Formation RNA-Seq

Formateurs

- **Bioinformatique :**
  - **Sarah Maman**
  - **Céline Noirot**
  - **Olivier Rué**
  - **Matthias Zytnicki**
- **Biostatistique :**
  - **Ignaccio Gonzales**
  - **Annick Moisan**
  - **Nathalie Villa-Vialaneix**

Organisation

- Formations de 3,5 jours
- Présentations seront entre-coupées de TP
- Horaires
  - Matin : 9h30 - 12h30
  - Après-midi : 14h00 – 17h

Tour de table

- Vous ? Votre labo ? Vos données ?
- Avez vous déjà traité des NGS ?
- Avez vous déjà utilisé Galaxy et R ?
- Vos attentes ?

Qu'allez vous apprendre …

- A l'issue de 3 jours de formation **vous saurez :**
  - Utiliser un environnement **Galaxy** pour **quantifier** et **découvrir de nouveaux transcrits**,
  - Utiliser **Rstudio** pour détecter le gène DE correspondant à un **plan d'expérience simple**.

- Vous **ne saurez pas** :
  - Étudier un transcriptome **sans génome** de référence,
  - (Traiter bioinformatiquement de **nombreuses librairies.**)
  - Détecter les gènes DE correspondant à un plan d'expérience **complexe**.

INRA
SCIENCE & IMPACT

TRAITEMENT BIOINFORMATIQUE DE DONNÉES RNA-Seq

Plan

- Mardi - Matin (9h30 - 12h30)
  - Introduction au RNAseq (Biologie & Protocole)
  - Présentation Galaxy et prise en main
- Mardi - Après-midi(14h – 17h)
  - Vérification de la qualité
  - Algorithmes d'alignement
- Mercredi - Matin (9h30 - 12h30)
  - Les formats de fichier d'alignement
  - Visualisation
  - Assemblage de transcrits
- Mercredi - Après-midi (14h00 - 17h)
  - Quantification de gènes
  - Quantification de transcrits
  - Créer vos workflow RNAseq

_01

Rappels biologiques

Un peu de vocabulaire

- **Transcriptome** : Ensemble des transcrits d'un organisme
- **RNAseq de novo** : Etude du transcriptome sans génome de référence.
- **Read** : Lecture
- **Fragment** : Paire de lecture

Rappels biologiques

Qu’est-ce qu’un gène ?

Rappels biologiques

Qu'est-ce qu'un gène ?

- **Gène :** unité fonctionnelle de l'ADN qui contient les instructions nécessaires à la création d'un produit fonctionnel

- **Promoteur :** zone de fixation des ribosomes
- **TSS :** site de départ de transcription
- **Exon :** région codante de l'ARNm inclus dans le transcrit
- **Intron :** région non codante

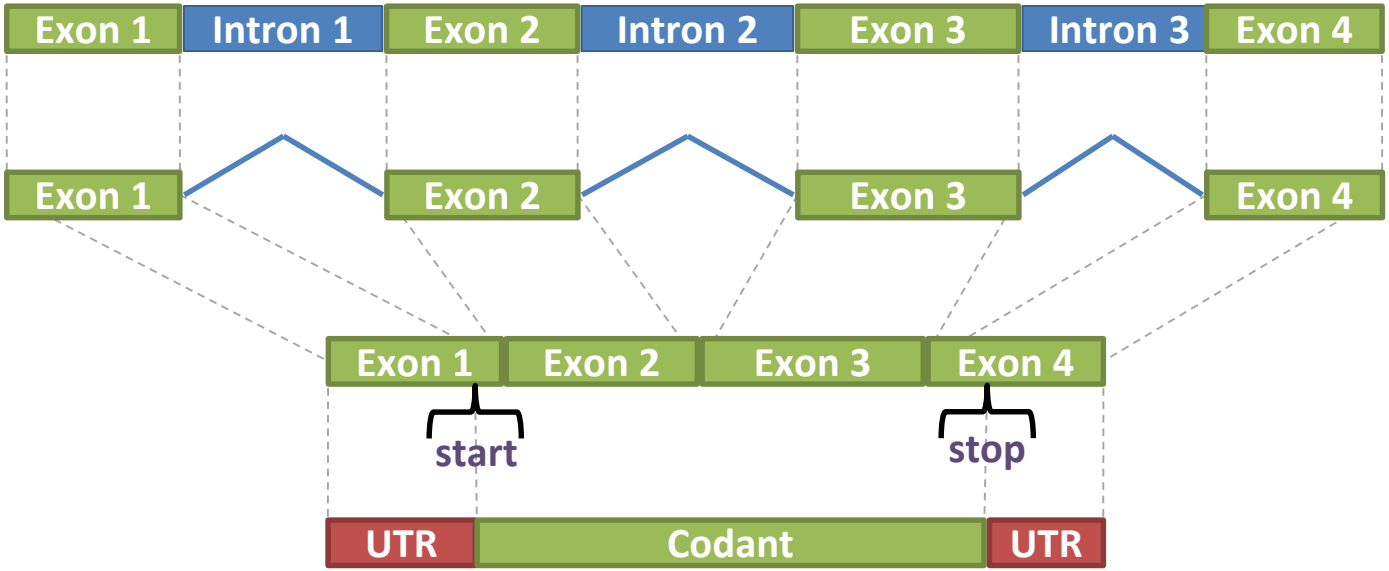
Rappels biologiques

Qu’est-ce qu’un transcrit?

Rappels biologiques

Qu'est-ce qu'un transcrit ?

- **Epissage :** Excision des introns avant traduction

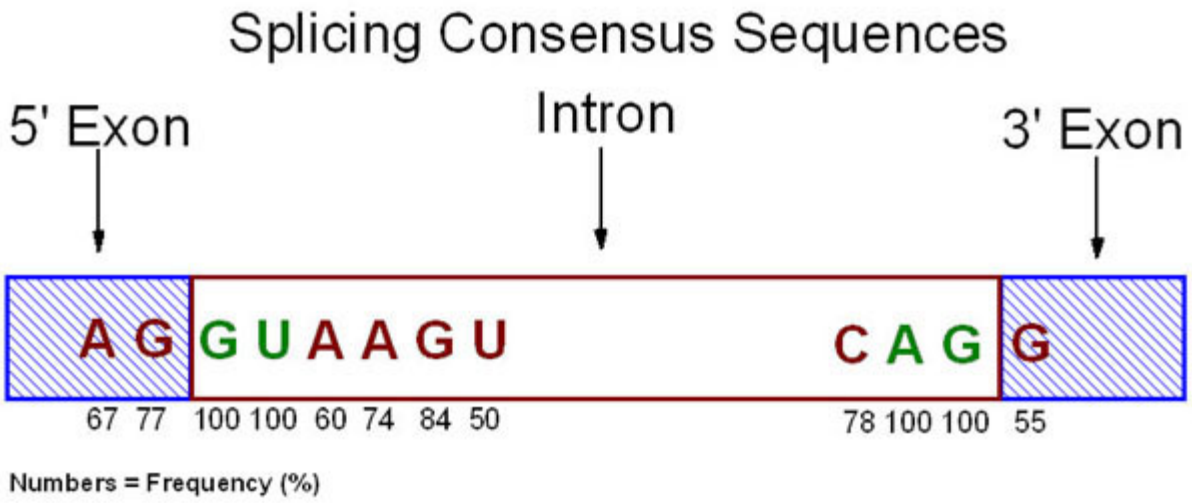


- **Transcrit :** portion d'ADN transcrite en molécule d'ARN
- **UTR :** région transcrite mais pas traduite

Rappels biologiques

Qu'est-ce qu'un site d'épissage?

