Séquençage Haut-Débit sur GeT

Olivier Bouchez, GeT-PlaGe
Responsable Séquençage Haut-débit

olivier.bouchez@toulouse.inra.fr

IBiSA

Commission nationale des outils Collectifs

GeT Génome et Transcriptome

Localisation des séquenceurs

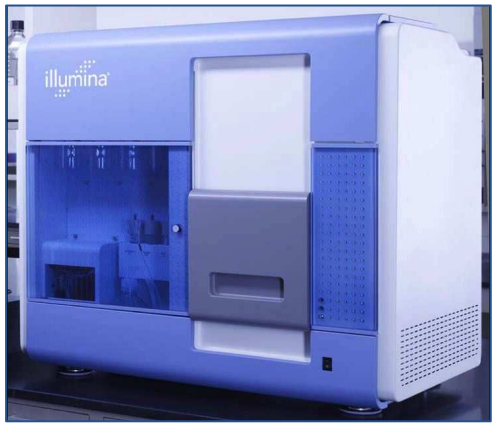
Plateforme Génomique
INRA Auzeville

GeT
Génome et Transcriptome

Séquenceurs 2ème génération

Séquenceurs de 2ème génération :

- Genome Analyser IIx (Solexa/Illumina)
- **HiSeq2000 (Illumina)**
- SOLiD (Applied Biosystem)
- HeliScope (Helicos BioScience Corporation)
- **GSFLX (454/Roche)**



GeT Génome et Transcriptome

Pourquoi deux séquenceurs HD ?

Roche 454 GS FLX

Spécificités :
-Lectures « longues » (450 pb)
-1,2 millions de séquences/run
-Pyroséquençage

Applications principales :
-Séquençage *de novo* (BACs, fosmides…)
-Re-séquençage (Régions ciblées..)
-Métagénomique (PCR 16S)

Illumina HiSeq 2000

Spécificités :
-Lectures « courtes » (100 pb)
-1 milliard de séquences/run
-Sequencing by synthesis

Applications principales :
-Analyses Transcriptomiques (=> génome de référence)
-Re-séquençage (Génomes entiers)
-Métagénomique (ADNg total)
-Analyses épigénétiques (Génome entier, ChipSeq)

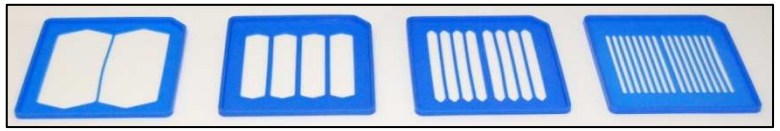
Analyses transcriptomiques (sans génome de référence)
Séquençage de novo (Génomes entiers)

Roche 454 GS FLX

Séquençage sur le Roche 454 GS FLX

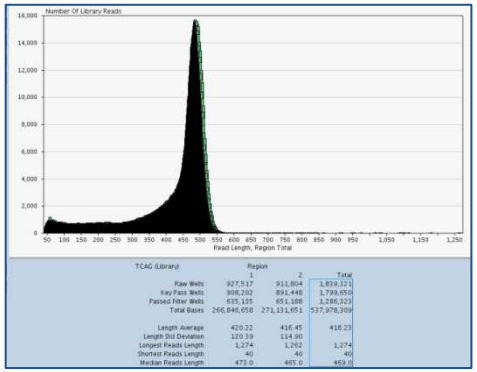
Spécifications :

- **1.2 M** de séquences/run
- Longueur moyenne des séquences : **450 pb**
- **500 Mb**/run
- Lames 2, 4, 8 et 16 régions
- Chimie **Titanium**
- Possibilité de réaliser des **multiplexages**



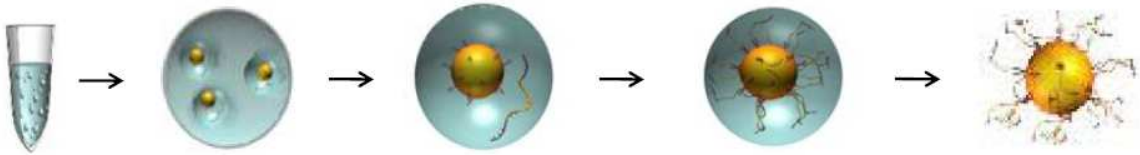
Applications:

- Séquençage ***de novo***
- **Reséquençage** de régions d'intérêt (par LR-PCR ou capture de séquences Nimblegen)
- **Transcriptomique**
- Analyses **Epigénétiques** (Etudes de Méthylation)
- Analyses **Métagénomiques**



Roche 454 GS FLX

Amplification clonale en émulsion



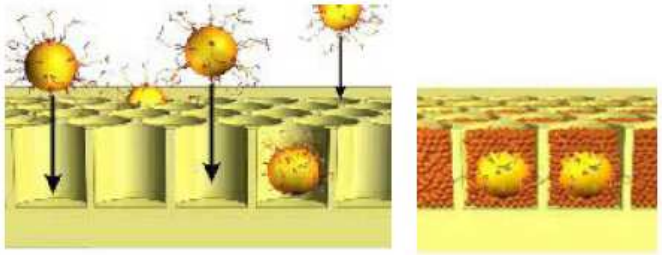
Mélane des fragments sstDNA à un excés de bille

Création de microréacteur avec les billes et les réactifs PCR

Amplification clonale à l'intérieur des microréacteurs

Enrichissement des billes "positives"

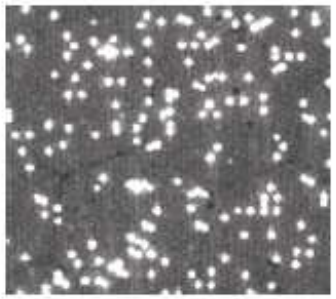
Dépot des billes portant la banque d'ADN simple brin dans la PicoTiterPlate

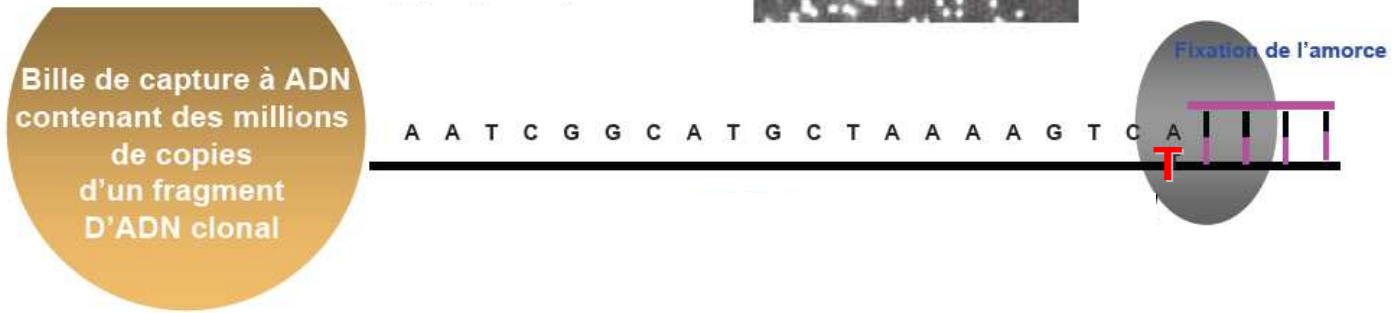


- Dépot dans chaque puits d'une seule bille portant l'ADN amplifié clonalement

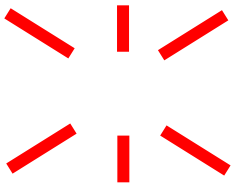
Roche 454 GS FLX

- Injection consécutive des nucléotides (A puis T, puis G, puis C)
- L'incorporation du nucléotide injecté induit une émission lumineuse





- Identification des puits donnant un signal (et quantification)



