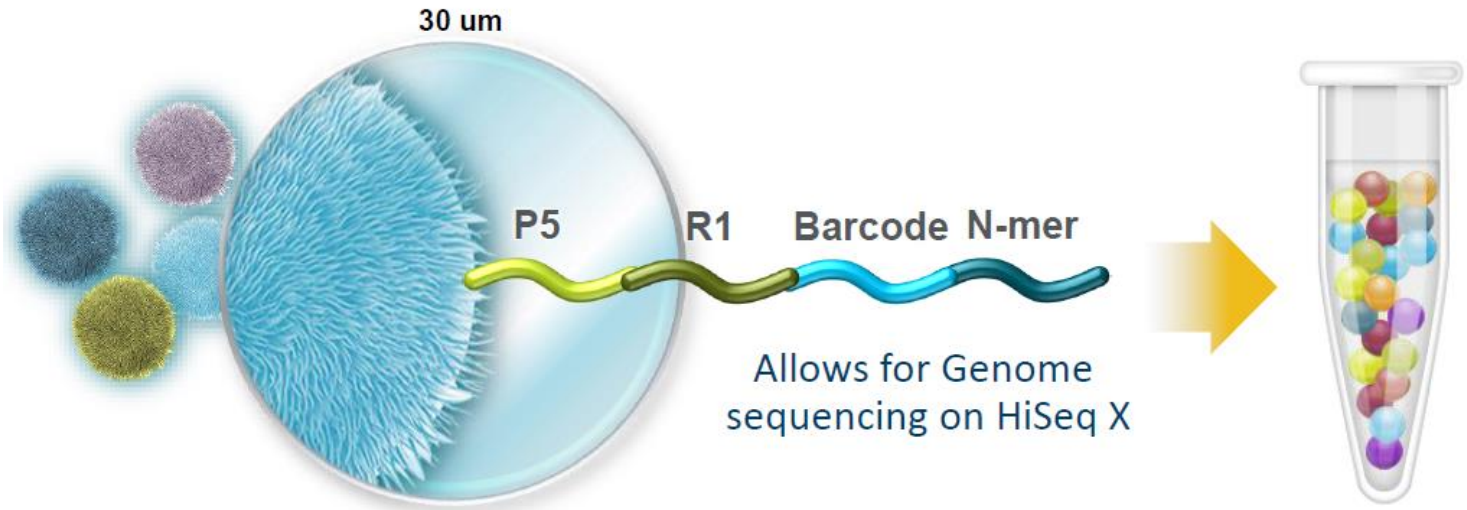
3G - long reads

PacBio RsII
70 000 reads - 20 kb

