
Les technologies de séquençage à haut débit



Patrick Wincker, Genoscope, Institut de Génomique du CEA

- **Séquençage Sanger** (méthode des dididéoxy terminateurs) : a permis les progrès de la génomique jusqu'en 2006
- Les applications sont limitées par le coût
- **Les Nouvelles Technologies de Séquençage (NTSs)** répondent d'abord à une diminution des coûts, mais sont aussi motivées par la possibilité de nouvelles applications



Applied Biosystems
ABI 3730XL



Roche / 454
Genome Sequencer FLX



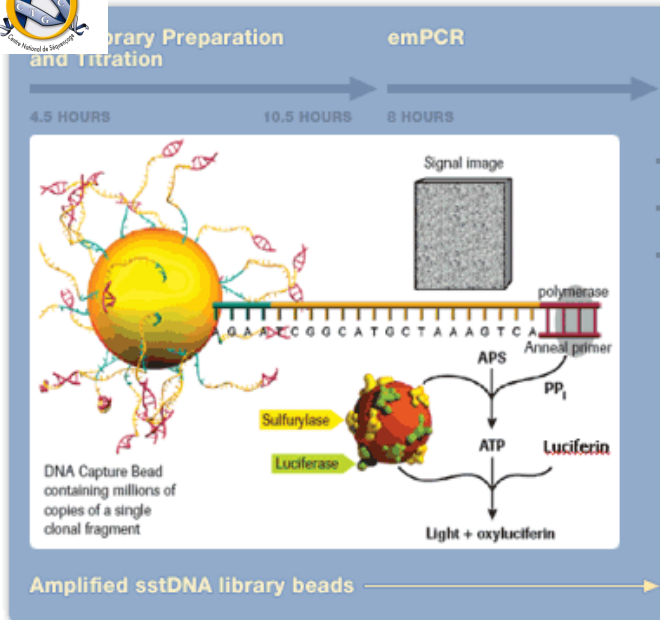
Illumina / Solexa
Genetic Analyzer



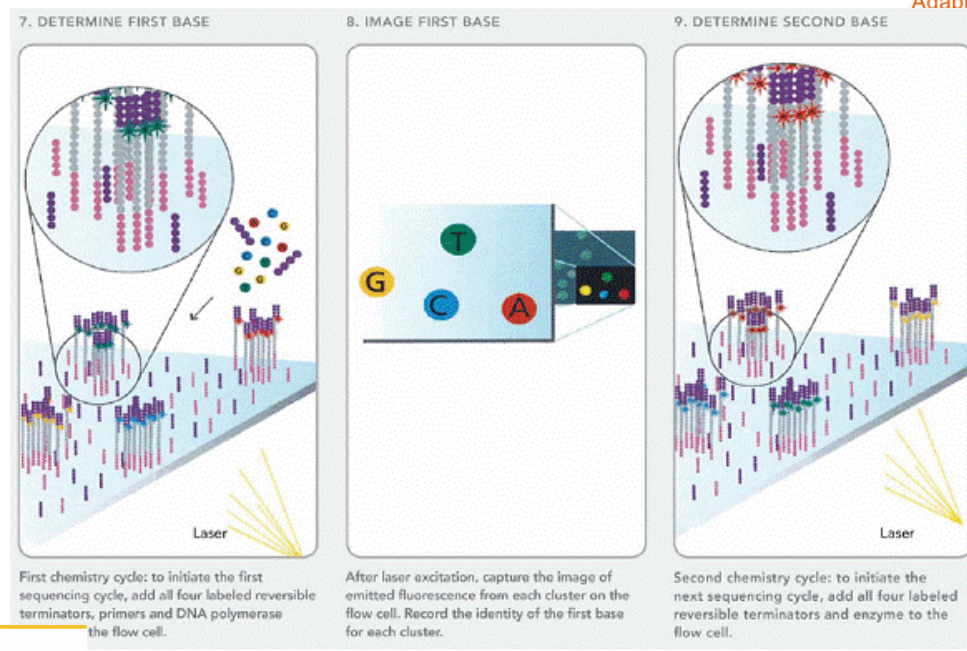
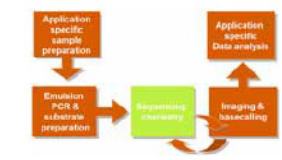
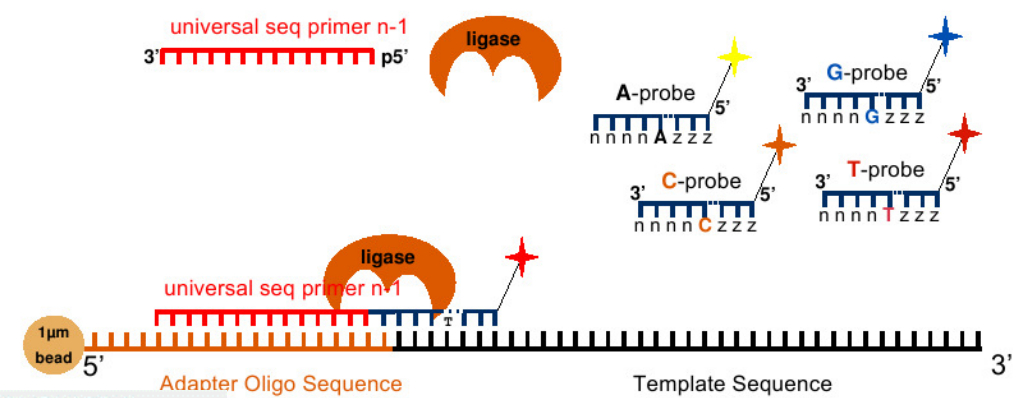
Applied Biosystems
SOLiD

Ce qui change :

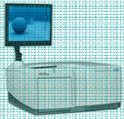
- La quantité et le type des données générées
- La qualité des données (type et taux d'erreurs)



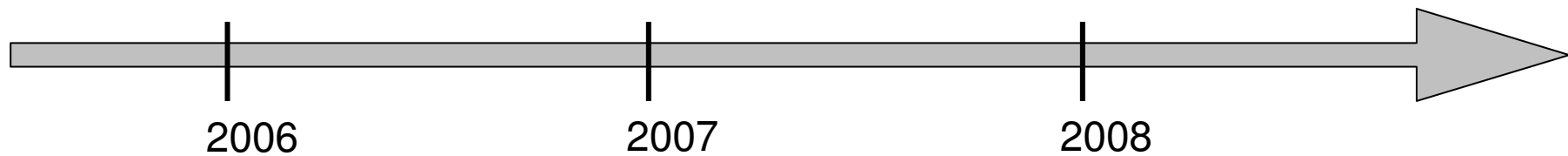
SOLiD 4-color ligation (1st cycle after reset)



- pas seulement une différence quantitative mais aussi qualitative : données intrinsèquement différentes
- d'où une réadaptation complète des procédures



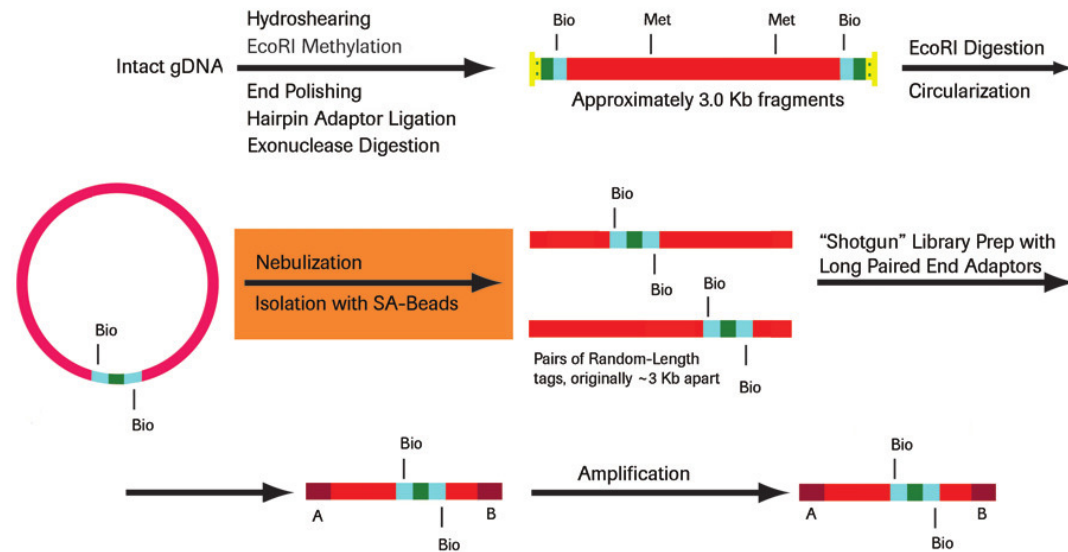
454 / Roche – Genome Sequence FLX



Gs20
20Mb / run
100pb / lecture

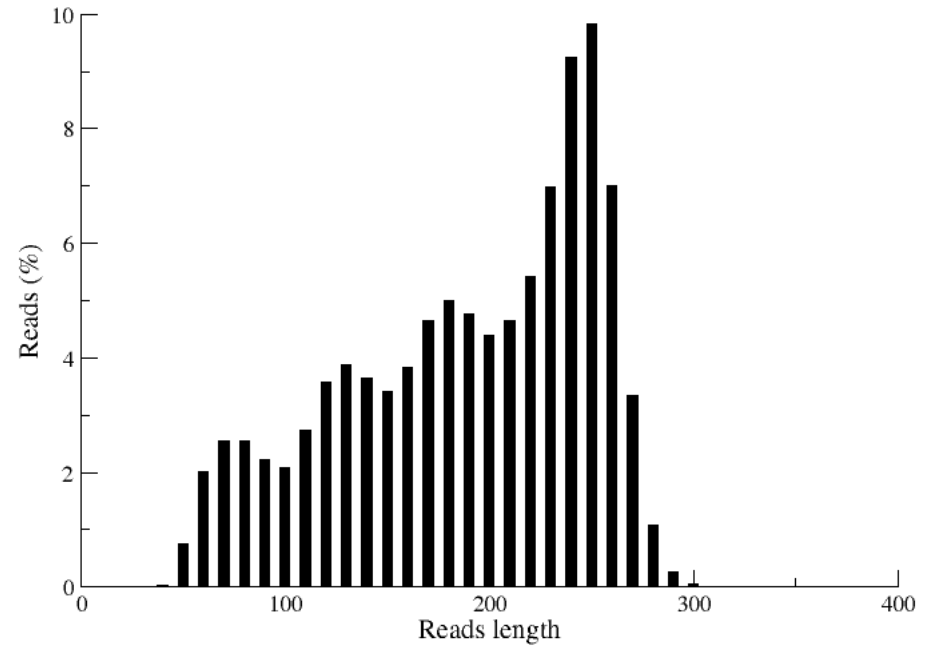
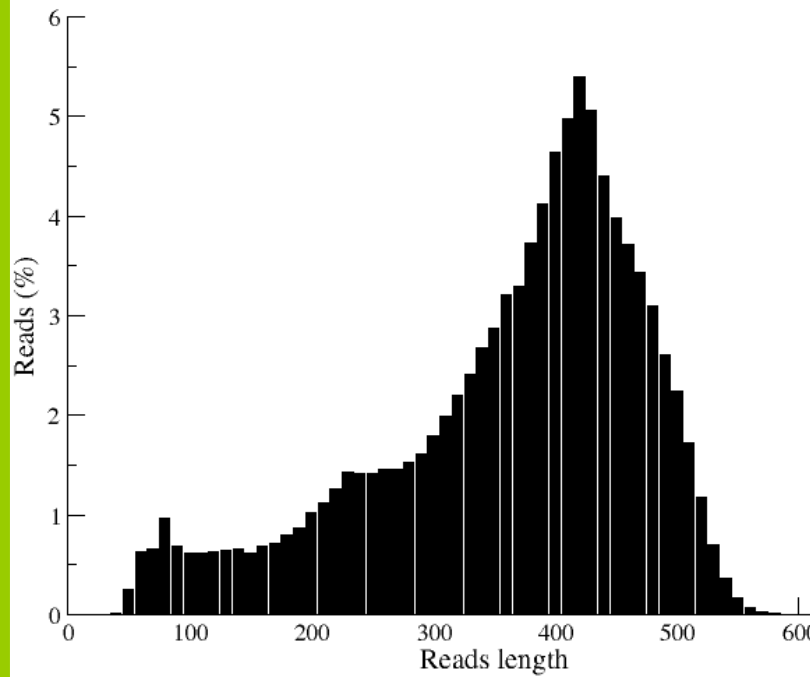
GsFLX
100Mb / run
250pb / lecture