- **Site** d'**épissage canonique :**
  - plus de **99%** de **GT** et **AG** comme sites **donneurs** et **accepteurs**



http://themedicalbiochemistrypage.org/rna.php

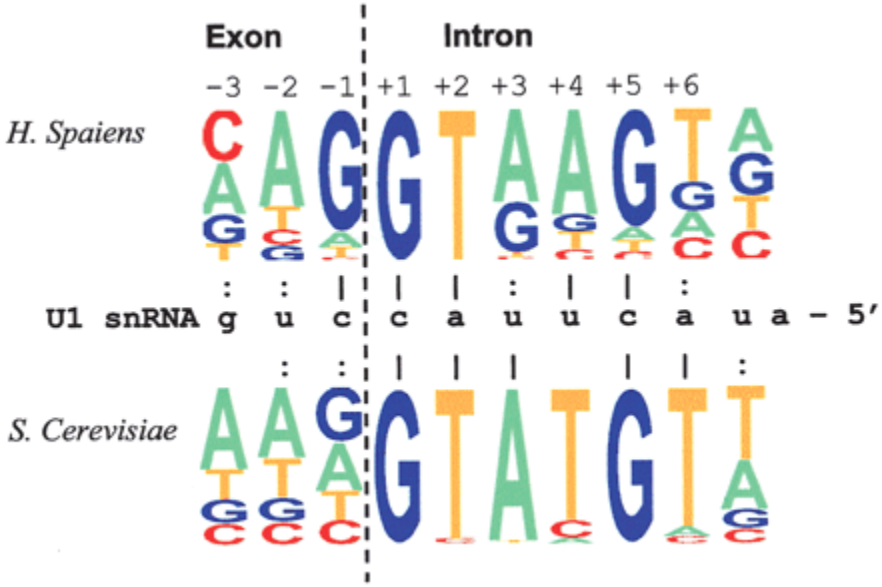
Rappels biologiques

Qu'est-ce qu'un site d'épissage?

- **Site** d'**épissage non-canonique :**
  - **GC-AG** ou **AT-AC** comme sites **donneurs** et **accepteurs**

- **Mammifère :**
  - 0.69% GC-AG
  - 0.05% AT-AC

- **Autre exemple :**



http://rnajournal.cshlp.org/content/10/5/828.full

Rappels biologiques

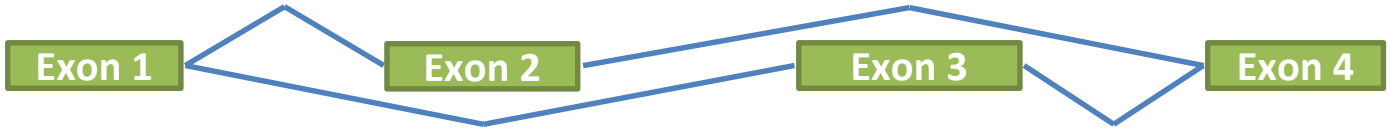
Epissage alternatif et isoformes

- Excision d'exon

- Rétention d'intron

- TSS alternatif

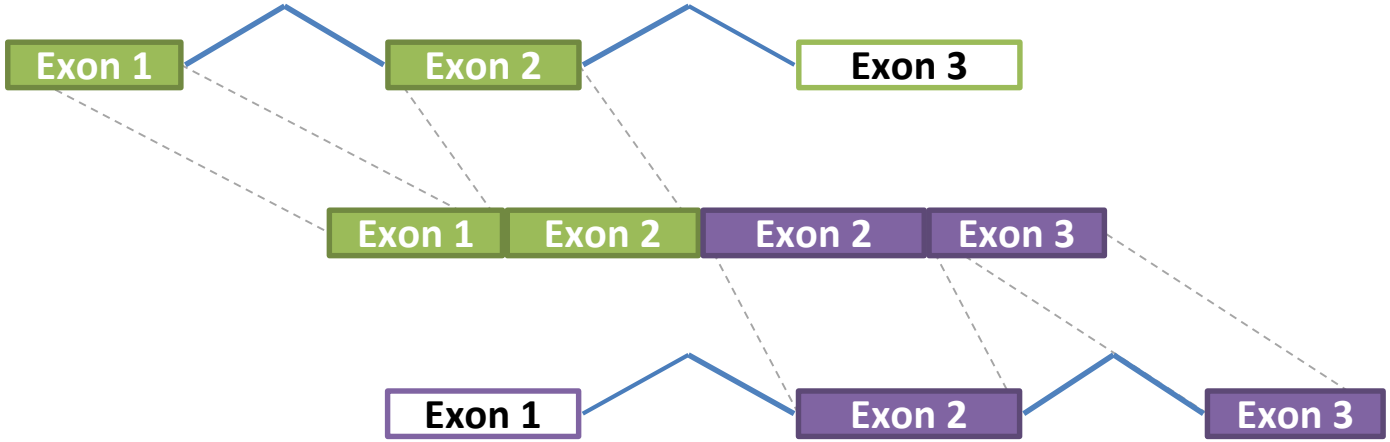
- Exons exclusifs



Rappels biologiques

Et plus encore ?