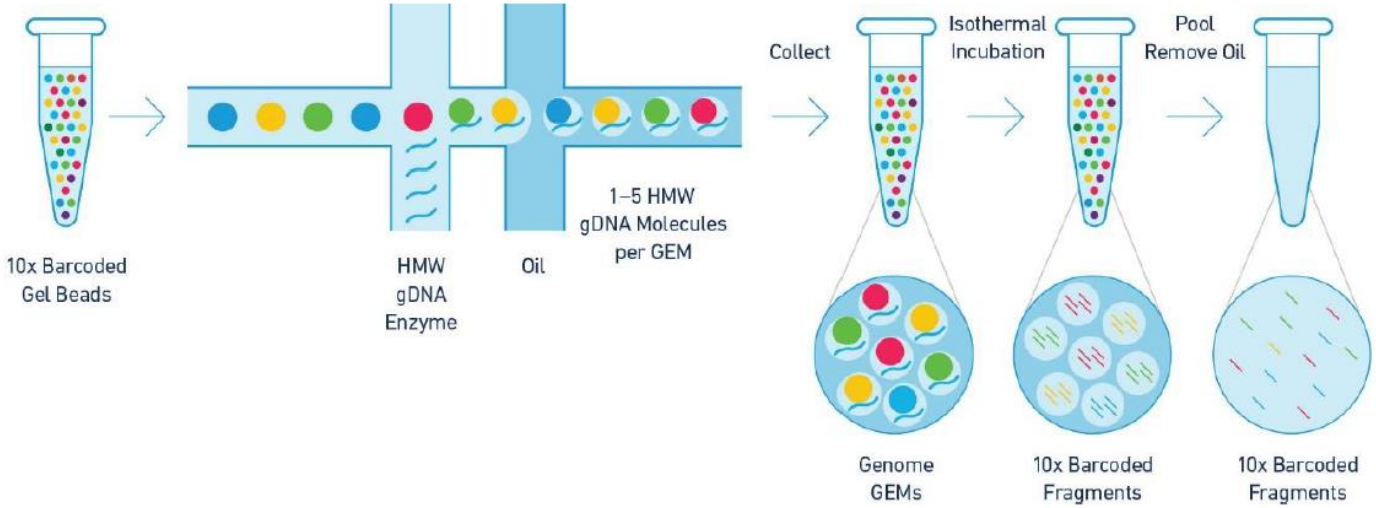
New techlonogy available

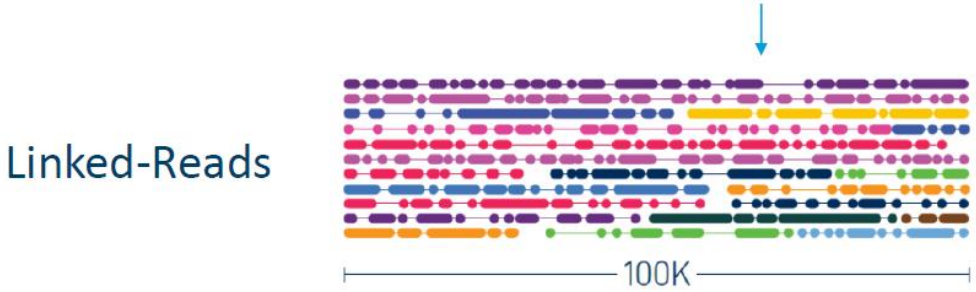
Chromium (10X genomics)