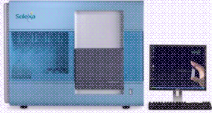
Titanium GS FLX
500Mb / run
400pb / lecture
+
Lectures en paires



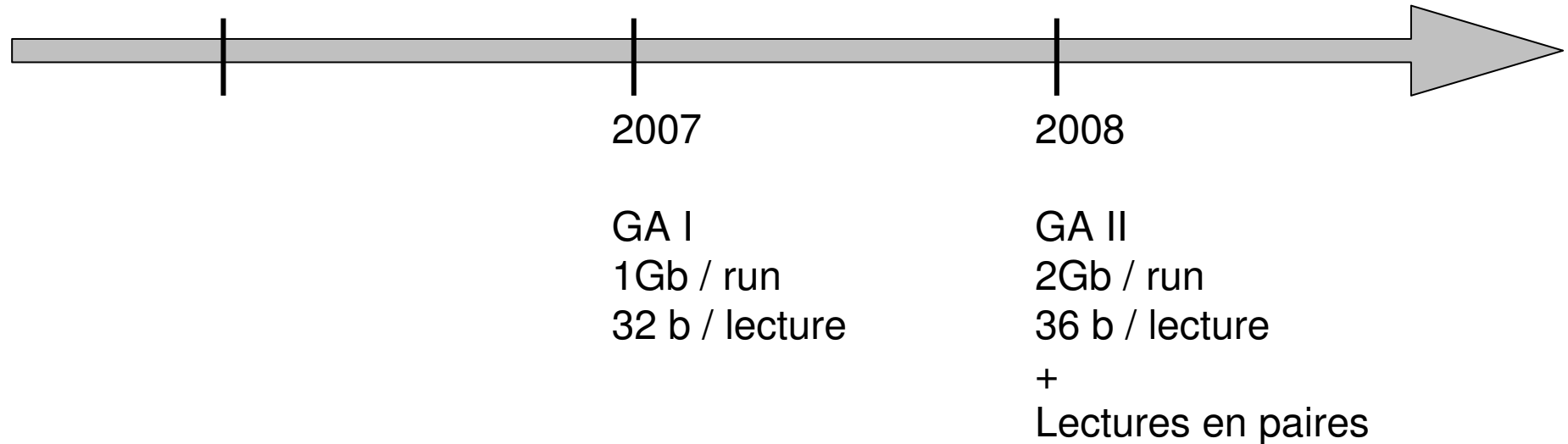
Données brutes

- ✓ Utilisation d'1/4 run Titanium : 141.771 lectures (51Mb
– 14,1X) : taille moyenne = 357pb





Illumina / Solexa – Genetic Analyzer 1G



✓ Version actuelle (Genome Analyzer II):

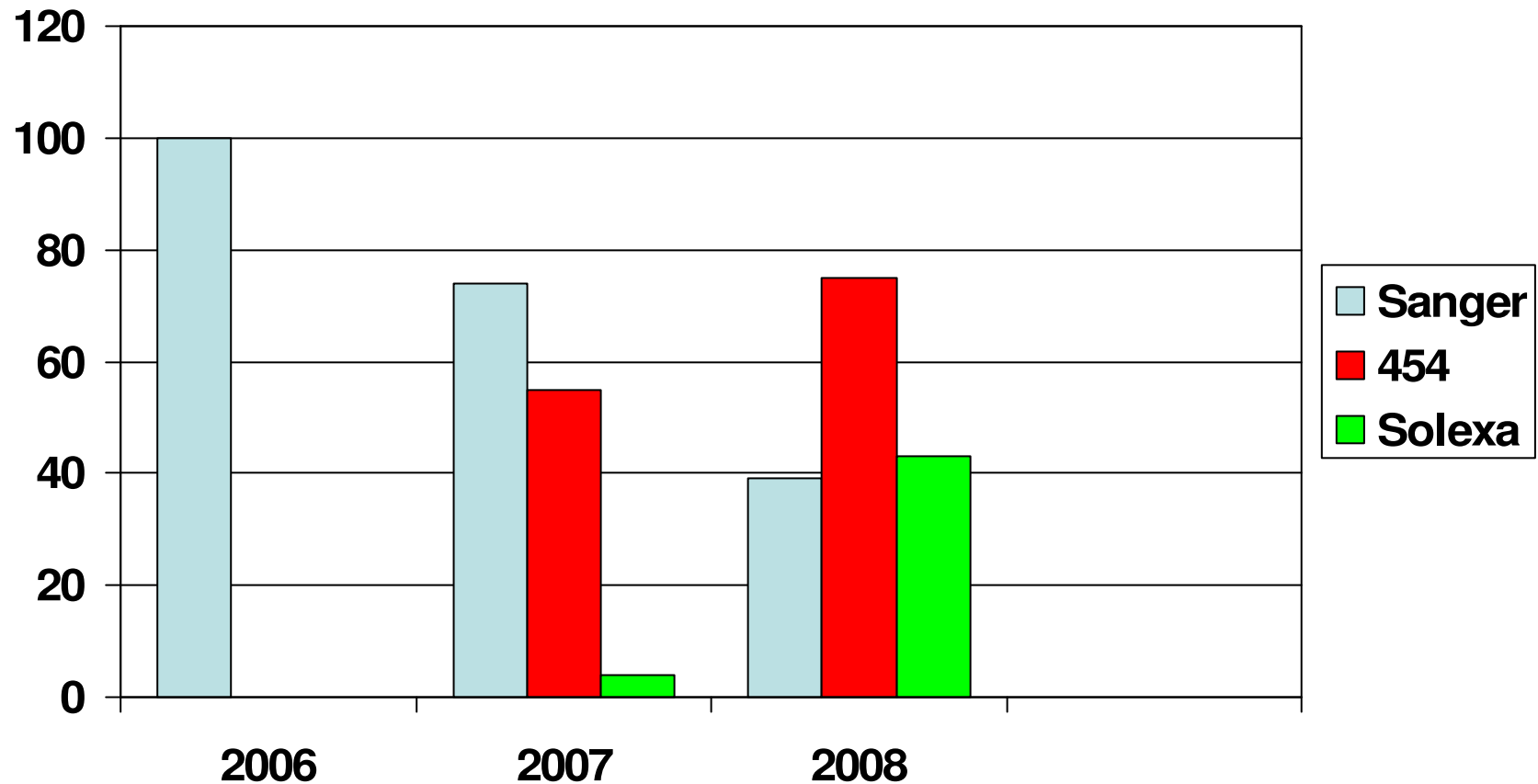
- ✓ Lectures de 36 à 108 bases
- ✓ Environ 60M lectures / run et ~2 à 6 Gbp / run
- ✓ 3-4 jours / run
- ✓ Taux d'erreurs très bas (~0,4%)
- ✓ Essentiellement des mismatches => complémentaires du GsFLX

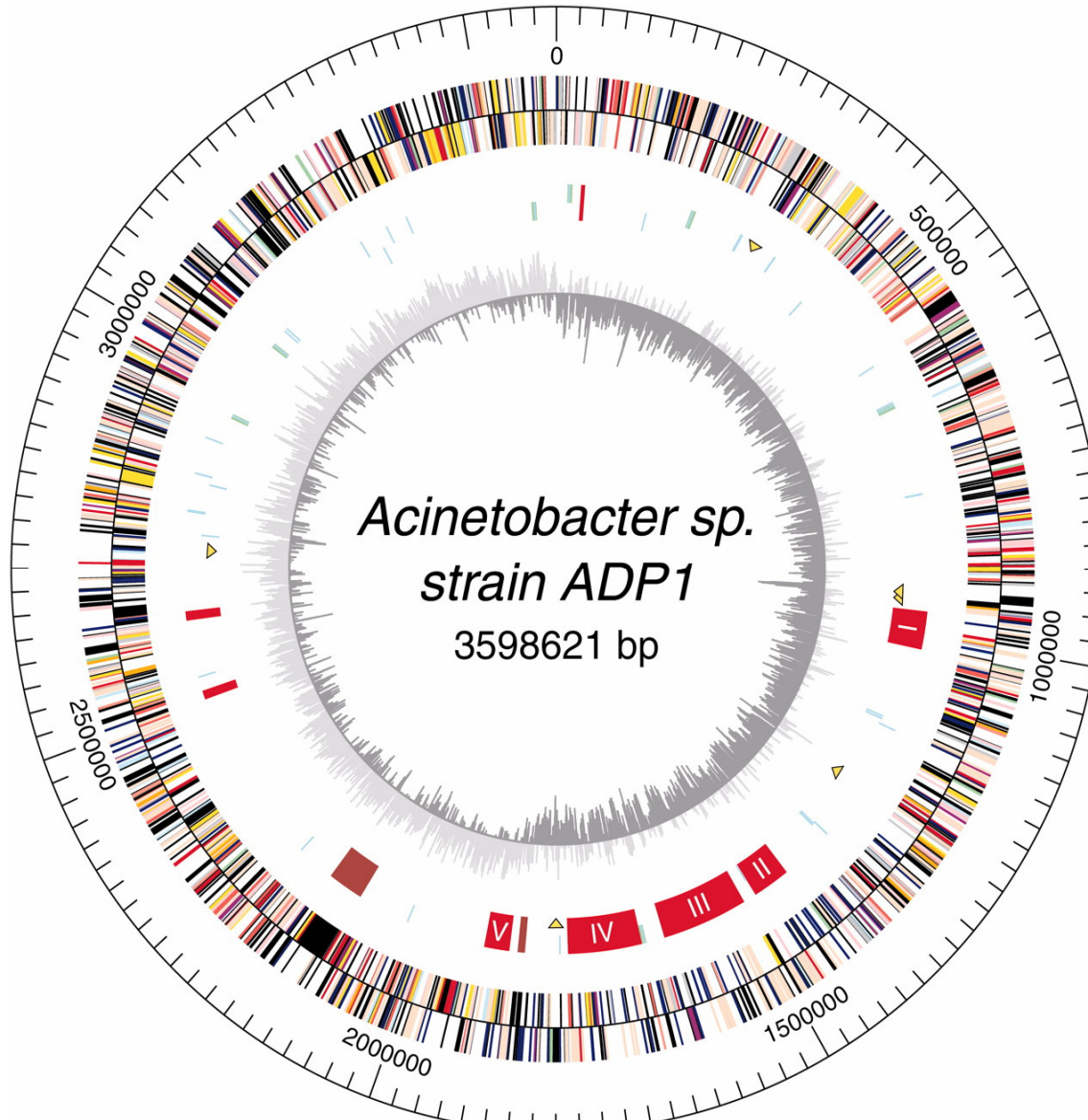
Evolution des besoins en grand séquençage : l'Appel à Projets Genoscope