- **Fusion** de **gènes** ou **Trans-épissage**

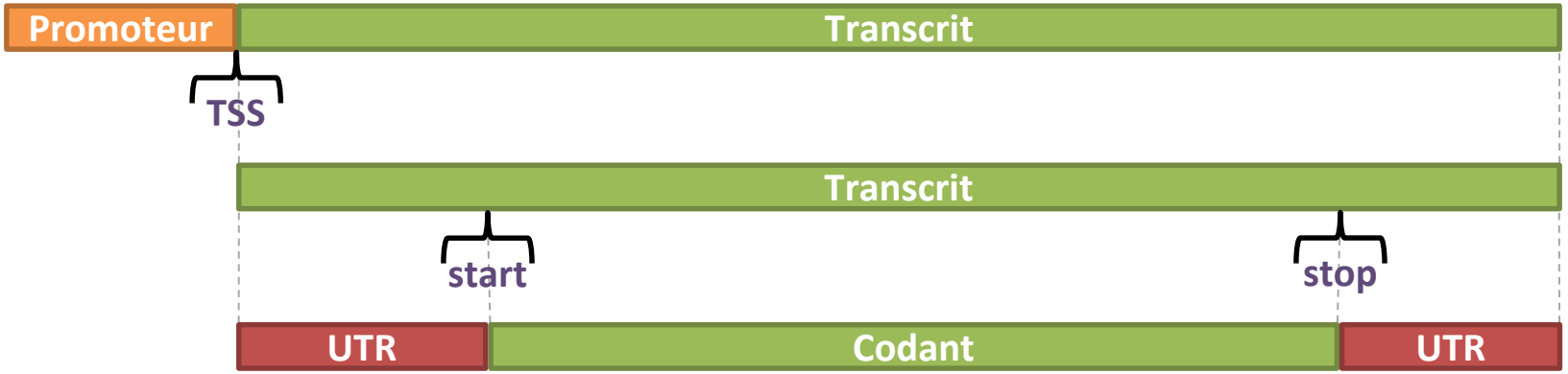


- **Chimère biologique**

Rappels biologiques

Gène procaryote / gène eucaryote

- Pas d'intron chez les procaryotes



Le RNA-Seq

Modes d'étude du transcriptome

- EST
- rt-PCRq
- puce d'expression
- tiling array

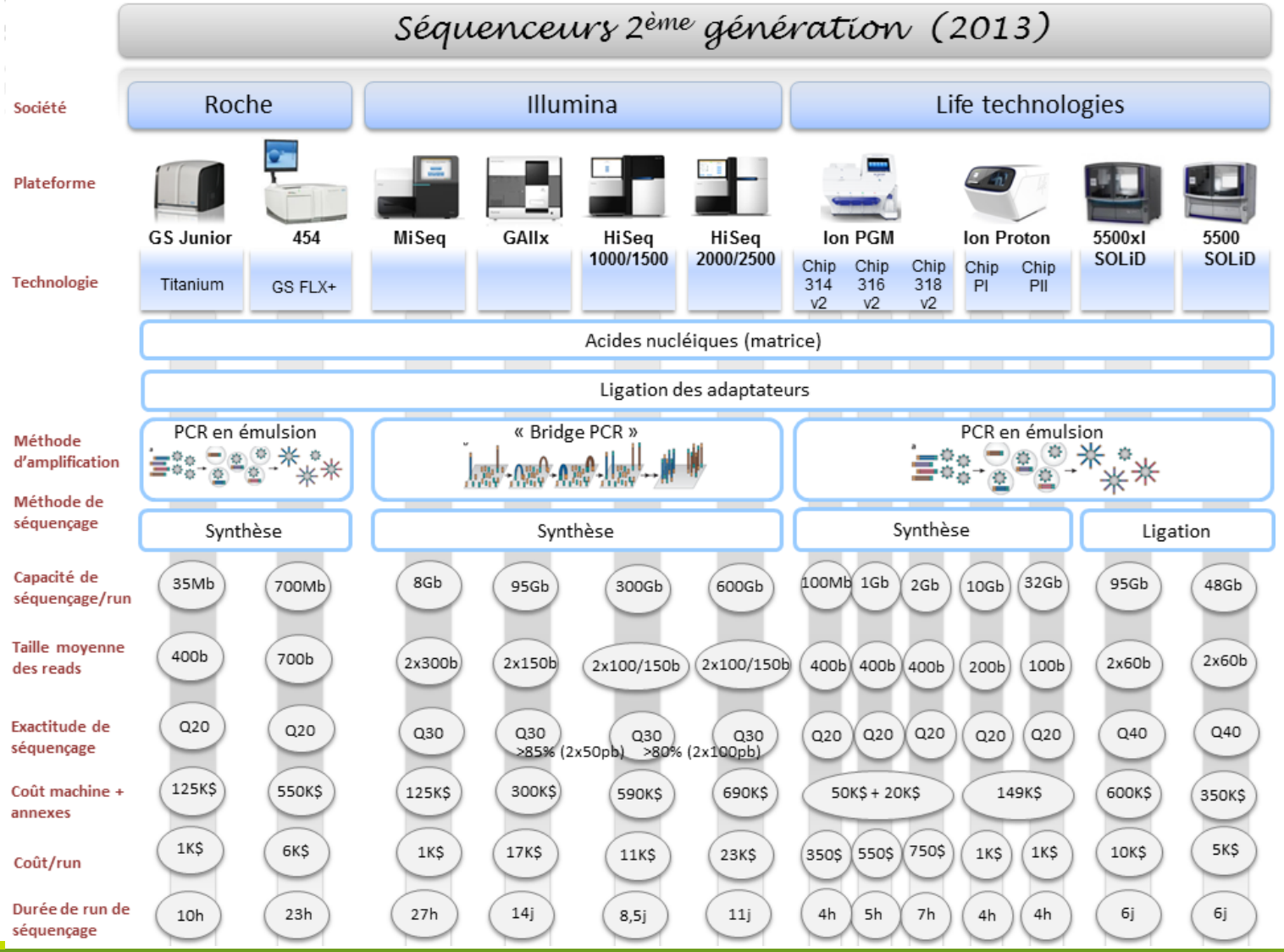
- RNA-Seq

Quelles sont les principales différences ?

Modes d'étude du transcriptome

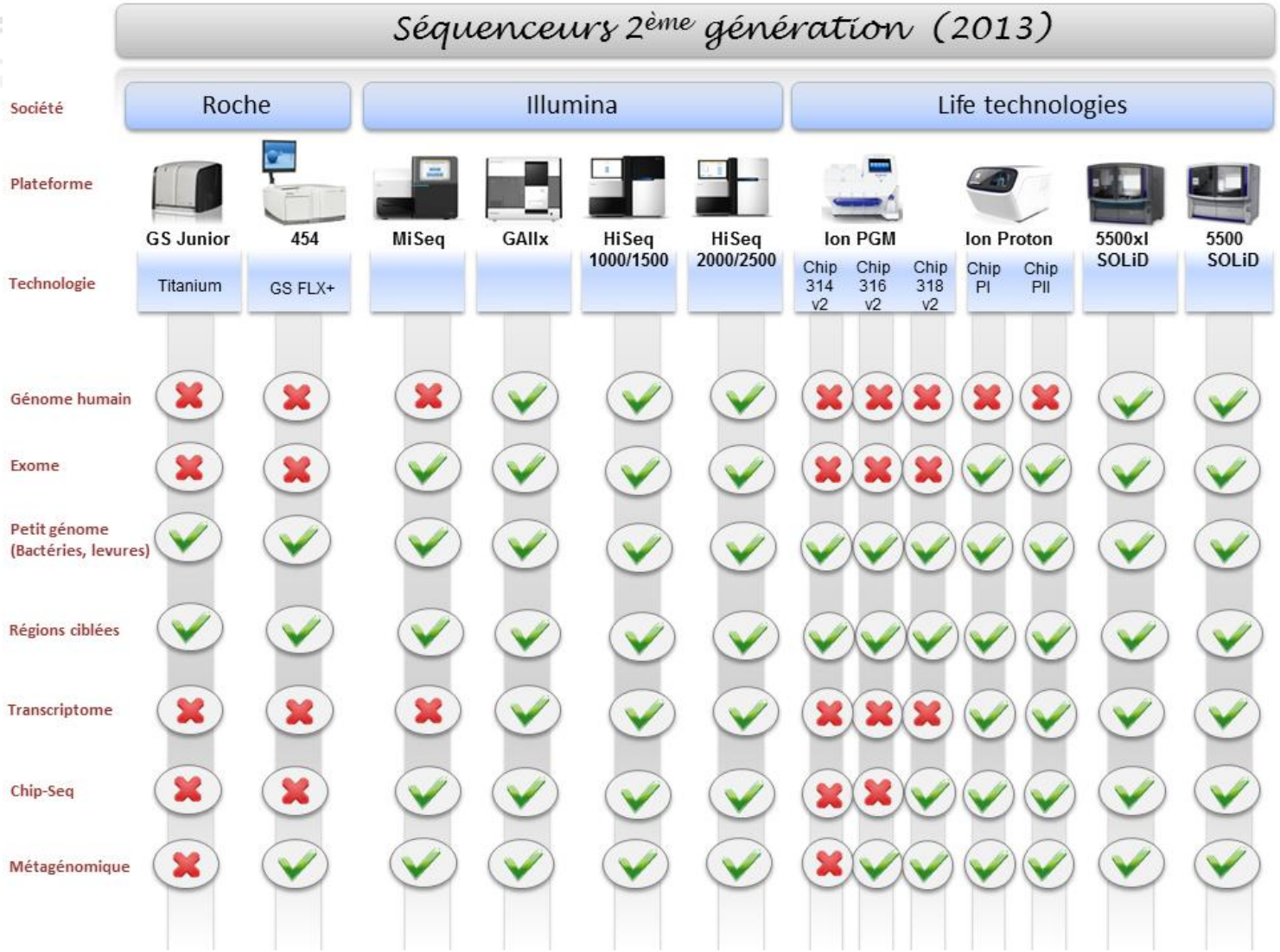
- Pas besoin d'avoir de connaissance sur la séquence
- Spécificité de ce que l'on mesure
- Augmente l'échelle de mesure
- Quantification directe
- Très bonne reproductibilité
- Différents niveau d'étude : gènes, transcrits, spécificité allèlique, variant de structure
- Découverte de nouveaux : transcrits, isoformes, (ncRNA), structures (fusion...)
- Détection possible of SNPs, ...