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

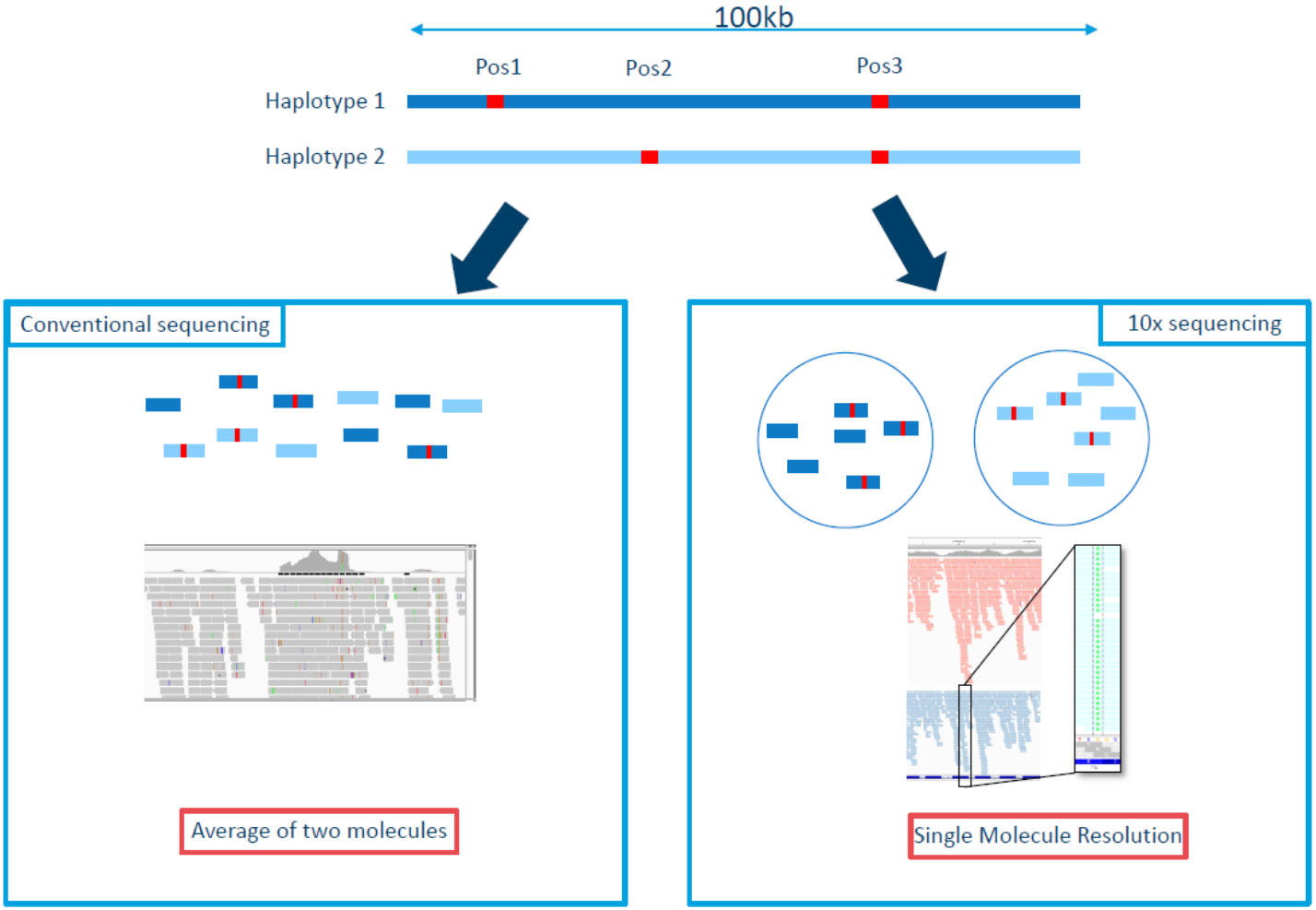


10XGENOMICS





10XGENOMICS



GeT Génome et Transcriptome

From Sanger to 3rd generation

Unique sequence

Sanger

16 or 48 capillaries

Next Genration Sequencing

NGS - short reads

Illumina

2xHiSeq 3000
700 Gb
2 x 150 pb

3 x MiSeq
15 Gb
2 x 300 pb

Long Reads Sequencing

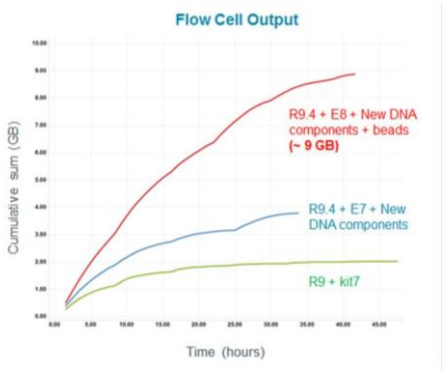
Chromium (10X genomics)