- Appel d'offres annuel depuis 1998
- Introduction des technologies 454 (2007), Solexa (2008) et SOLiD (2009)

Evolution des demandes AP Genoscope par technologie



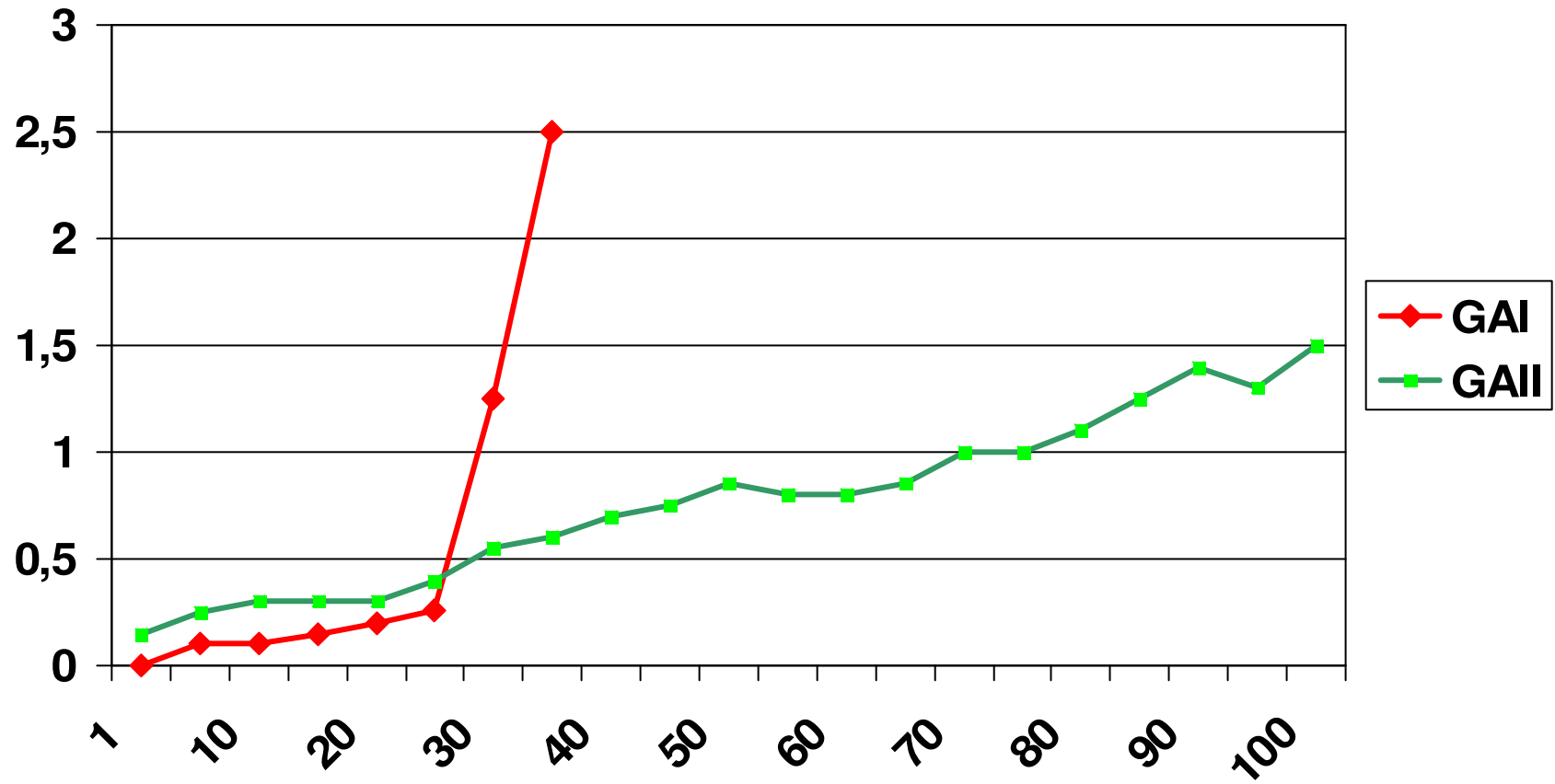


GC% = 40.3

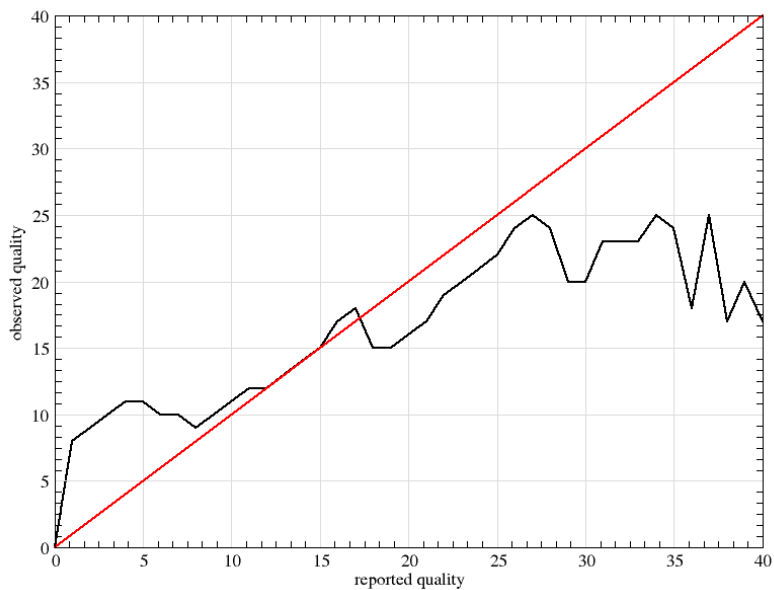
1.6% repetitive DNA

Manually finished sequence

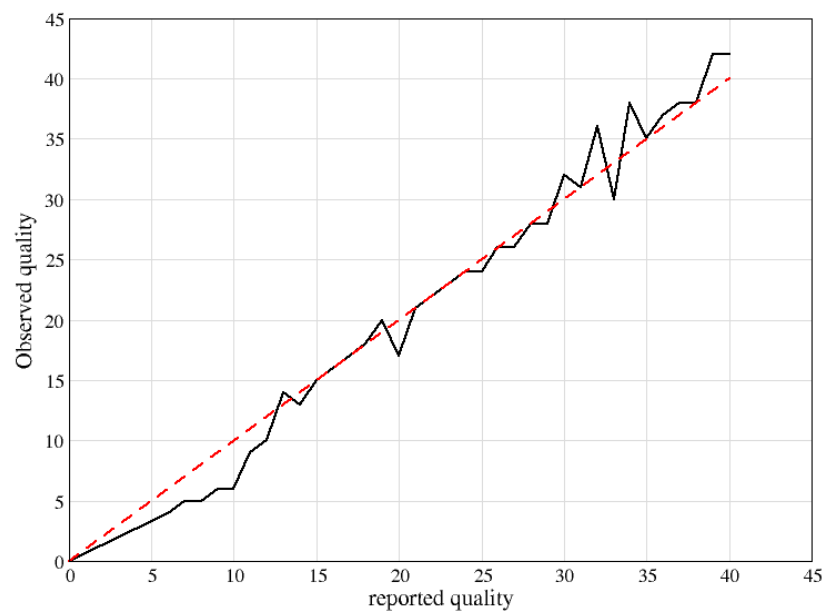
Taux d'erreurs GAI vs. GAI



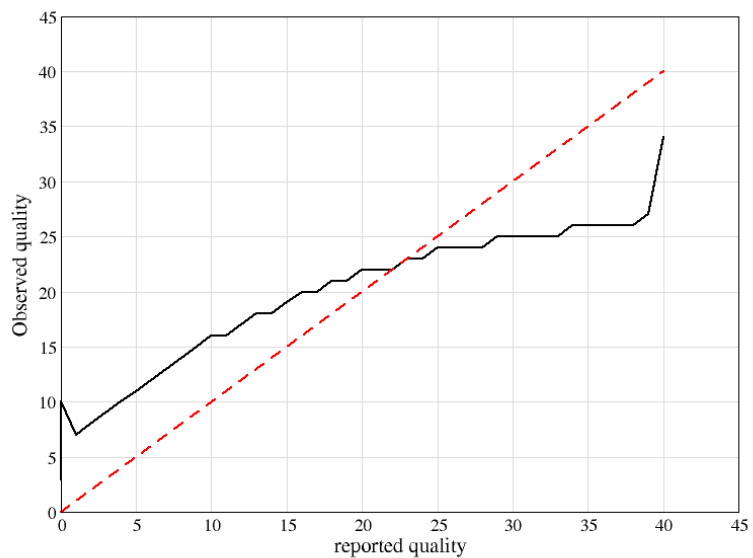
GS FLX base calls : initial version



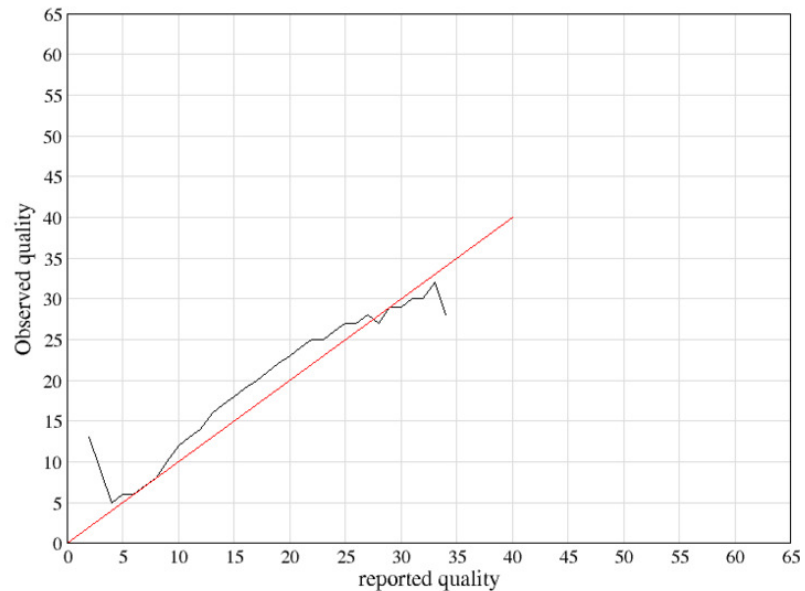