Les séquenceurs



http://www.biorigami.com/?p=5492

Les séquenceurs



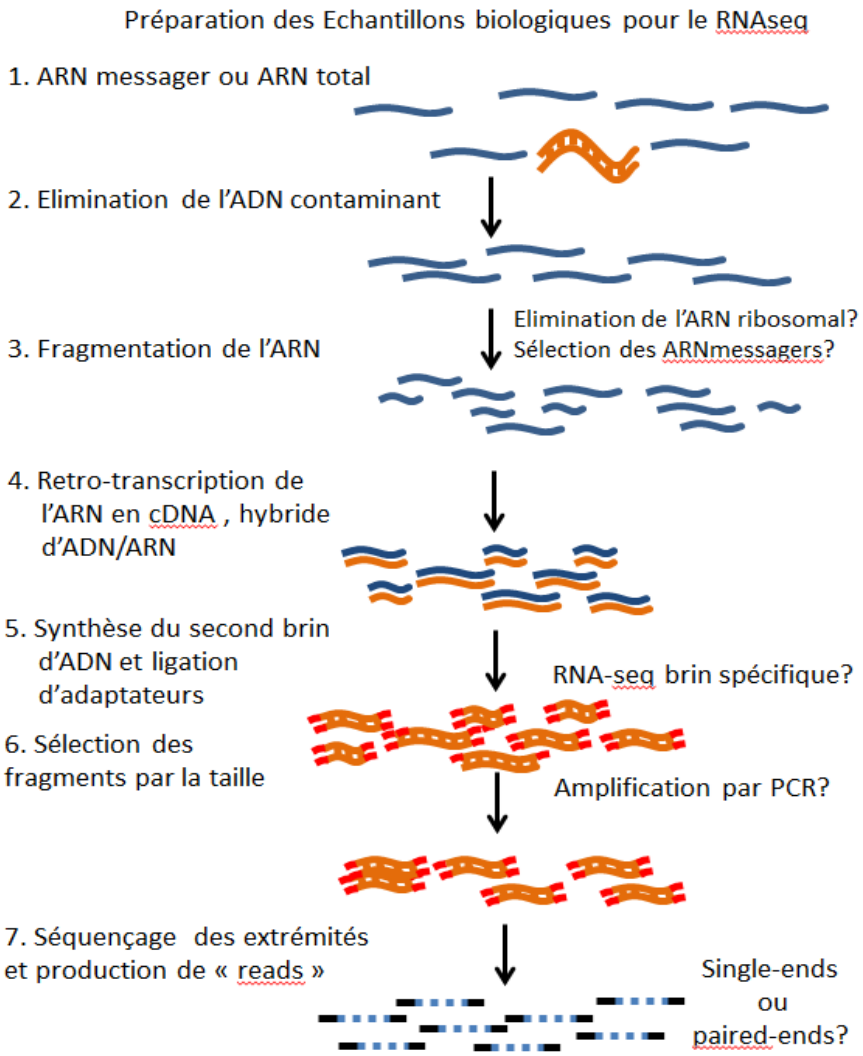
http://www.biorigami.com/?p=5492

Illumina sequencing vocabulary

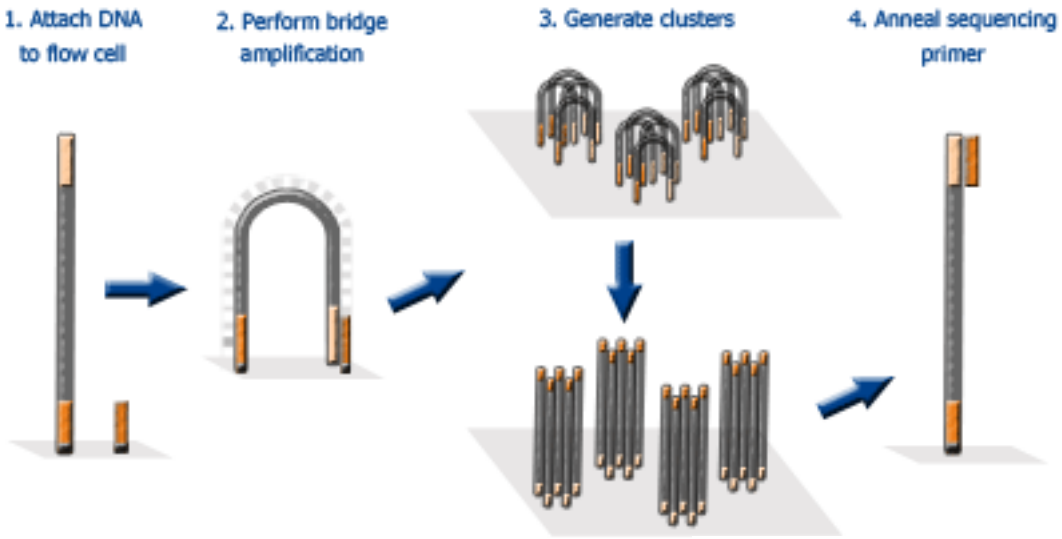
- **Flowcell : 1 plaque (en général 1 run)**
- Lane : ligne de séquençage
- 1 Flowcell : 8 Lane

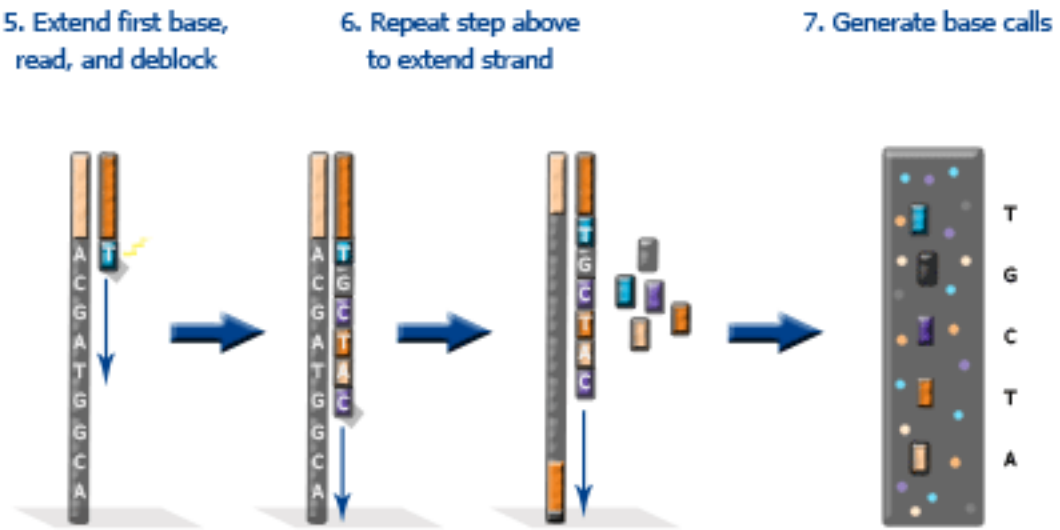
- 1 flowcell Hiseq 2500 : 2 Milliard de reads single ou 4 Milliard de reads paired.
- Hiseq 2000 / Hiseq 2500 : séquençage possible de 2 flowcells en parallèle.

Le protocole RNAseq



Séquençage illumina





Quels choix quand on fait du RNA-Seq ?