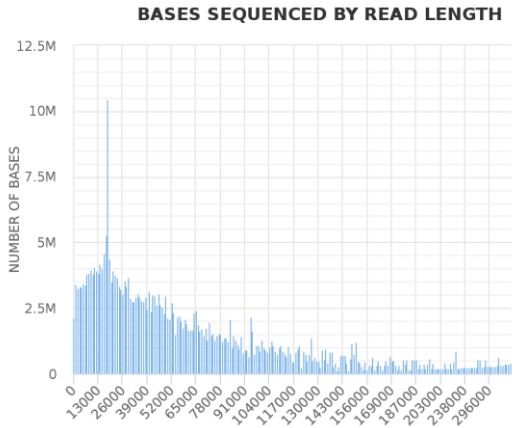
Oxford Nanopore
MinION

3G - long reads

PacBio RsII
70 000 reads - 20 kb

Oxford Nanopore





MinION

smidgION

PromethION

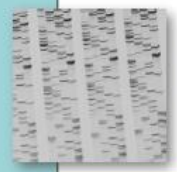
genotoul

GeT
Génome et
Transcriptome

L'ANALYSE QUALITÉ

Comparaison des technologies

1ère génération SANGER



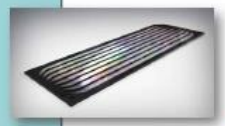
Fiche signalétique

- utilisé depuis : **1977**
- nombre max de séquences en parallèle : **96**
- longueur séquence : **1000 bases**
- pourcentage d'erreur: 1/1000
- coût séquence : €€€
- productivité :

Spécificités

- Génère et lit les séquences **UNE PAR UNE**
- A permis le séquençage du 1er génome humain

2nde génération ILLUMINA



Fiche signalétique

- utilisé depuis : **2003**
- nombre max de séquences en parallèle : **5 milliards**
- longueur séquence : **100 → 400 b**
- pourcentage d'erreur: 2/100
- coût séquence : €
- productivité :

Spécificités

- Génère et lit les séquences **PAR MILLIARDS EN PARALLELE**
- Les séquences sont de **petits fragments courts** d'ADN
- L'analyse des séquences nécessite des serveurs de calculs

3ème génération PacBio

Fiche signalétique

- utilisé depuis : **2013**
- nombre max de séquences en parallèle : **70 000**
- longueur séquence : **18 000 b**
- pourcentage d'erreur: 15/100
- coût séquence : €
- productivité :

Spécificités

- Sans photocopie de l'ADN
- Lit **DIRECTEMENT** des séquences **ULTRA-LONGUES**

Les formats de fichiers

Illumina	PacBio	Nanopore
Fastq	H5 - Bam	Fast5 (H5)

Fastq :
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!"*((((***+))%%%++)(%%%%).1***-+*"))**55CCF>>>>>>CCCCCCC65

Qualité : Score PHRED

Score de qualité phred	Probabilité d'une identification incorrecte	Précision de l'identification d'une base
10	1 pour 10	90 %
20	1 pour 100	99 %
30	1 pour 1000	99.9 %
40	1 pour 10000	99.99 %
50	1 pour 100000	99.999 %

H5 : HDF5 pour Hierarchical Data Format permet de structurer et de sauvegarder des fichier contenant de très grandes quantités de données

Critères de qualité

- **Critères de qualité**
  - ✓ Le nombre de reads produites correspondant au nombre attendu
  - ✓ Pas de contamination
  - ✓ Longueur des reads correcte
  - ✓ Bonne qualité, mais ce n'est pas un critère rédhibitoire
  - ✓ Bon alignement (re-séquençage avec génome de référence « propre ») : peu de reads non alignées

Analyses bioinformatiques et statistiques

- **Différentes approches d'alignement des séquences:**
  - ✓ **De novo : pas de génome de référence, transcriptome non disponible, très coûteux en terme de calculs, résultats très variables**
  - ✓ **Transcriptome de référence : la plupart sont incomplets**
  - ✓ **Génome de référence : le plus utilisé, permet l'alignement de reads sur des parties non annotées, nécessite un « spliced aligner » pour eucaryotes**

    → Etude à différents niveaux : gènes, transcripts, spécificité allèlique

    → Découvertes de nouveaux transcripts, nouveaux isoformes, nouvelles structures de gènes (fusion)

- **IMPORTANT : Discuter de la question biologique et du plan expérimental avec des Bioinformaticiens et Biostatisticiens <u>AVANT</u> de mettre en place l'expérience**