GS FLX base calls : upgrade March 2008



Solexa base calls GA I

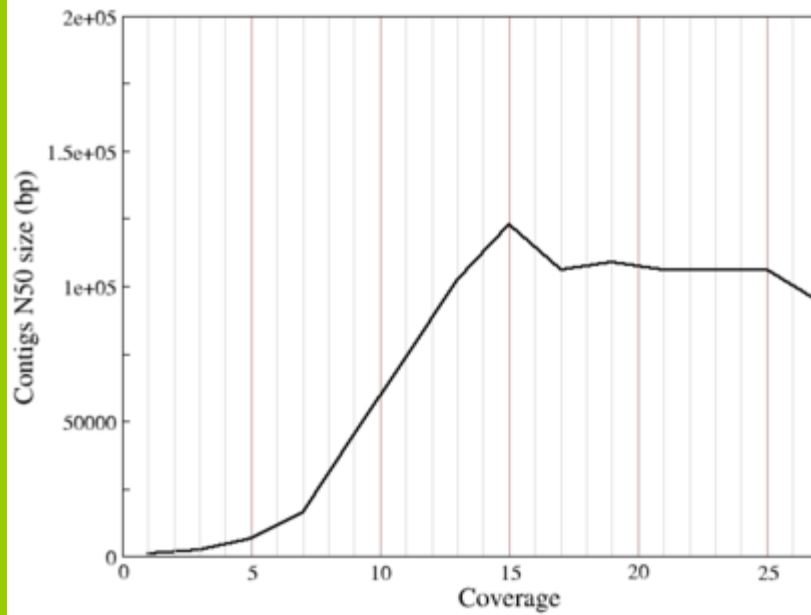


Solexa base calls GA II

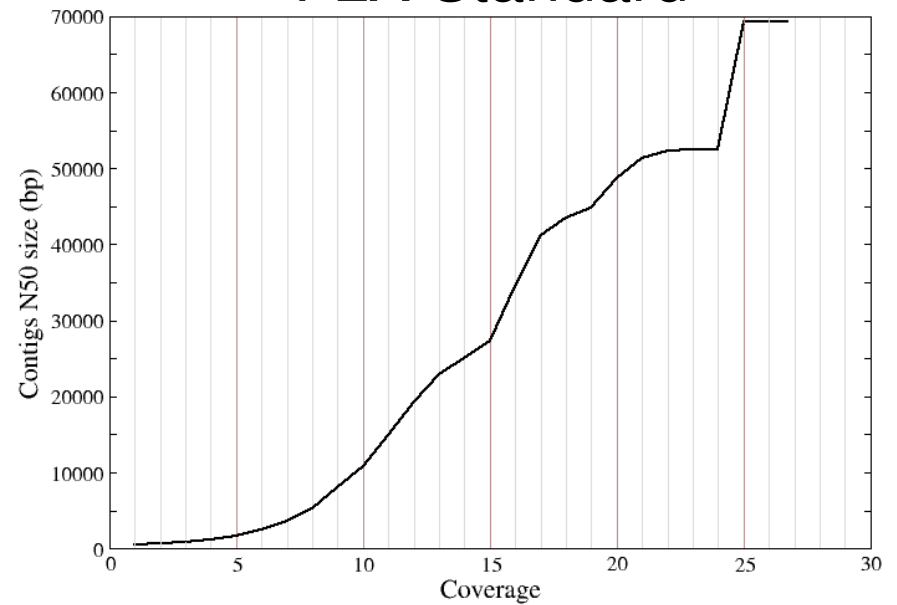


Assemblage 454

FLX Titanium



FLX Standard



Stratégie pour le séquençage de novo



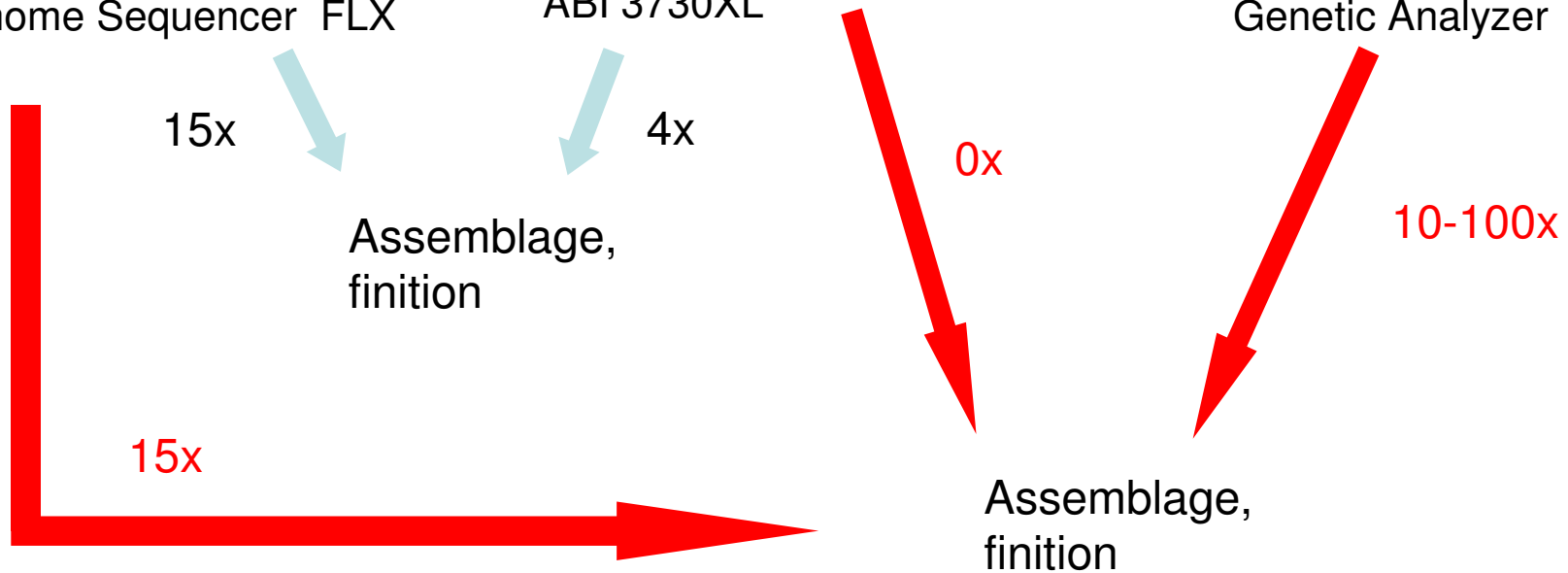
Roche / 454
Genome Sequencer FLX



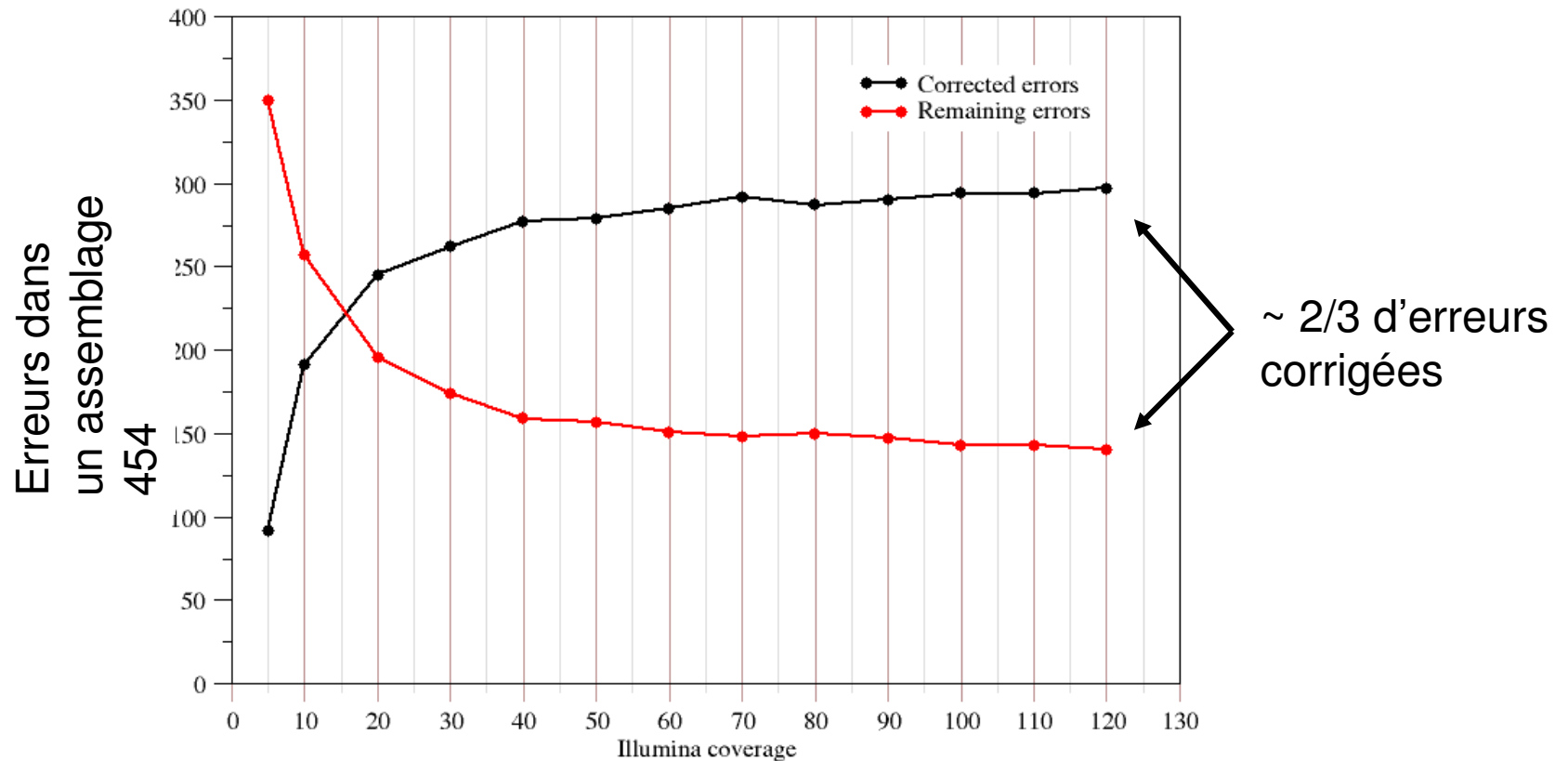
Applied Biosystems
ABI 3730XL



Illumina / Solexa
Genetic Analyzer

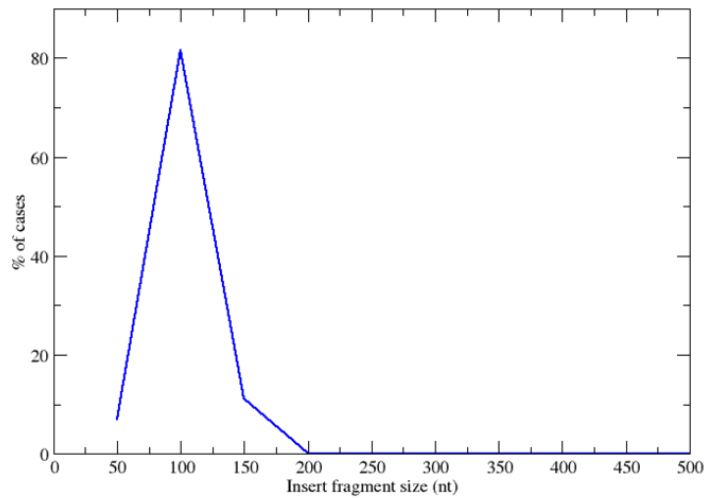