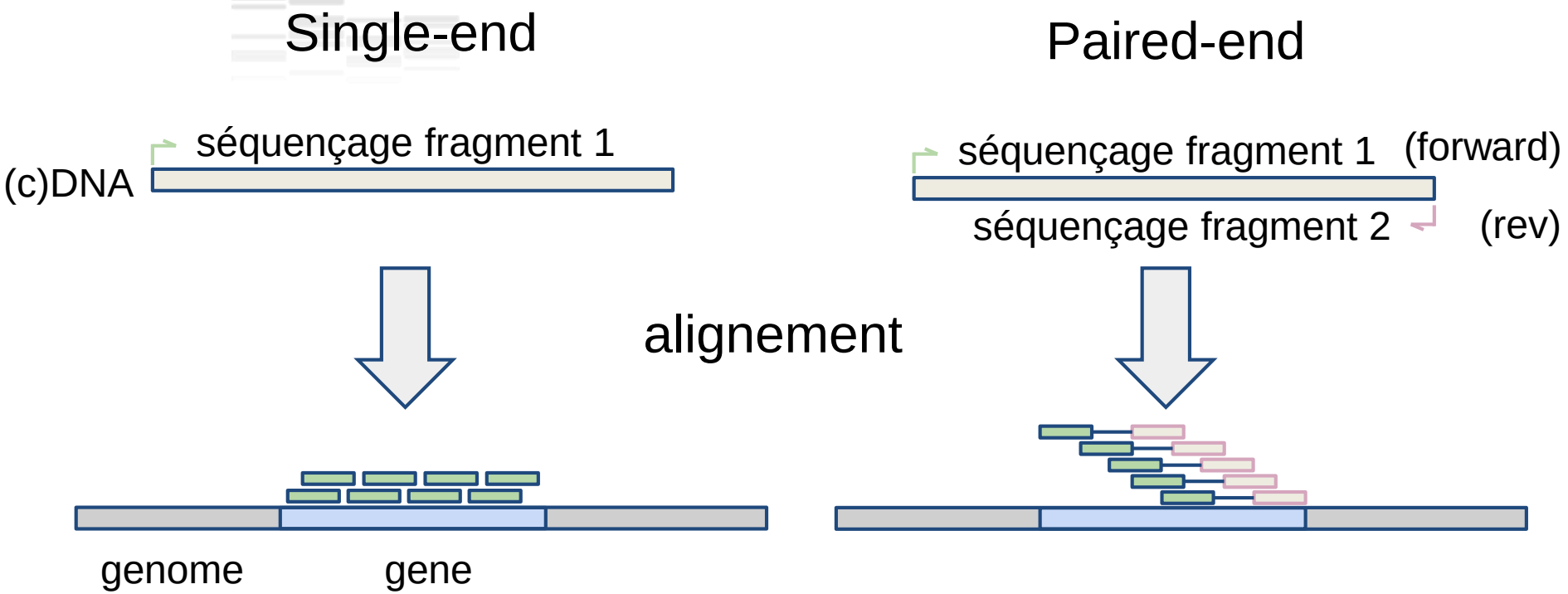
- **Déplétion / enrichissement**
- **Paired-end / single-end**
- Séquençage en tenant compte du **sens du brin**
- **Nombre de séquence / de réplicats**
- **Multiplexage**

Déplétion / Enrichissement

- Déplétion :
  - Suppression des rRNA
- Enrichissement polyA :
  - Pas de transcrits sans queue PolyA ou partiellement dégradés
- Résultats semblables d'après :

Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling, BMC Genomics , 2014

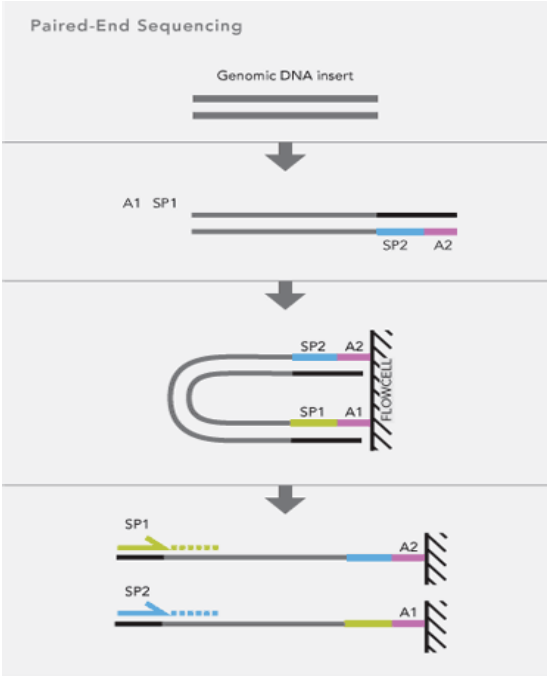
Single-end *vs* Paired-end



- La taille des cDNA détermine la taille d'insert (p. ex. 200-500 pb).
- Les fragments sont habituellement en Forward-Reverse.

Paired-end

- Protocole différent (Adaptateurs spécifiques)
- Améliore le mapping
- Aide a la détection de variant alternatif
- Plus généralement aide à la détection de : variation structurale de génome (insertion/délétion), CNV, réarrangement génomique



Adapter (A1 and A2) with sequencing primer sites (SP1 and SP2) are ligated onto DNA fragments. Template clusters are formed on the flow cell by bridge amplification and then sequenced by synthesis from the paired primers sequentially.

L'intérêt des librairies brin spécifique

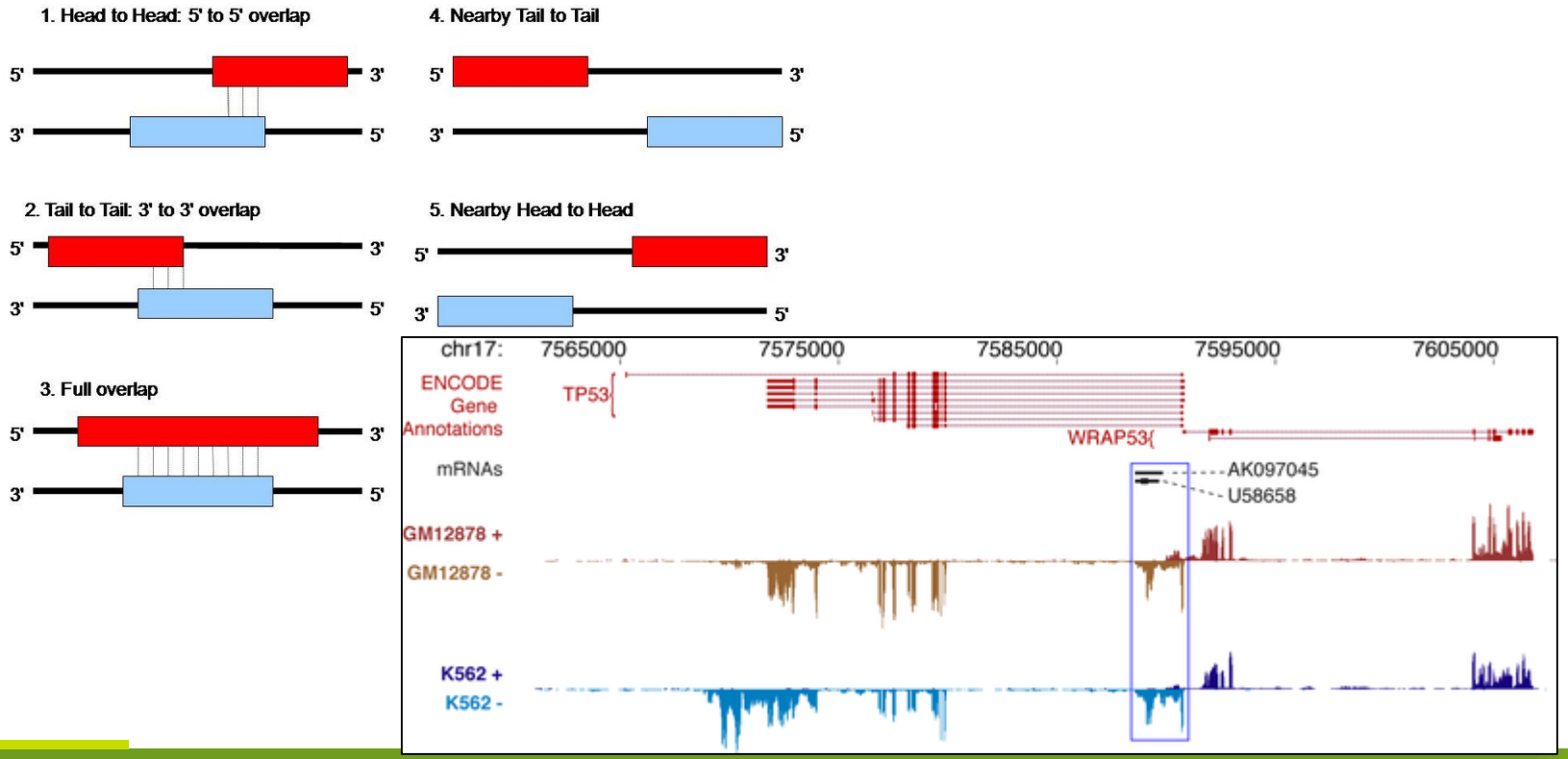
Nat Methods. 2010 Sep;7(9):709-15. Epub 2010 Aug 15.