GeT Génome et Transcriptome

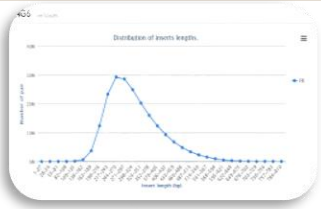
Main quality control workflows for Illumina's data

Core pipeline

- Statistics about demultiplexing *DemultiplexStat*
- Filter CASAVA 1.8/2.16 FASTQ file *Illumina Filter*
  - Contamination Search
  - Statistics on reads and their qualities *Fastqc*

DNA Resequencing *illumina_qc*

Alignment Statistics) → Visualization of insert Size Distribution



RNA Resequencing *illumina_rnaseq*

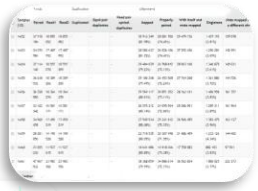
Alignment Statistics → Visualization of insert Size Distribution

Amplicon/16S sequencing *illumina_diversity_qc*

Paired-end reads merging → Assignation on a subset of sequences

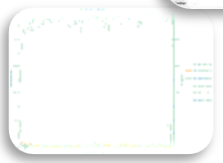
Mate Pair Sequencing *illumina_matepair*

Nextera MatePair Adapters trimming → Alignment Statistics → Visualization of insert Size Distribution



Whole Genome Bisulfite Sequencing *methylseq*

Adapters and quality trimming → Alignement and methylaytion calling (Bismark)

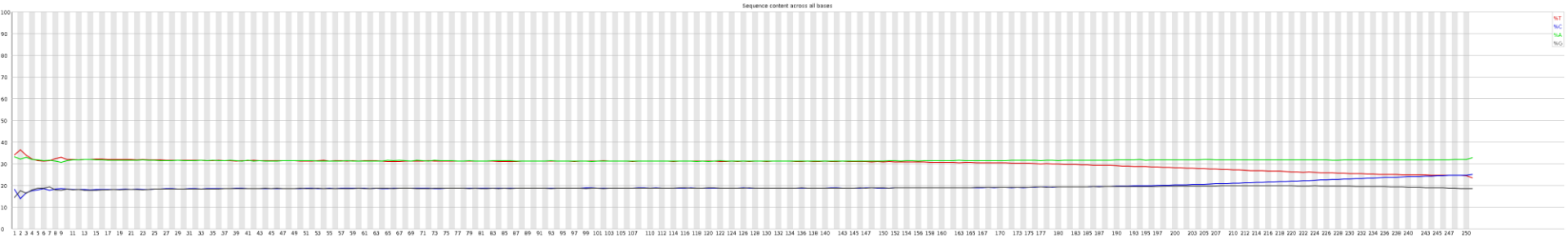


C1/mRNASeq

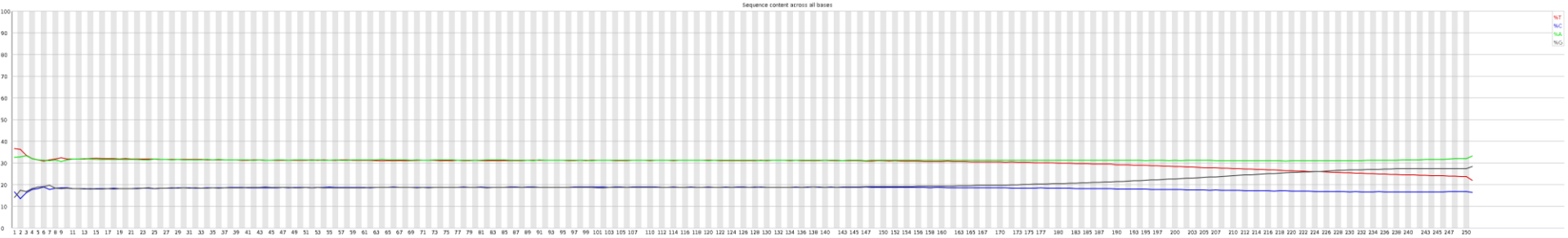
Transcription differences → Contamination Search / Spike

These NGS pipelines are provided by NG6[1], an information system built upon the jflow workflow management system.