Complémentarité des technologies 454 / Roche et Solexa

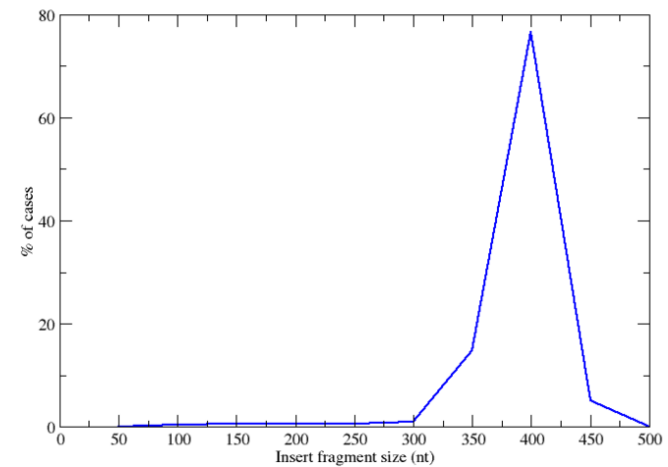


Le séquençage paired-end Illumina GAll, courts fragments

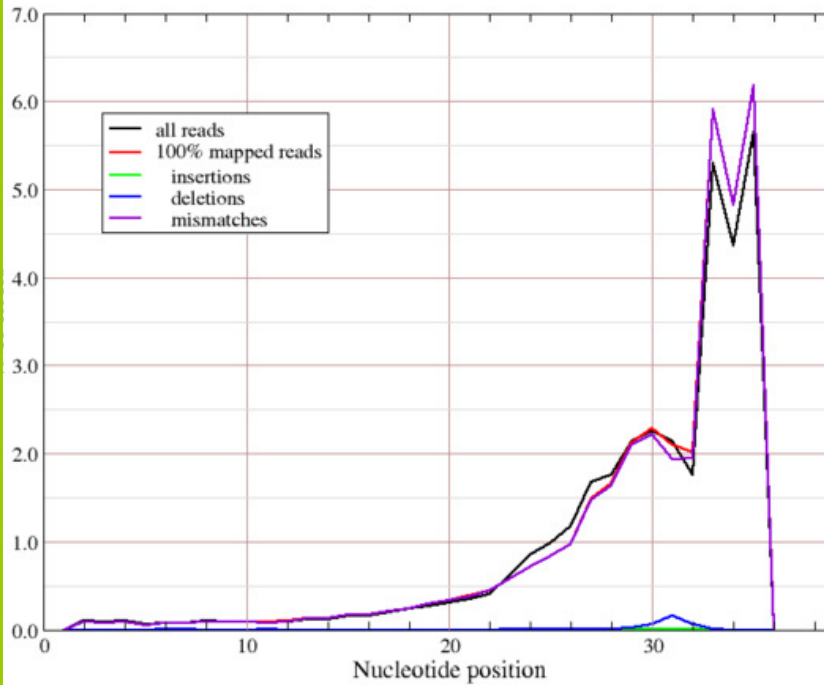
Lane 1



Lane 4

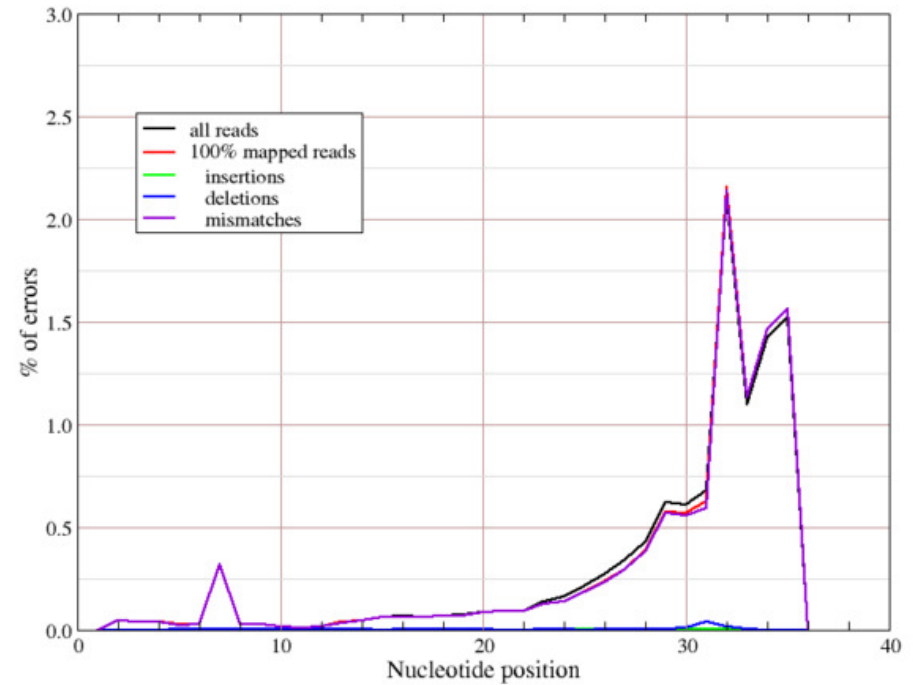


Solexa paired-end (GA I)



Mapping des extrémités droite

Number of errors : 203099 (**0.88%**)

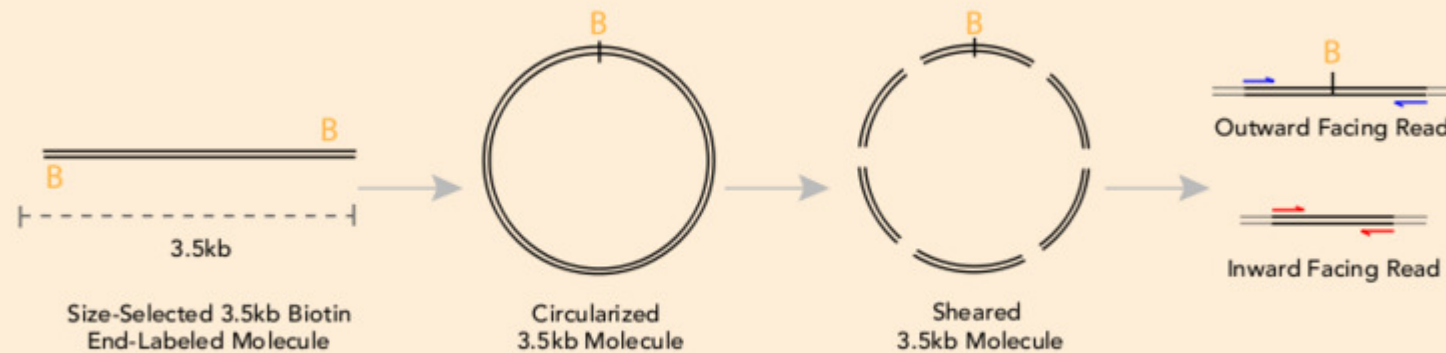


Mapping des extrémités gauche

Number of errors : 124183 (**0.29%**)