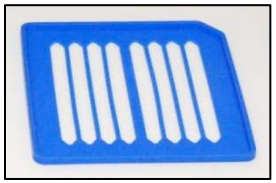
Roche 454 GS FLX

Depuis février 2009 :

- 60 runs réalisés
- 30 équipes différentes

Tarifs :
~10000 € /run 2 régions
~2000 € 1/8 run
~170 € /librairie

Depuis octobre 2009 : mise en place de runs de séquençage commun où différentes équipes peuvent passer sur le même run (8 régions).
Prévision : **1 run/mois => 6 runs réalisés**



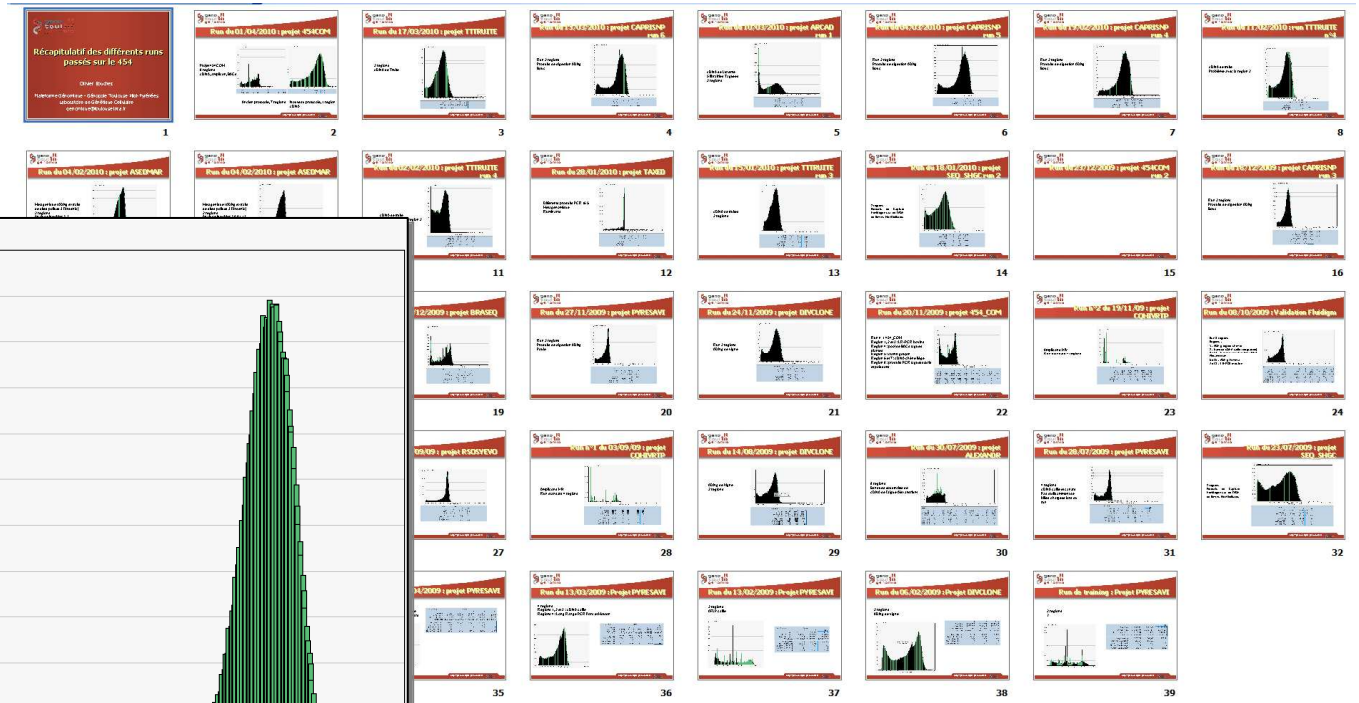
En 2009-2010 : Le Roche en production

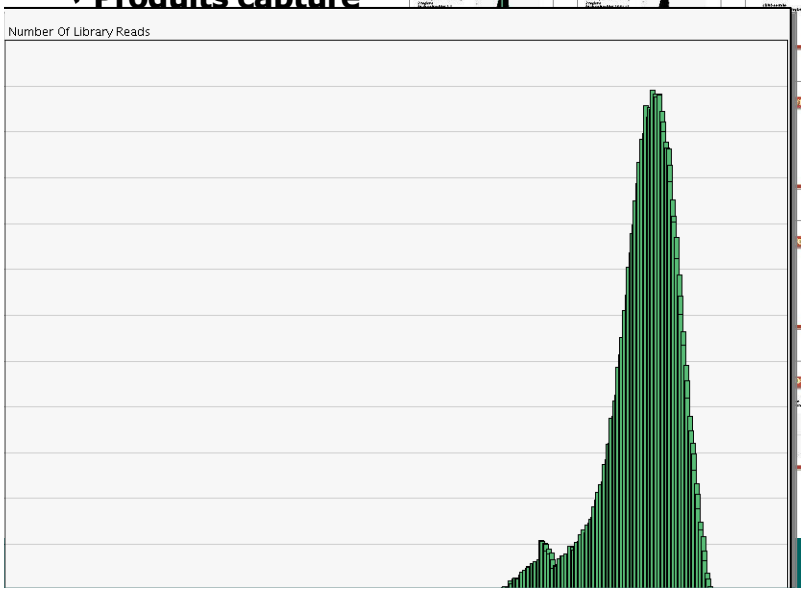
- Séquençage 454 : 60 runs dont 6 runs communs

✓**ADNg**

✓**cDNA**

✓**Produits capture**





Illumina HiSeq2000

GeT Génome et Transcriptome

HiSeq2000 (Illumina)

- 3ème machine installée en France, en cours de validation par le fournisseur

- Technologie « sequencing by synthesis »

- Possibilité de multiplexage