Comprehensive comparative analysis of strand-specific RNA sequencing methods.

Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A.

Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts, USA.
jlevin@broadinstitute.org

Abstract



Profondeur / Répétitions ?

- Equilibre **profondeur /** nombre de **répétitions** :
  - directives du consortium ENCODE en 2011
  - **plus de deux répétitions biologique**

- Chez l'**humain 100M de lectures** sont suffisantes pour détecter **90 % des transcrits** de **81 % des gènes** du **transcriptome humain**.

- **20M de lectures** (75bp) permettent de détecter des **transcrits exprimés à un niveau moyen ou faible** chez le **poulet**.

- **10 M de lectures** permettent que **90% des transcrits** (**humain, zebrafish**) soient **couverts par 10 lectures en moyenne**.

(Plus d'informations : Toung et al. 2011 ; Wang et al. 2011 ; Hart et al. 2013)

Profondeur / Répétitions ?

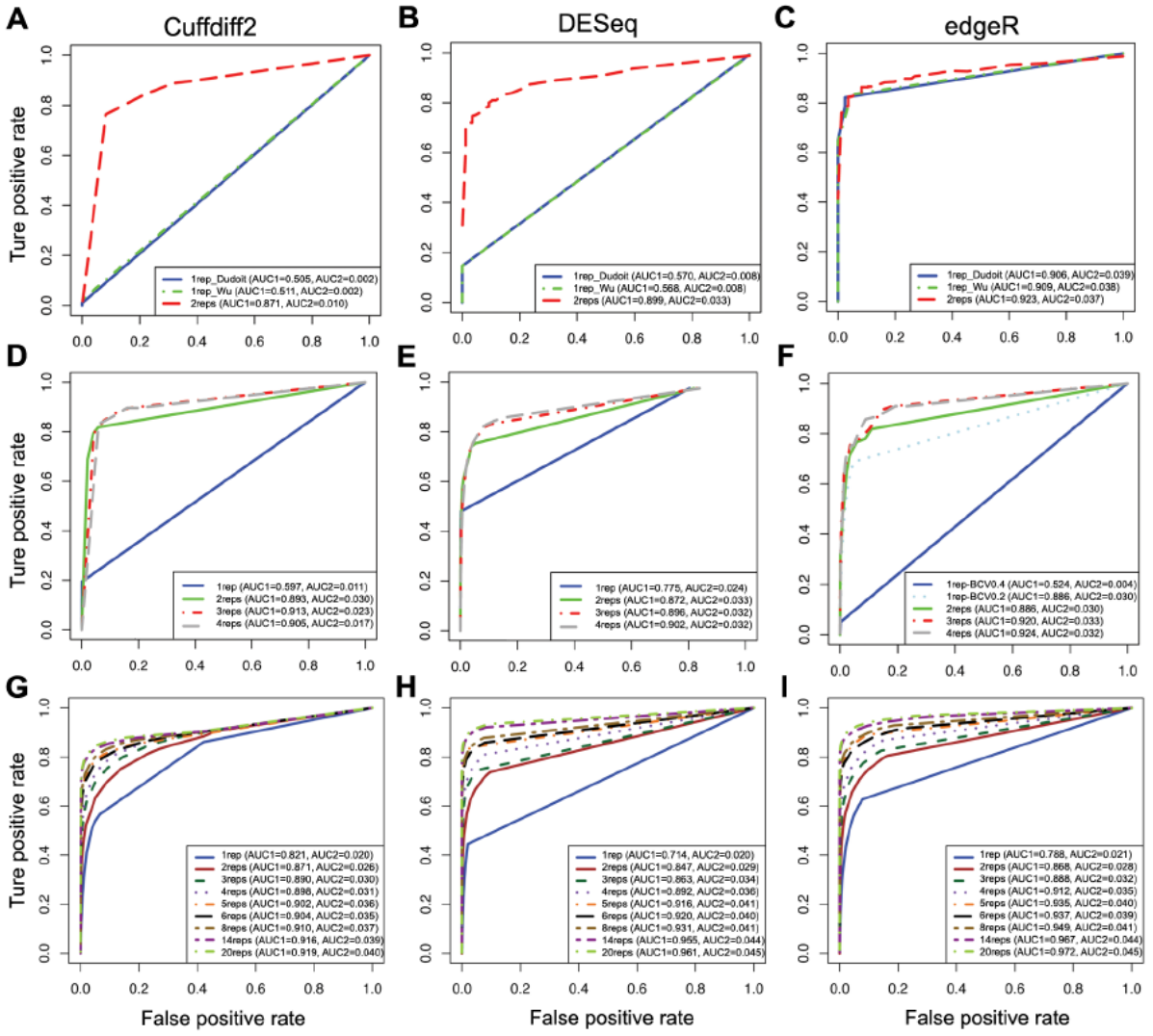
- **Pourquoi augmenter le nombre de répétitions biologiques ?**
  - Généraliser les résultats à la population
  - Estimer avec plus de précision la variation de chaque transcrit individuellement *(Hart et al. 2013)*
  - Améliorer la détection des transcrits différentiels et le contrôle du taux de faux positifs : **VRAI à partir 3** *(Zhang et al. 2014,Sonenson et al. 2013, Robles et al 2012)*

Profondeur / Répétitions ?