Illumina GAI : long paired-end

Origin of Inward and Outward Facing Reads

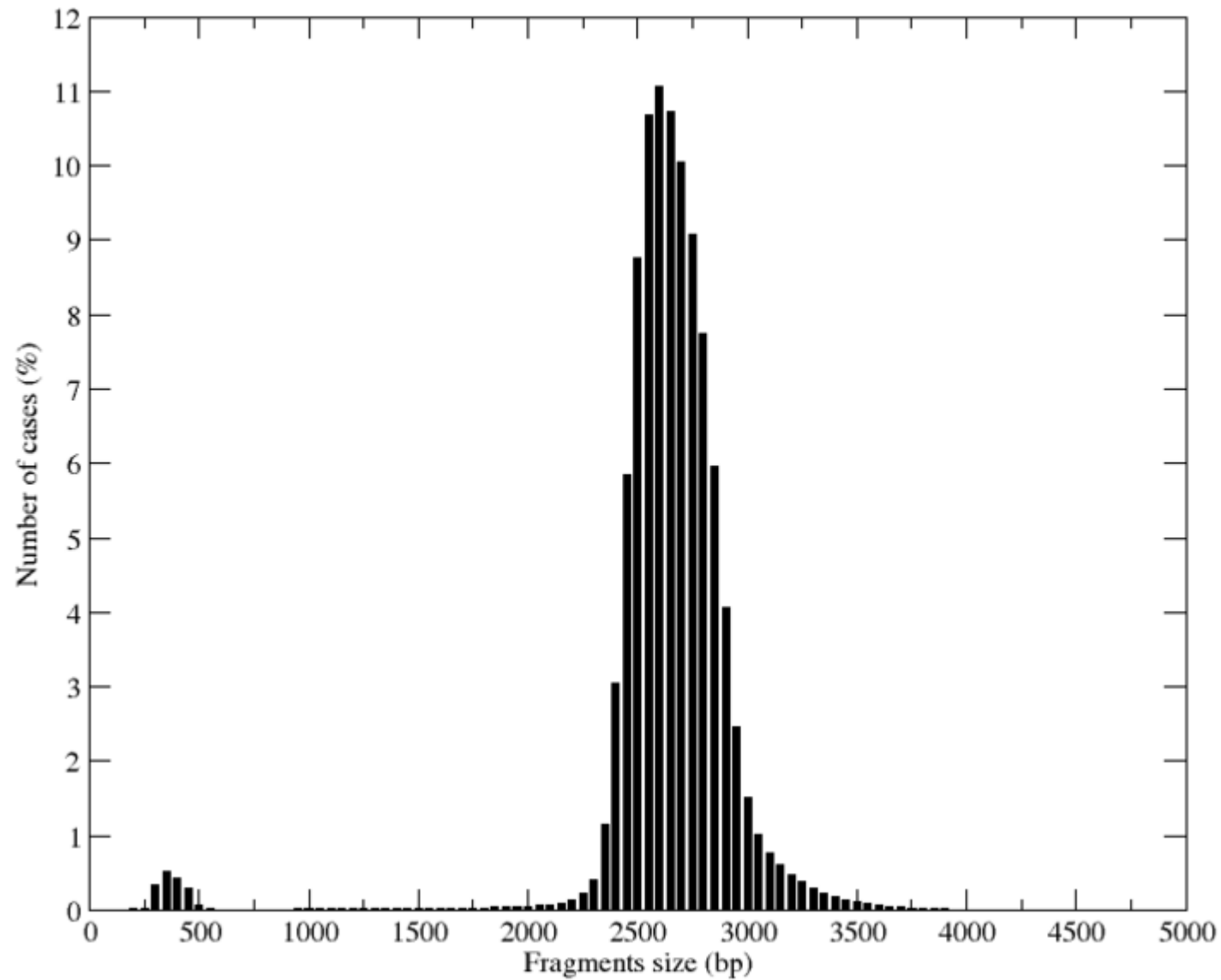


Alignment of Inward and Outward Facing Reads



Alignment of larger gap sized Outward Facing Reads (blue arrows) and shorter gap sized Inward Facing Reads (red arrows) back onto A) Fragmented 3.5kb circularized molecule B) Linear size-selected molecule C) Genomic Reference sequence.

Illumina GAI : long paired-end

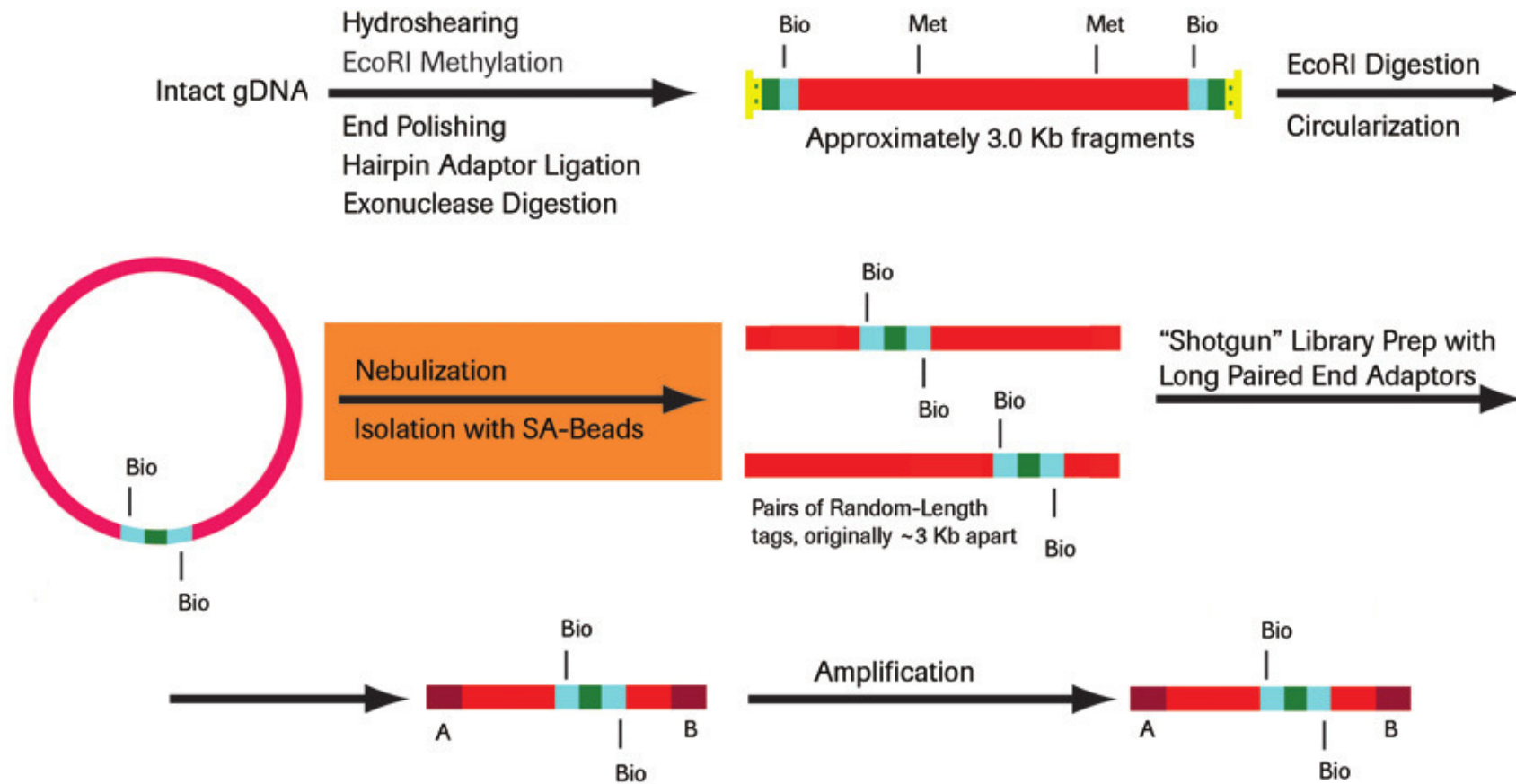


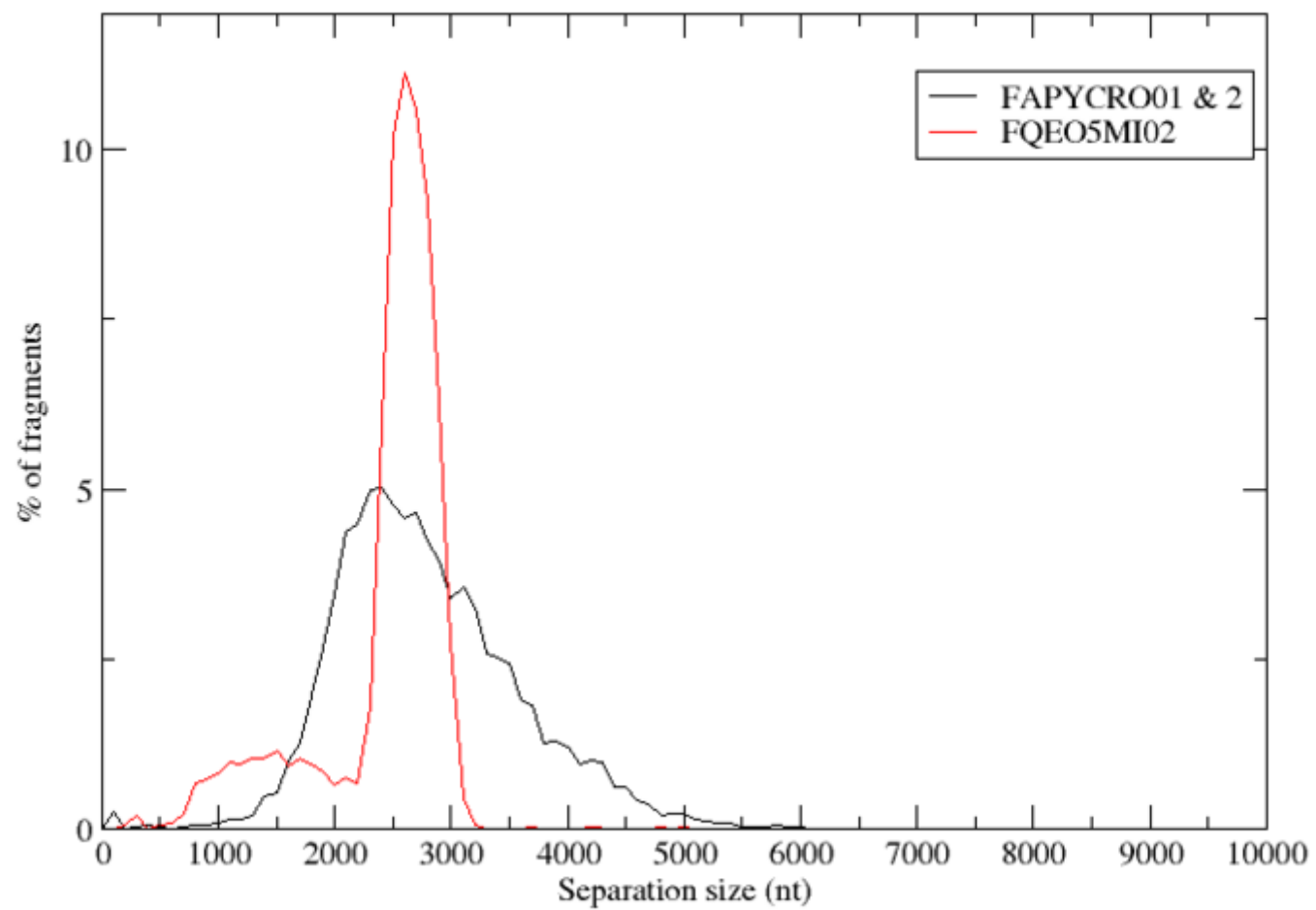
Assemblages d'un génome de levure

Comparaison GAI paired-end vs. 454+Sanger

	Solexa	454 single + Sanger
• Couverture initiale	~45X	~30X
• Nb de supercontigs	109,995	360
• Nb (%) de bases	15,848,259	12,343,951
• Taille moyenne	144bp	34Kb
• Taille N50	62Kb	156Kb
• Taille max	364Kb	599Kb