- Applications :
  - séquençage de novo
  - re-séquençage génomes entiers/régions candidates
  - analyses épigénétiques
  - ChiP-seq
  - analyses transcriptomiques
  - identification et quantification de petits ARN

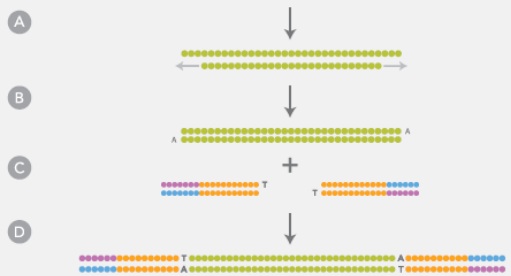
- Single read :
  - 100 Gb/run
  - Read size : 100 pb
  - 1 run = 4 jours

- Paired-end reads :
  - 200 Gb/run
  - Read size : 2x100 pb
  - 1 run = 8 jours

GeT
Génome et Transcriptome

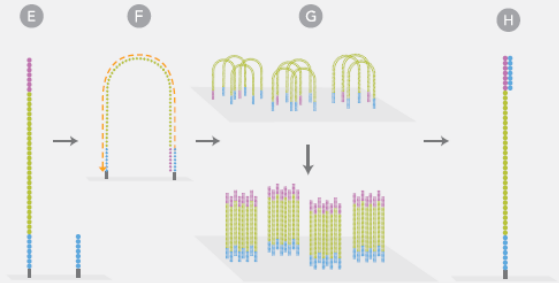
HiSeq2000 (Illumina)

1 LIBRARY PREPARATION



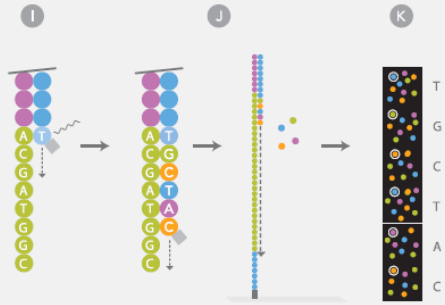
- A Fragment DNA
- B Repair ends Add A overhang
- C Ligate adapters
- D Select ligated DNA

2 CLUSTER GENERATION



- E Attach DNA to flow cell
- F Perform bridge amplification
- G Generate clusters
- H Anneal sequencing primer