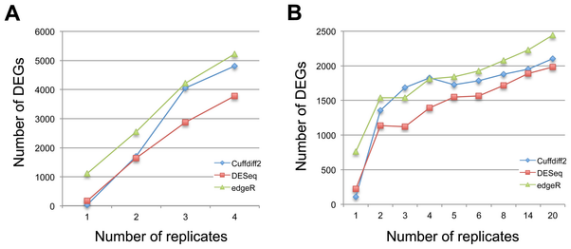
MAQC

L'effet du nombre de réplicats sur le taux de vrai positifs et de faux positifs

K_N



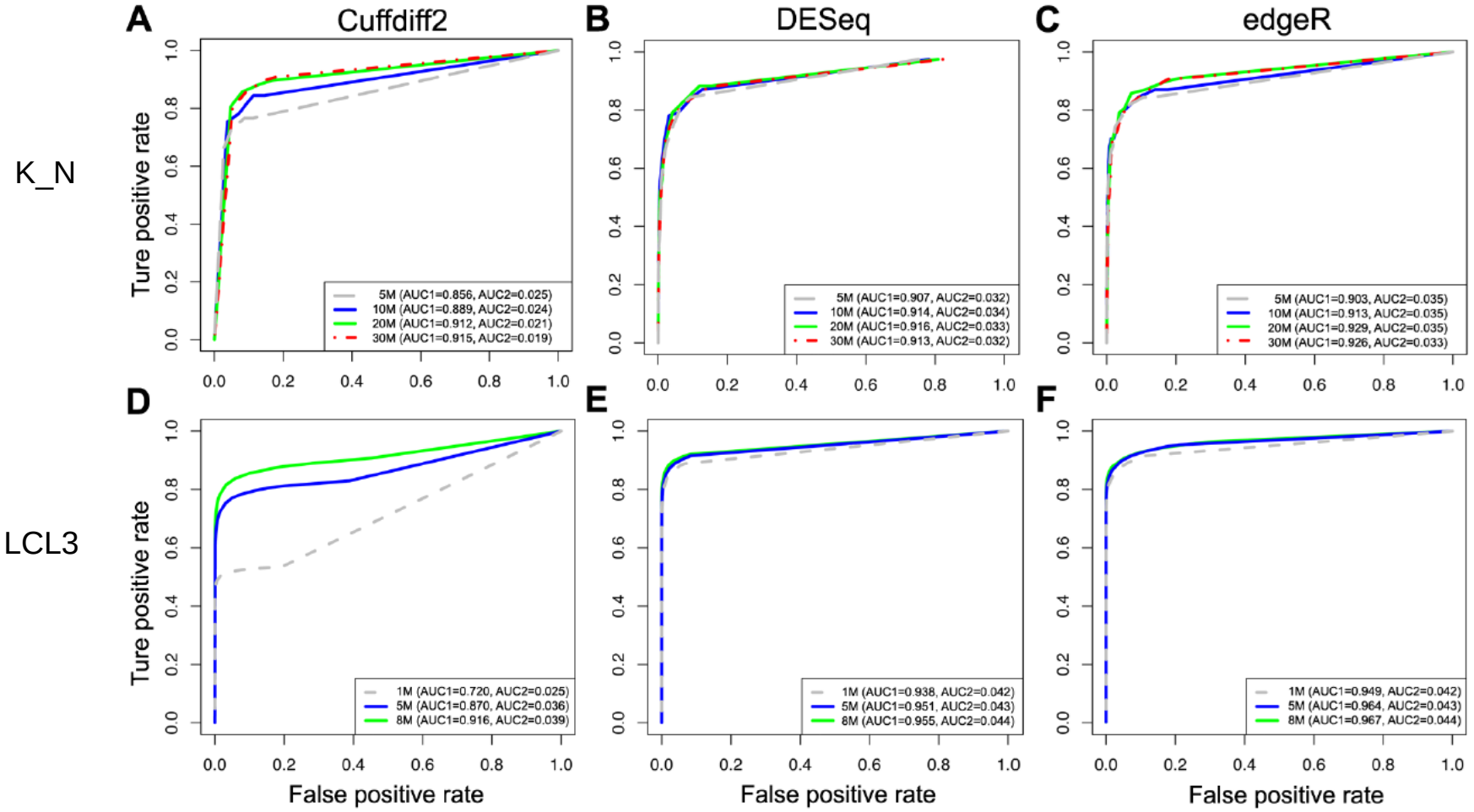
K_N



LCL2

Profondeur / Répétitions ?

L'effet de la profondeur.



Profondeur / Répétitions ?

- **Quel choix ? Plus de profondeur *ou plus de* répétition ?**

- **Ça dépend !** (Haas et al. 2012, Liu Y. et al 2013)

  - Détection de transcrits différentiels :
    (+) répétitions biologiques

  - Construction/annotation transcriptome :
    (+) profondeur & (+) conditions

  - Recherche de variants :
    (+) répétitions biologiques & (+) profondeur

A quelles questions biologiques PEUT répondre le RNA-seq ?

- L'**analyse d'expression différentielle** (différence d'expression) au niveau du transcriptome
- L'étude de l'**épissage alternatif** (isoformes) et recherche de **nouveaux transcrits**
  - amélioration des annotations structurales existantes
- La recherche d'**allèles spécifiques** et la **quantification** de leur **expression**
- La construction d'un **transcriptome *de novo*** (organismes non modèles)

Stratégie d'analyse en fonction des données disponibles

- **De novo :**
  - Pas de génome/transcriptome de référence
  - Outils en évolution permanente
  - Ressources (cpu/disque) +++
- **Transcriptome de reference**
  - Dépendant de la qualité de l'annotation structurale
  - Peu couteux
- **Génome de référence**
  - Permet une approche combinée :
    - sur **transcriptome**
    - recherche de **nouveaux transcrits**
  - Ressources ++
  - Alignement épissé

Pipeline d'analyse RNA-Seq : avec référence

- **Contrôle qualité**
- **Pre-nettoyage** des lectures
  - **suppression** des **adaptateurs** de **séquençage**
  - (**suppression** des **adaptateurs** de **multiplexage**)
- **Nettoyage** des lectures
  - **tronquer** les **extrémités** de **mauvaise qualité** des lectures
- **Alignement** des **lectures sur** la **référence**
  - gènes ou génome complet
- **Comptage** des gènes / transcrits

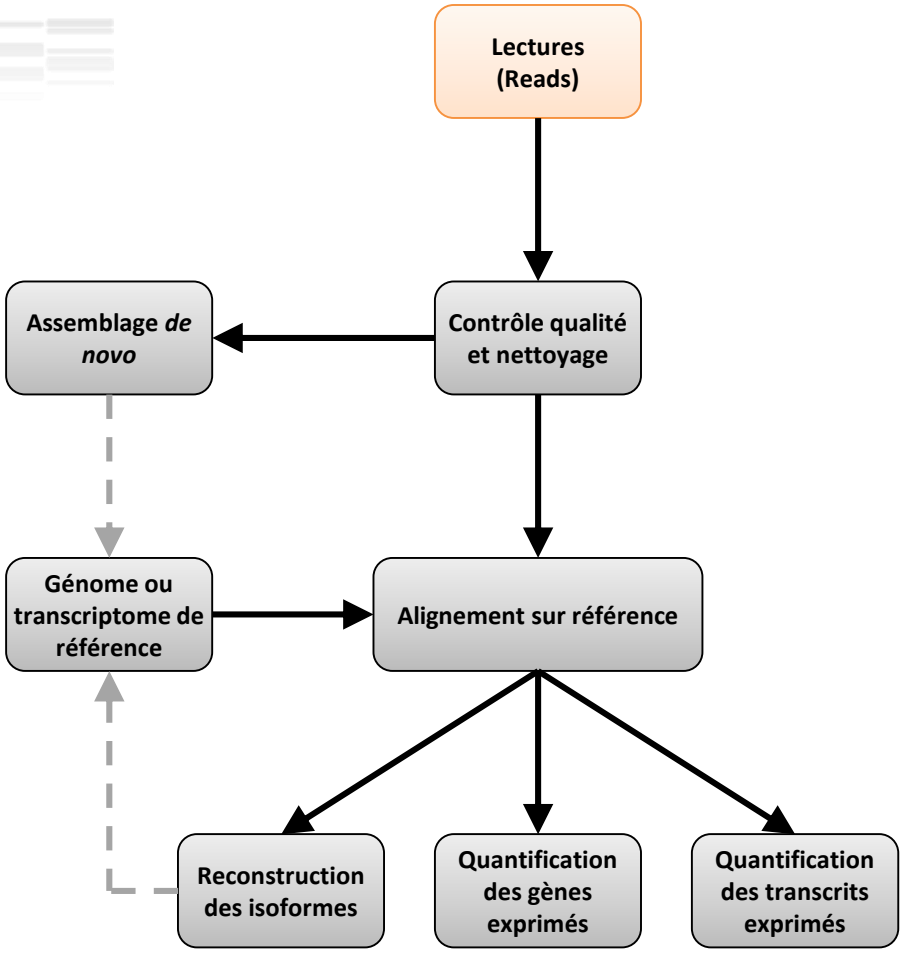
Galaxy : un outil de traitement de données