454 GSFLX : Paired-end sequencing

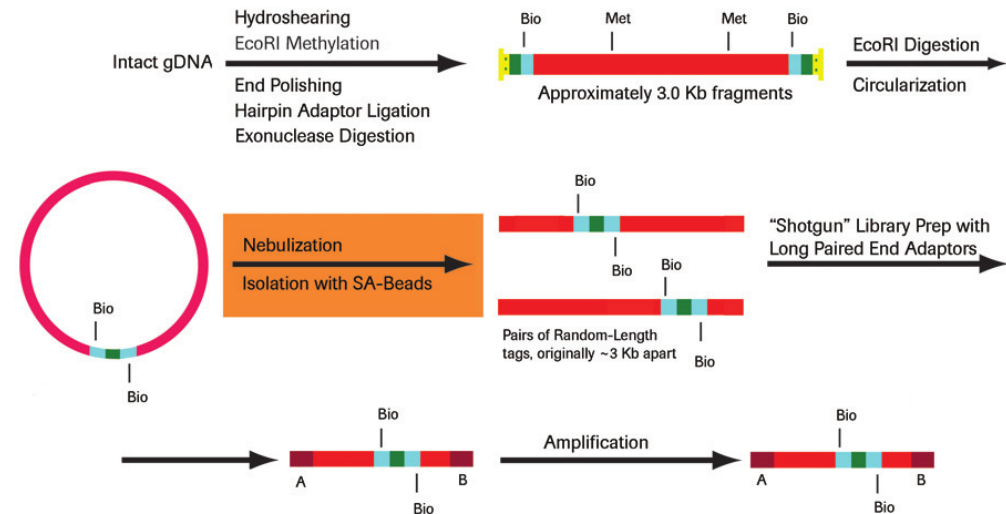






GSFLX paired-end sequencing

- Nb de lectures pairées (en % lectures totales) :



- Protocole standard : 26.34%
- Protocole amélioré : 41.93%

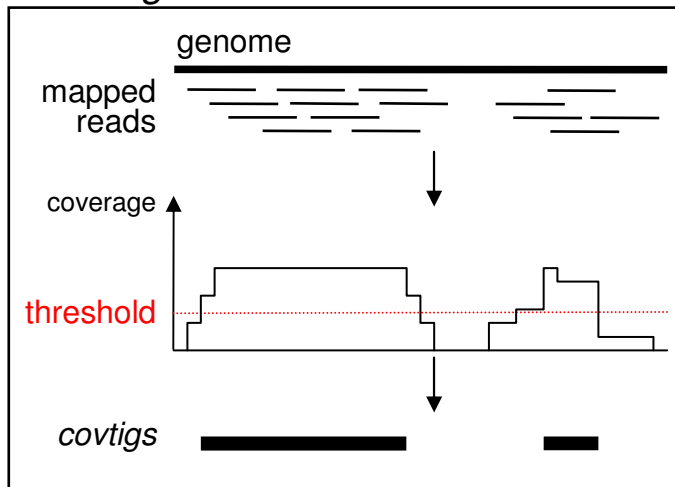


Use of High-Throughput data for annotation of the grape genome

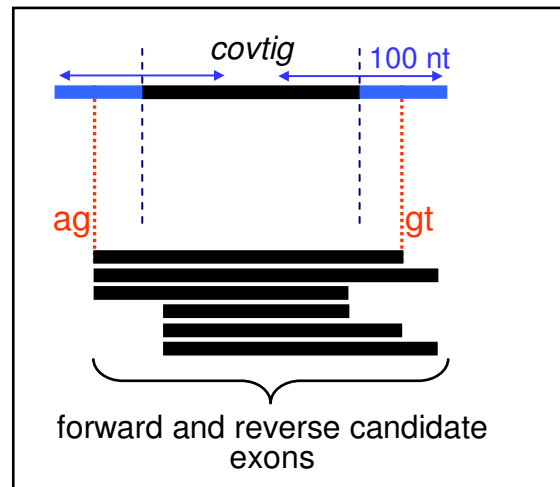
- 173 millions Solexa/Illumina GA1 reads
- Many putative genes are tagged
- But the reads also map in non-genic regions
- Development of a method to discriminate genes from background transcription, and to create gene models by connecting short sequence tags

Use of short reads to build gene models

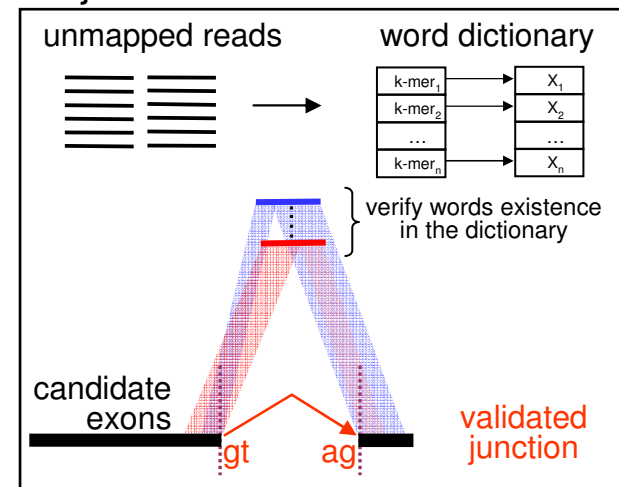
1. *covtigs* construction



2. candidate exons

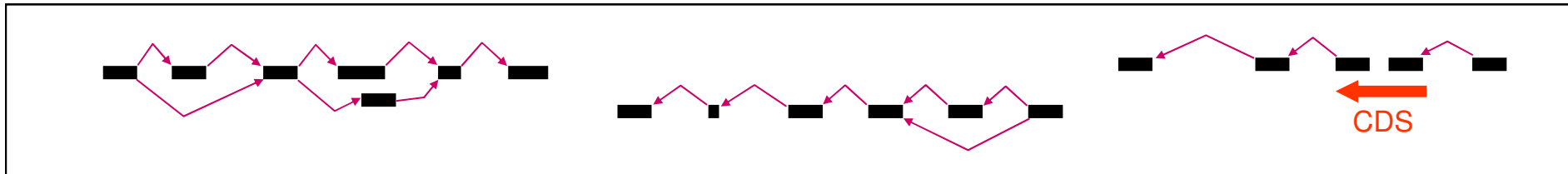


3. junction validation

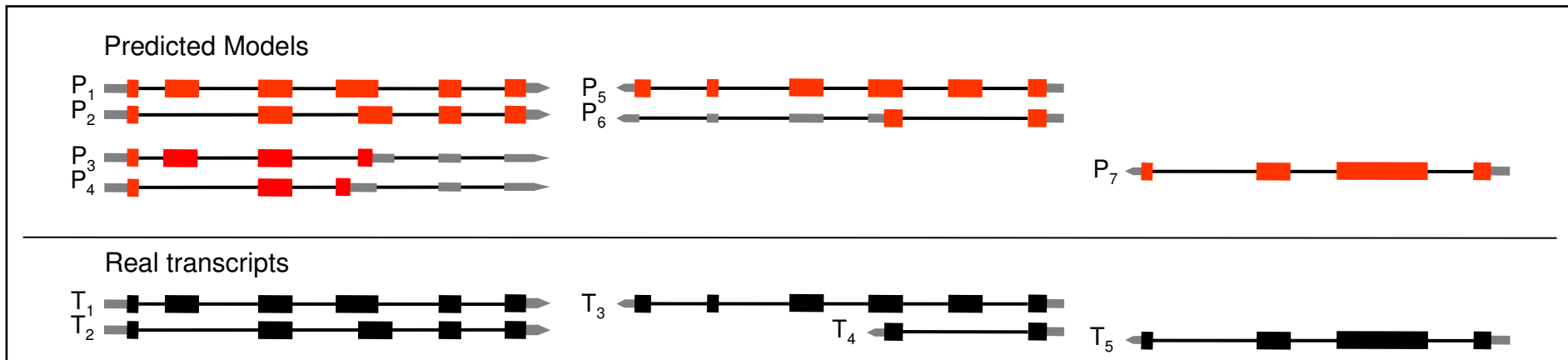


Use of short reads to build gene models /2

4. graph of candidates exons linked by validated junctions



5. models construction and coding sequences detection



Usefulness of short reads in annotating the grape genome

- New splice variants discovery (11,842 versus 2,171 using existing cDNA resources)
- Bring cDNA confirmation for ~50% of the gene models
- 675 new gene models (small size)
- Test with paired-end reads



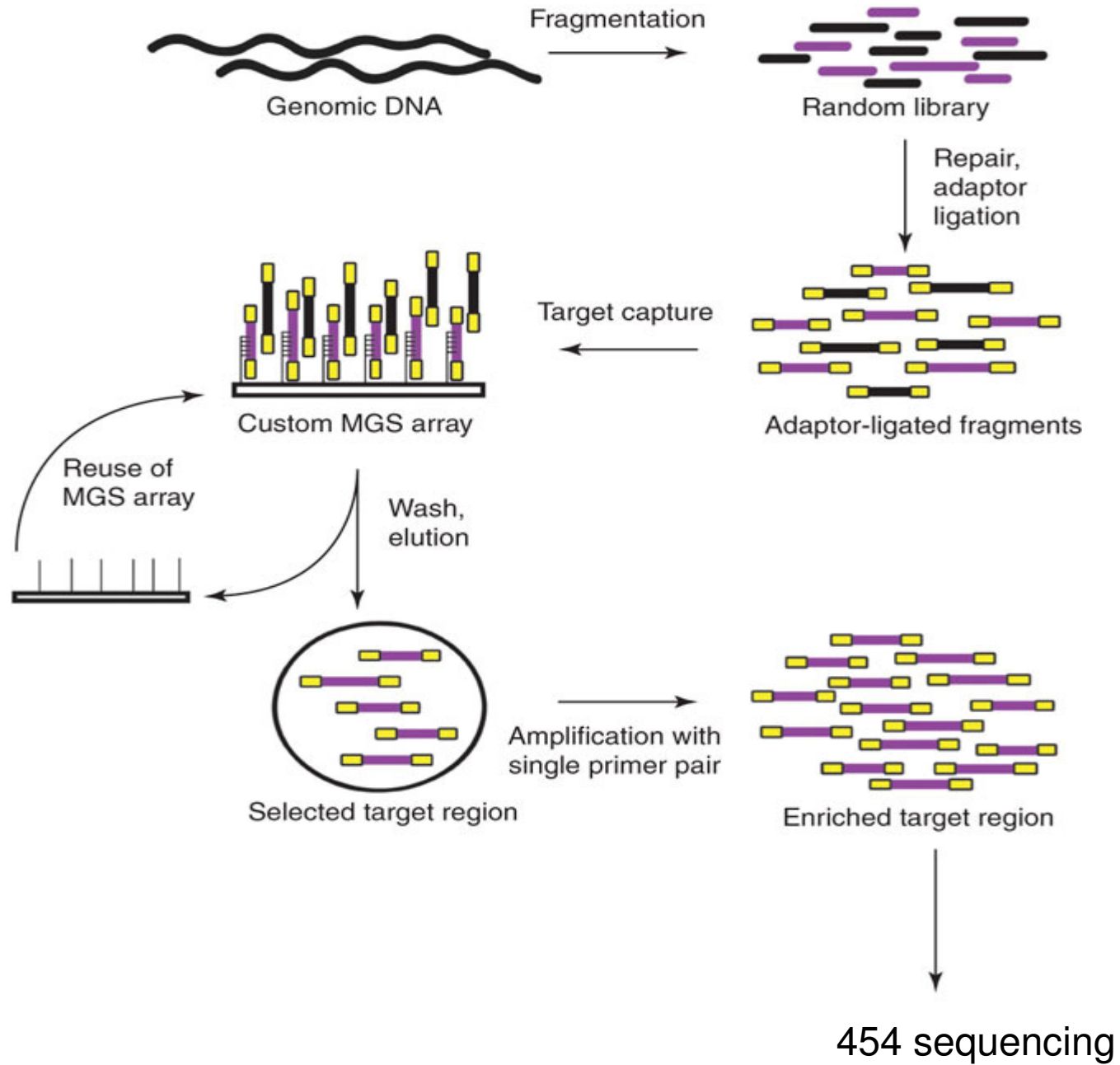
Reséquençage / SNPs/ Détection de mutations

Trois approches :

- ✓ Reséquençage d'un génome entier sous forme de microreads
- ✓ Amplification de régions génomiques cibles et séquençage des produits de PCR en 454
- ✓ Capture d'exons : Restreindre l'ADN à séquencer à la région d'intérêt



Plateforme Recherche de Mutations



- Les NTSs actuelles (2^{ème} génération) vont vers des lectures de + en + longues de - en - chères
- Mais la question se pose : quelle pérennité face aux séquenceurs de 3^{ème} génération?