3 SEQUENCING



- I Extend first base, read, and deblock
- J Repeat step above to extend strand
- K Generate base calls

GeT
Génome et Transcriptome

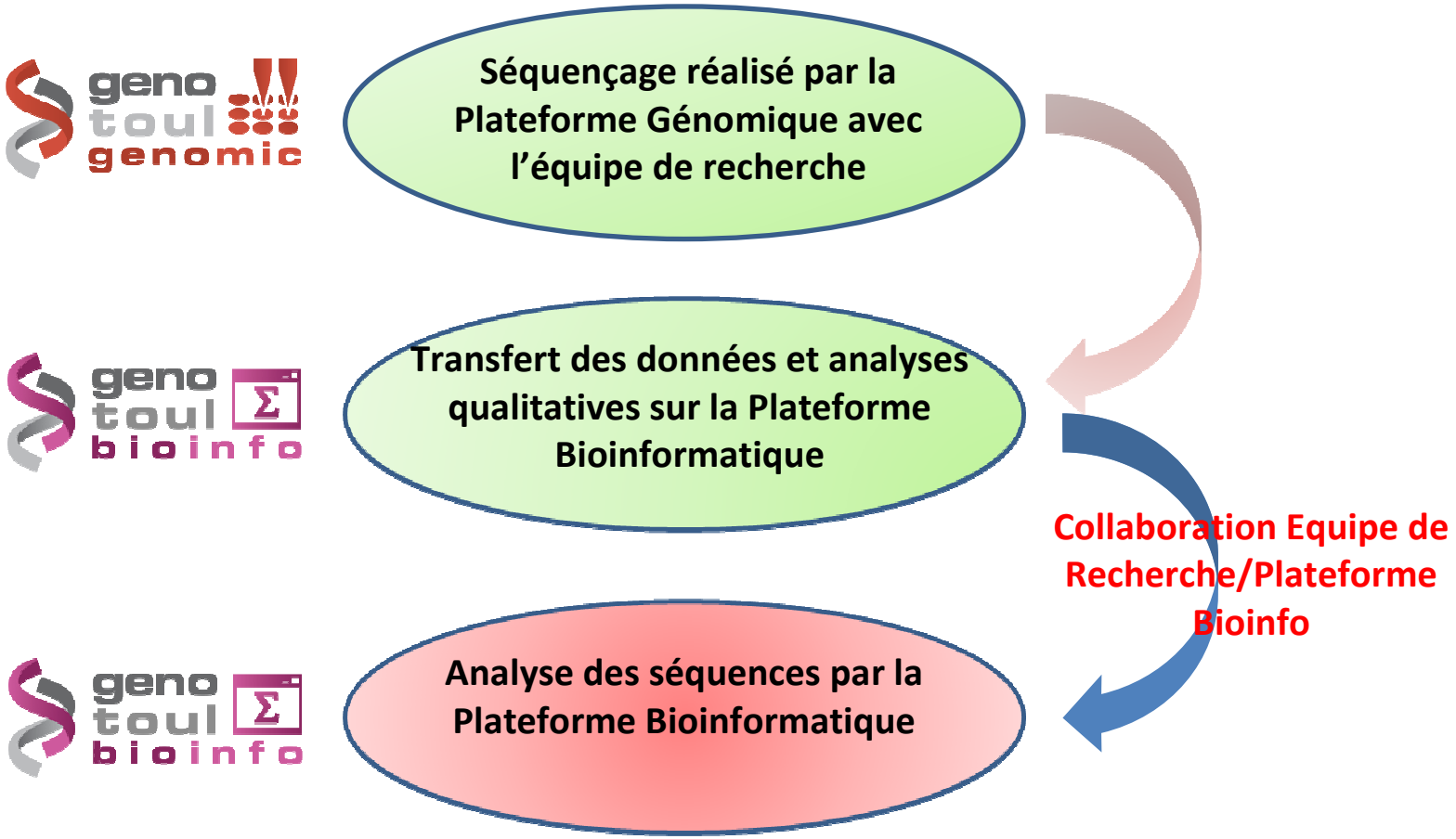
HiSeq2000 (Illumina)