For every specific protocol, a new workflow is built with jflow components, based on open source software or our in-house tools.


Sequence content across all bases
Position in read (bp)

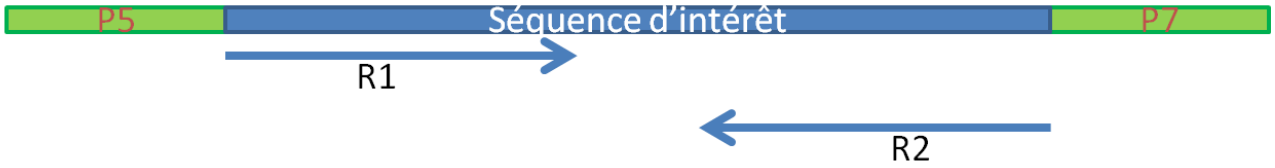
R1


Sequence content across all bases
Position in read (bp)

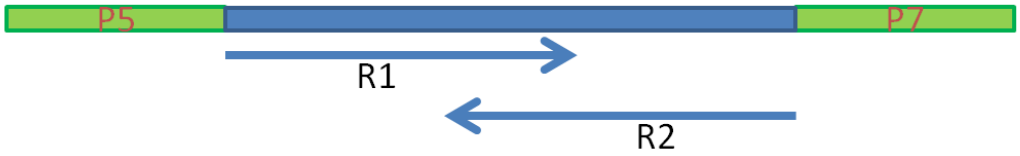
R2

Inserts trop petits

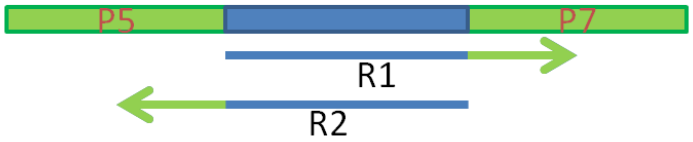
Séquençage classique sans overlap

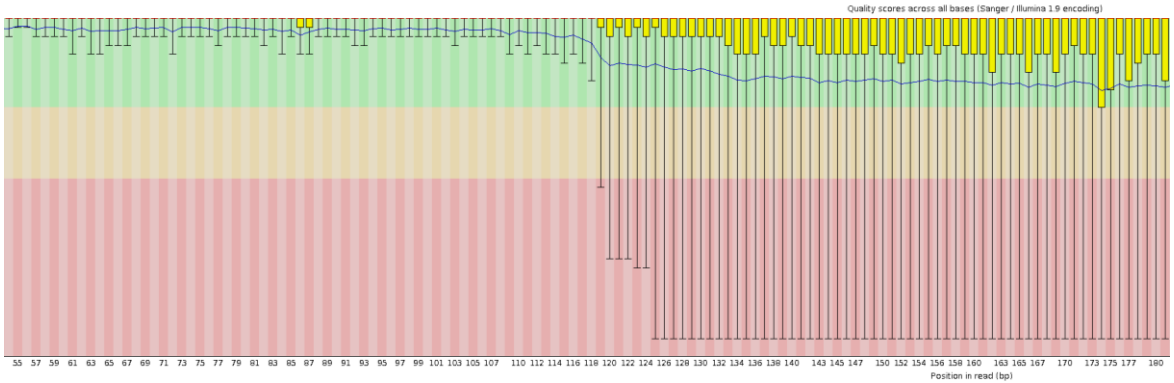