Perspective : Séquenceurs de 3^{ème} génération

Le séquençage d'ADN
simple-molécule

En temps réel plutôt que par
cycles



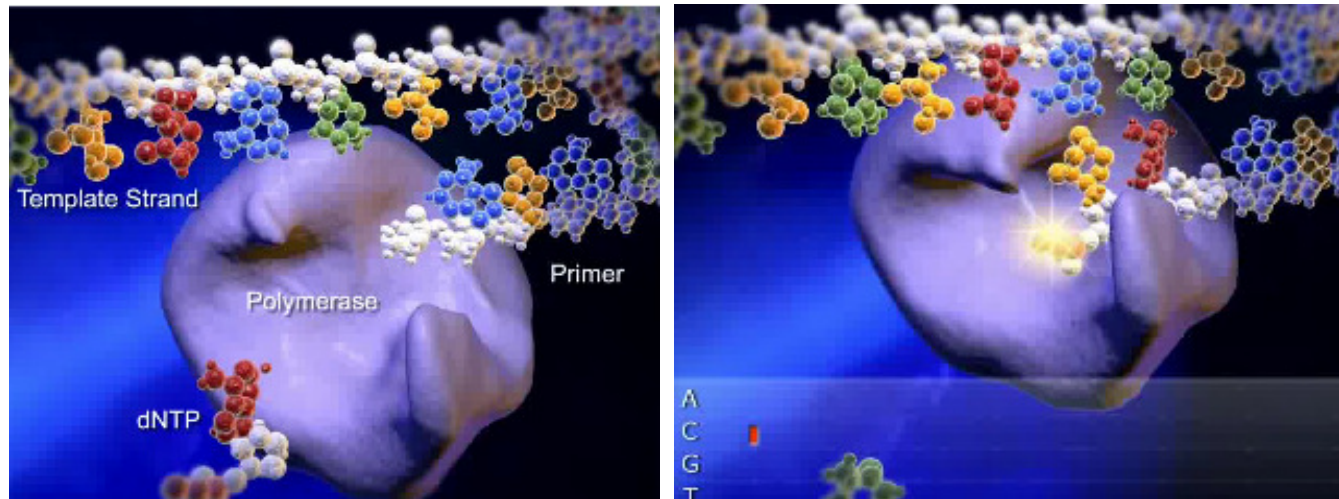
Heliscope – Helicos Biosciences



- Performances selon le constructeur :
 - 2 Gb par jour ; lectures jusqu'à 55 bases
 - Pour des longueurs de 20 bases, 78% des séquences passent les filtres de qualité
 - Problème des homopolymères (mais pas de déphasage) : virtual terminators

Life Technologies : FRETing

✓ FRET : Fluorescence Resonance Energy Transfer. Transfert d'énergie d'une polymérase à un nucléotide.



✓ Quand un nucléotide est incorporé, la proximité (entre le fluorophore donneur de la polymérase et le fluorophore accepteur du nucléotide) déclenche un signal FRET.

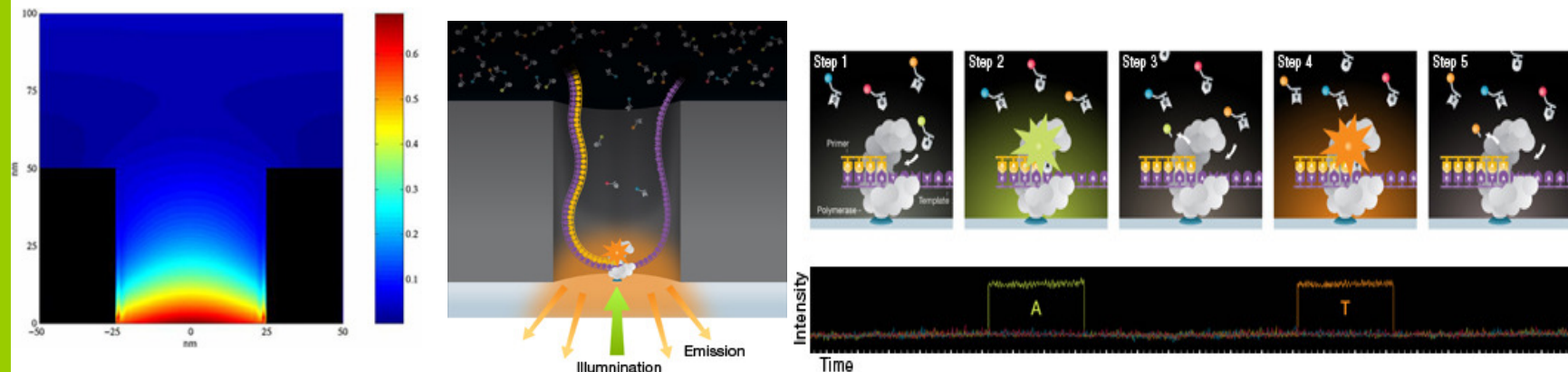
✓ La molécule d'ADN s'illumine et la couleur indique la base lue.

✓ L'objectif est de ~ 1 b / seconde

Pacific Biosciences : SMRT (Single Molecule Real Time)

La polymérase est attachée au fond d'un puits, d'une dizaine de nanomètres de diamètre et d'un volume de 20 zeptolitres (10^{-21} litres).

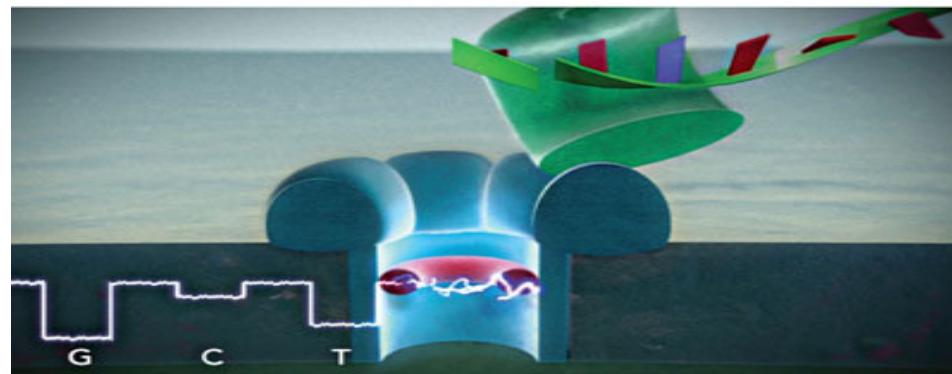
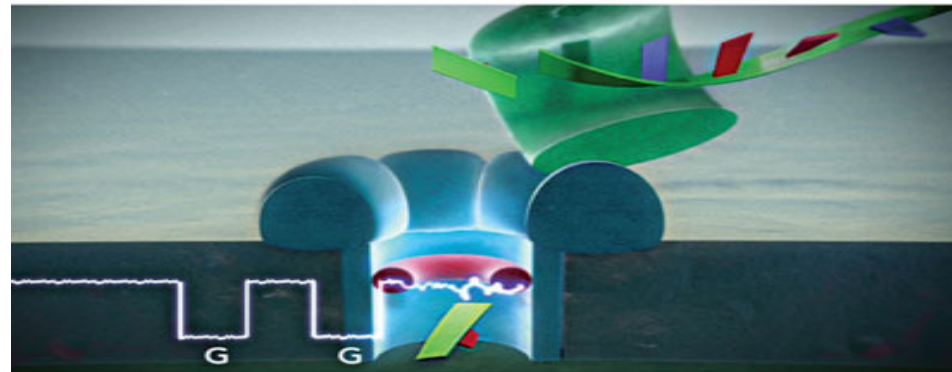
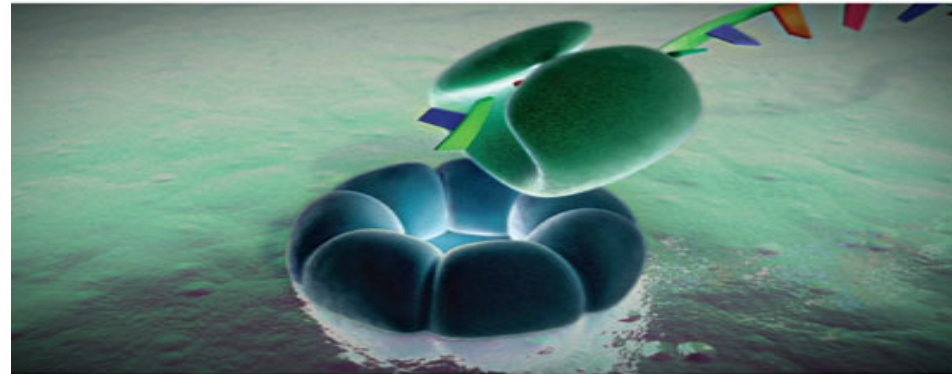
✓ Dans ce volume, l'activité d'un seul nucléotide peut être détecté parmi un bruit de fond de plusieurs centaines de nucléotides fluorescent.



✓ Vitesse de 10 bases / secondes et 1000 puits par puce (36Mb / heure), pour des lectures > 1000pb

✓ Prévision à 5 ans : 50 bases / secondes, avec des puces à 1M de puits, pour un débit d'environ 100Gb / heure

Illumina Oxford Nanopore





- *Director* : J. Weissenbach
- *Sequencing coordination* : P. Wincker
- *Production Sequencing*: J. Poulain
- *Finishing* : V. Barbe
- *NST development* : C. Cruaud, A. Alberti
- *Informatics*: V. Anthouard, M. Haquelle, F. Gavory, A. Couloux, C. Scarpelli
- *Bio-informatics* : J.M. Aury, G. Samson, O. Rogier, C. Battail, F. Artiguenave
- *Mutation Platform* : G. Gyapay, B. Noel, V. Meyer, M. Wessner