- ✓ En cours de validation
- ✓ Formation en novembre, puis début de la phase pilote
- ✓ Fin de la validation ADNg : février 2011
- ✓ Début de validation mRNA-seq : février 2011
- ✓ Ouverture machine prévue vers juillet 2011

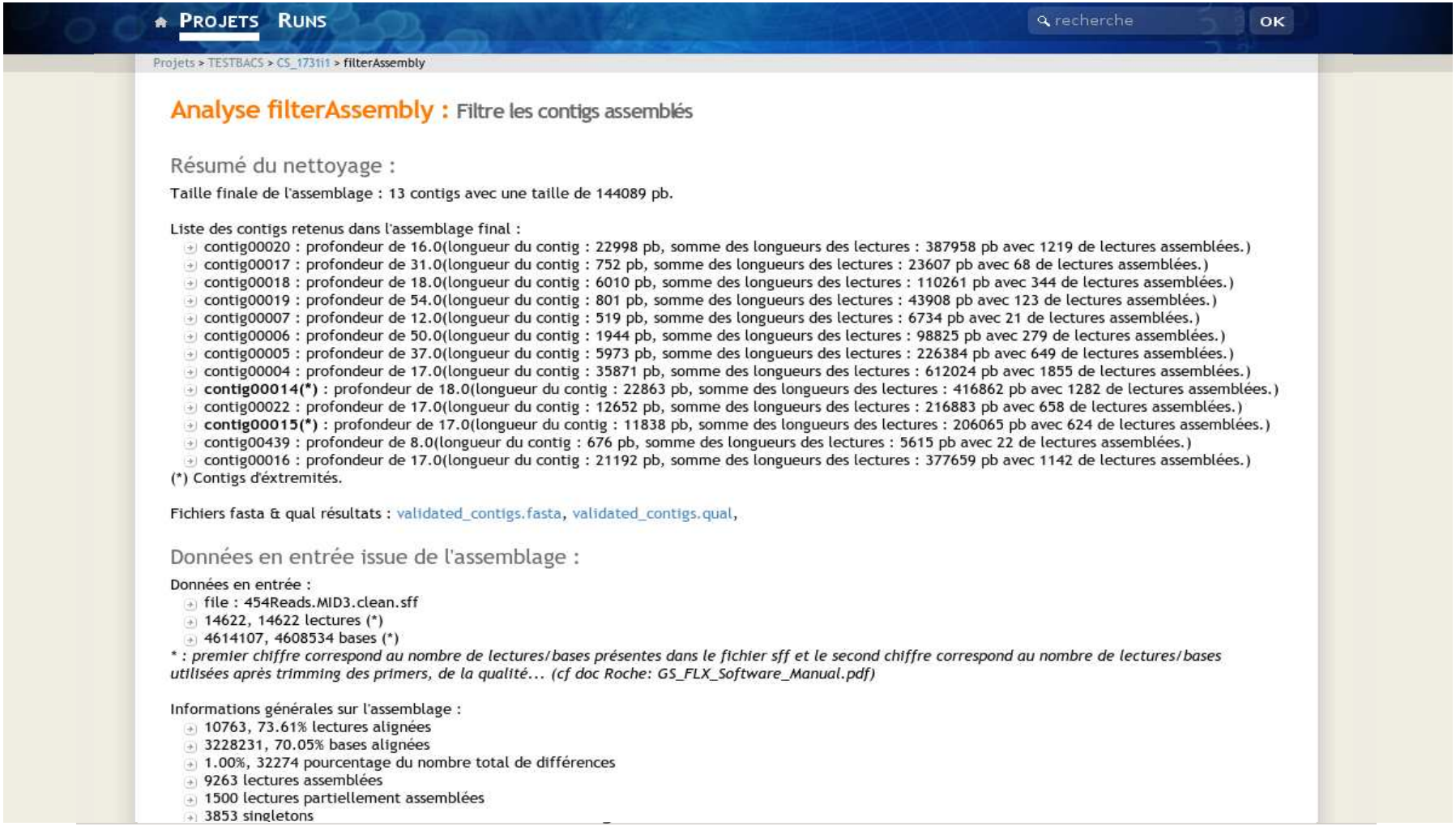
- ➢ Tarifs : ~10000 € / flowcell (=1/2 run) pour 1 librairie
  1 librairie = > 400 €
- ➢ Nouveaux protocoles bientôt disponibles