Séquençage classique avec overlap

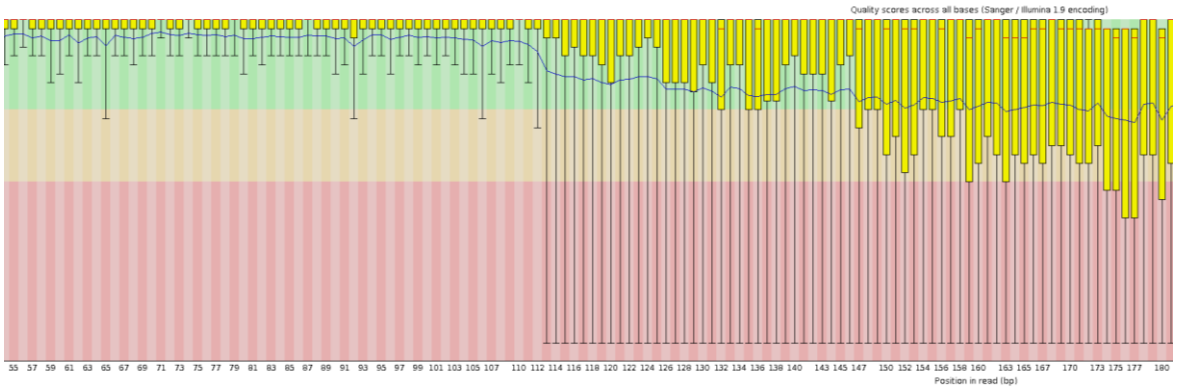


Séquençage avec overlap et adaptateurs

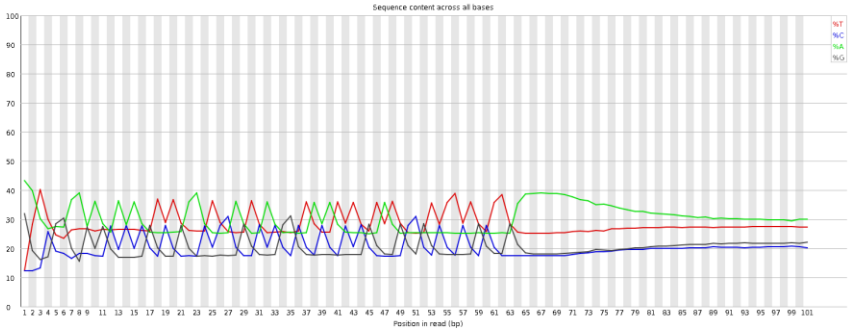


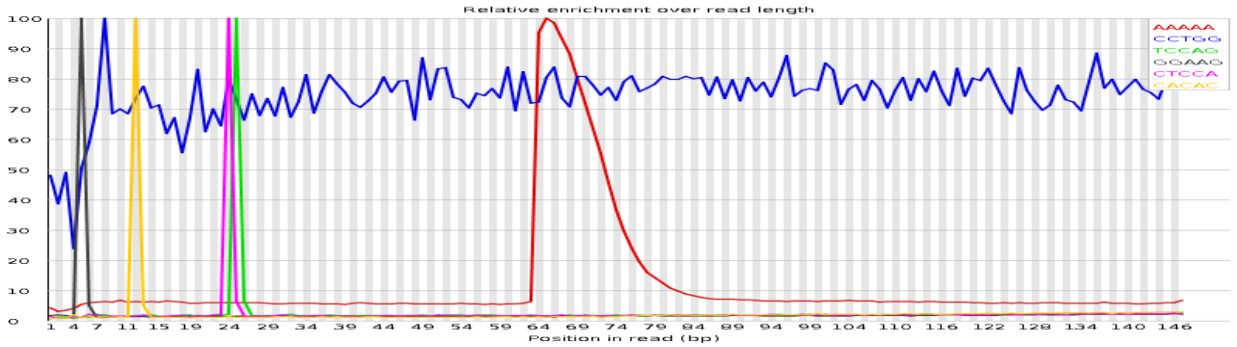

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
55 57 59 61 63 65 67 69 71 73 75 77 79 81 83 85 87 89 91 93 95 97 99 101 103 105 107 110 112 114 116 118 120 122 124 126 128 130 132 134 136 138 140 143 145 147 150 152 154 156 158 160 163 165 167 170 173 175 177 180
Position in read (bp)