GeT Génome et Transcriptome

Séquençage THD : Réalisation et analyse



Pipeline de traitement des données



PROJETS RUNS recherche OK

Projets > TESTBACS > CS_1731i1 > filterAssembly

Analyse filterAssembly : Filtre les contigs assemblés

Résumé du nettoyage :

Taille finale de l'assemblage : 13 contigs avec une taille de 144089 pb.

Liste des contigs retenus dans l'assemblage final :

- contig00020 : profondeur de 16.0(longueur du contig : 22998 pb, somme des longueurs des lectures : 387958 pb avec 1219 de lectures assemblées.)
- contig00017 : profondeur de 31.0(longueur du contig : 752 pb, somme des longueurs des lectures : 23607 pb avec 68 de lectures assemblées.)
- contig00018 : profondeur de 18.0(longueur du contig : 6010 pb, somme des longueurs des lectures : 110261 pb avec 344 de lectures assemblées.)
- contig00019 : profondeur de 54.0(longueur du contig : 801 pb, somme des longueurs des lectures : 43908 pb avec 123 de lectures assemblées.)
- contig00007 : profondeur de 12.0(longueur du contig : 519 pb, somme des longueurs des lectures : 6734 pb avec 21 de lectures assemblées.)
- contig00006 : profondeur de 50.0(longueur du contig : 1944 pb, somme des longueurs des lectures : 98825 pb avec 279 de lectures assemblées.)
- contig00005 : profondeur de 37.0(longueur du contig : 5973 pb, somme des longueurs des lectures : 226384 pb avec 649 de lectures assemblées.)
- contig00004 : profondeur de 17.0(longueur du contig : 35871 pb, somme des longueurs des lectures : 612024 pb avec 1855 de lectures assemblées.)
- **contig00014(*)** : profondeur de 18.0(longueur du contig : 22863 pb, somme des longueurs des lectures : 416862 pb avec 1282 de lectures assemblées.)
- contig00022 : profondeur de 17.0(longueur du contig : 12652 pb, somme des longueurs des lectures : 216883 pb avec 658 de lectures assemblées.)
- **contig00015(*)** : profondeur de 17.0(longueur du contig : 11838 pb, somme des longueurs des lectures : 206065 pb avec 624 de lectures assemblées.)
- contig00439 : profondeur de 8.0(longueur du contig : 676 pb, somme des longueurs des lectures : 5615 pb avec 22 de lectures assemblées.)
- contig00016 : profondeur de 17.0(longueur du contig : 21192 pb, somme des longueurs des lectures : 377659 pb avec 1142 de lectures assemblées.)

(*) Contigs d'extremités.

Fichiers fasta & qual résultats : validated_contigs.fasta, validated_contigs.qual,

Données en entrée issue de l'assemblage :

Données en entrée :