R1


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
55 57 59 61 63 65 67 69 71 73 75 77 79 81 83 85 87 89 91 93 95 97 99 101 103 105 107 110 112 114 116 118 120 122 124 126 128 130 132 134 136 138 140 143 145 147 150 152 154 156 158 160 163 165 167 170 173 175 177 180
Position in read (bp)

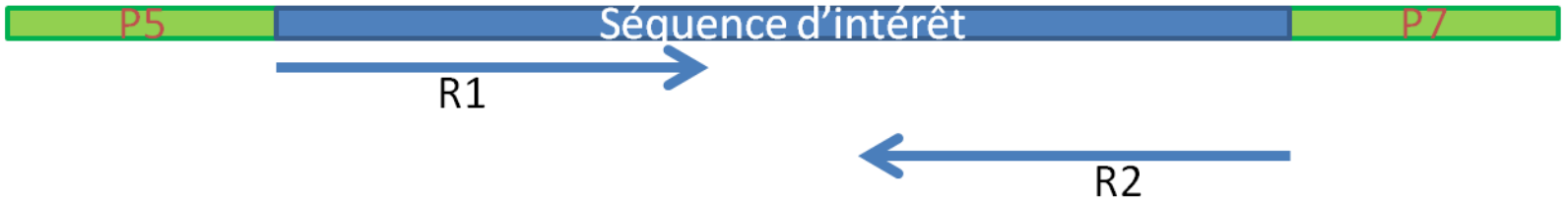
R2


Sequence content across all bases
%T
%C
%A
%G
Position in read (bp)


Relative enrichment over read length
AAAAA
CCTGG
TCCAG
GGAAG
CTCCA
GACAC
Position in read (bp)

Dimères d'adaptateurs

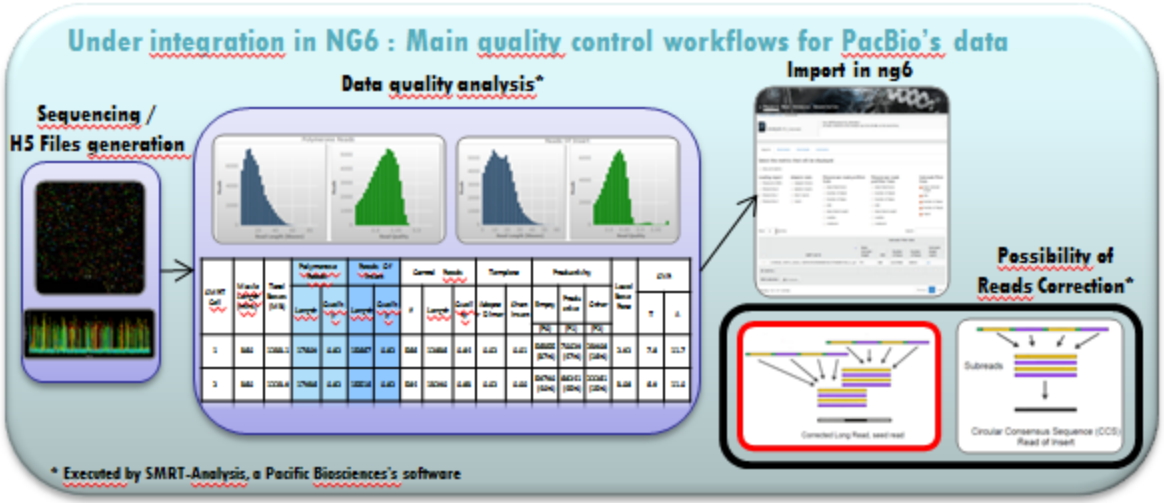
Séquençage classique



Séquençage de dimères d'adaptateurs

LIMS development

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



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