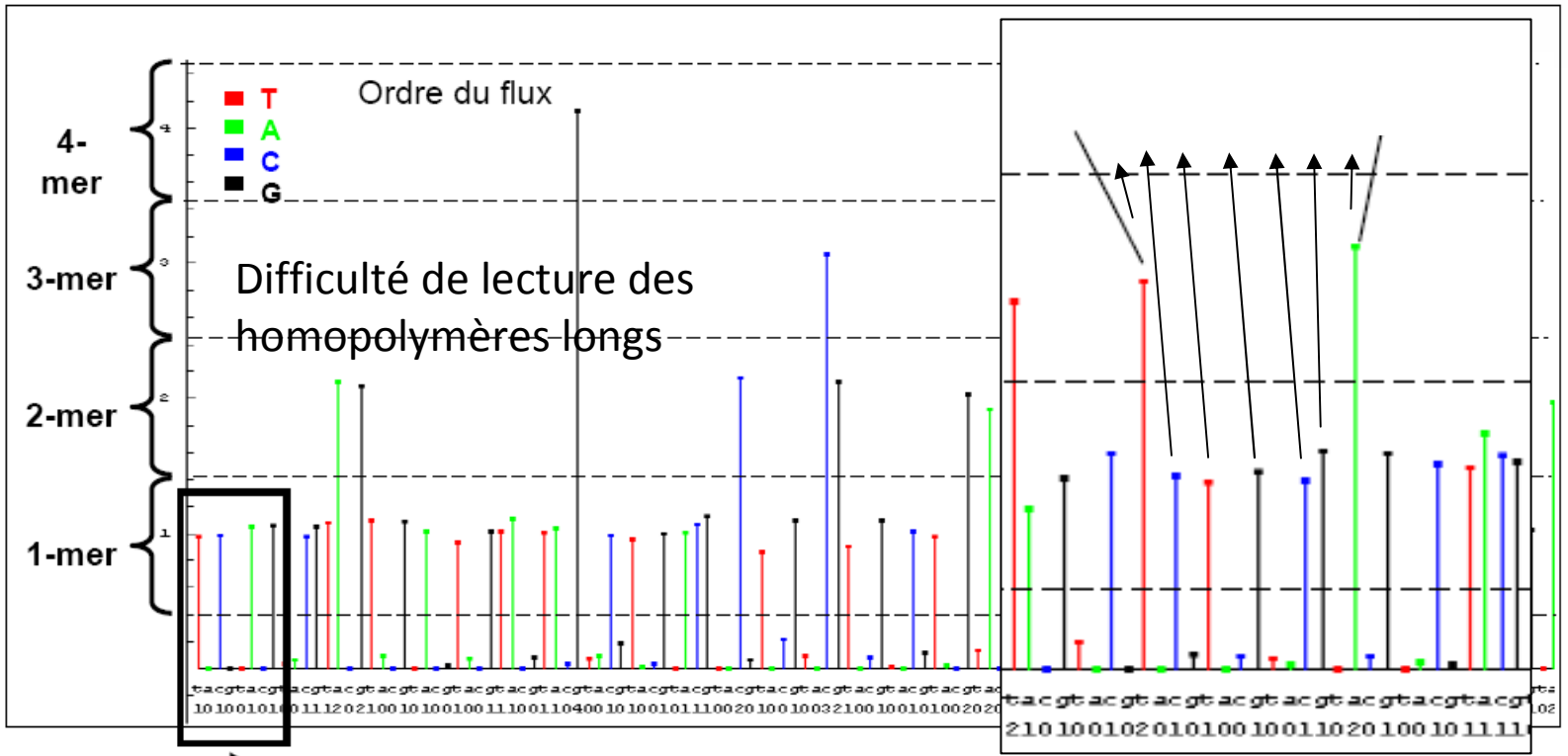
- file : 454Reads.MID3.clean.sff
- 14622, 14622 lectures (*)
- 4614107, 4608534 bases (*)

** : premier chiffre correspond au nombre de lectures/bases présentes dans le fichier sff et le second chiffre correspond au nombre de lectures/bases utilisées après trimming des primers, de la qualité... (cf doc Roche: GS_FLX_Software_Manual.pdf)*

Informations générales sur l'assemblage :

- 10763, 73.61% lectures alignées
- 3228231, 70.05% bases alignées
- 1.00%, 32274 pourcentage du nombre total de différences
- 9263 lectures assemblées
- 1500 lectures partiellement assemblées
- 3853 singletons

Roche 454 GS FLX



Séquence clé dans l'échantillon = AGTC pour calibration du signal