02

DONNÉES BRUTES

Obtenir des séquences de qualités

Workflow d'analyse RNA-Seq



Plan

- Le format fastq
- Les biais connus
- Vérification de la qualité avec FastQC
- Nettoyage des lectures avec Sickle

Format Fastq

- 1 séquence = 4 lignes dans le fichier

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((****+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

- 1 ère ligne = identifiant de la séquence

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
Y	Y if the read fails filter (read is bad), N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

Format Fastq

- 4ème ligne = Qualité

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((****+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

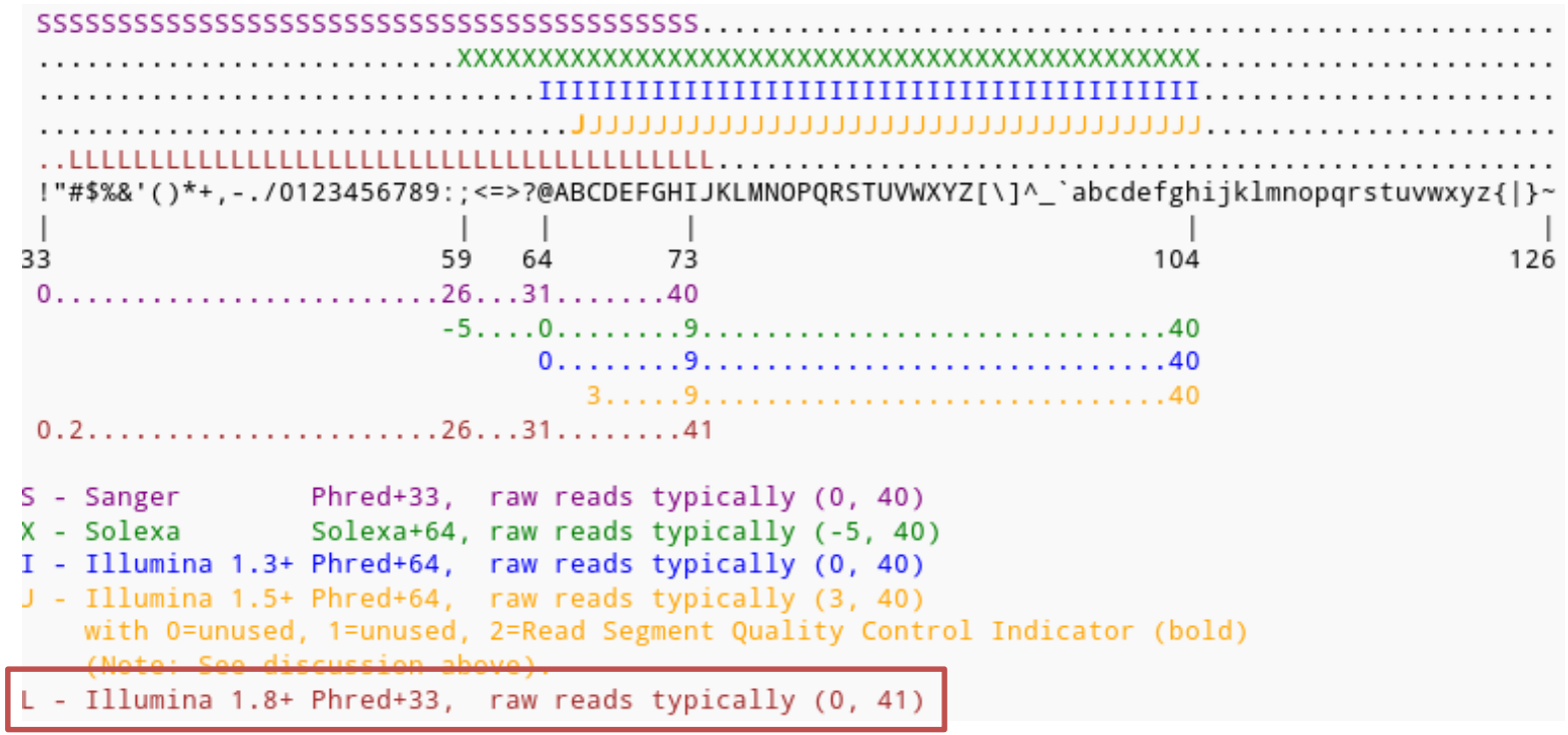
- Appelée aussi Phred quality score (Sanger format)

$$Q_{\text{sanger}} = -10 \log_{10} p$$

Probabilité qu'une base soit incorrecte

Format Fastq

- Qualité Encodée en ASCII



Objectif du contrôle qualité

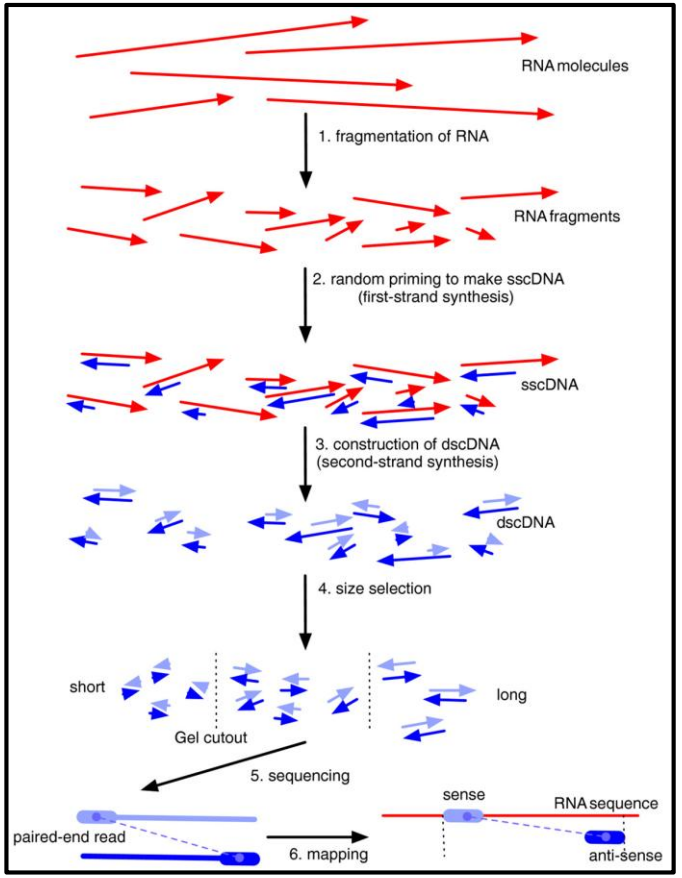
- Vérifier que les séquences sont **conformes au niveau de prestation attendu (taille, nombre, qualité,…)**
- Vérifier que les séquences peuvent **répondre au questions biologiques** posées :
  - **Biais techniques**
  - **Biais biologiques**
- Aider au paramètres pour le nettoyage des données

Biais spécifiques au RNA-Seq

- Influence du mode de préparation de la banque
  - amplification hexamèrique aléatoire (**Random hexamer priming**)
- Influence du séquençage
  - biais de position, de composition en séquence (contenu en GC)
  - influence de la longueur des transcrits
- « Mapabilité » du génome/transcriptome

Préparation de la banque

- **Extraction ARN total**
- **Déplétion** (queue polyA)
- **Fragmentation**, **reverse transcription** avec des **hexamères aléatoires** -> **dscDNA**
- **Séquençage**



Roberts et al. Genome Biology 2011, 12:R22

Biais : *random hexamer priming*

- Fort biais de composition des 13 premières nucléotides en 5’
  - spécificité de séquence de la polymérase

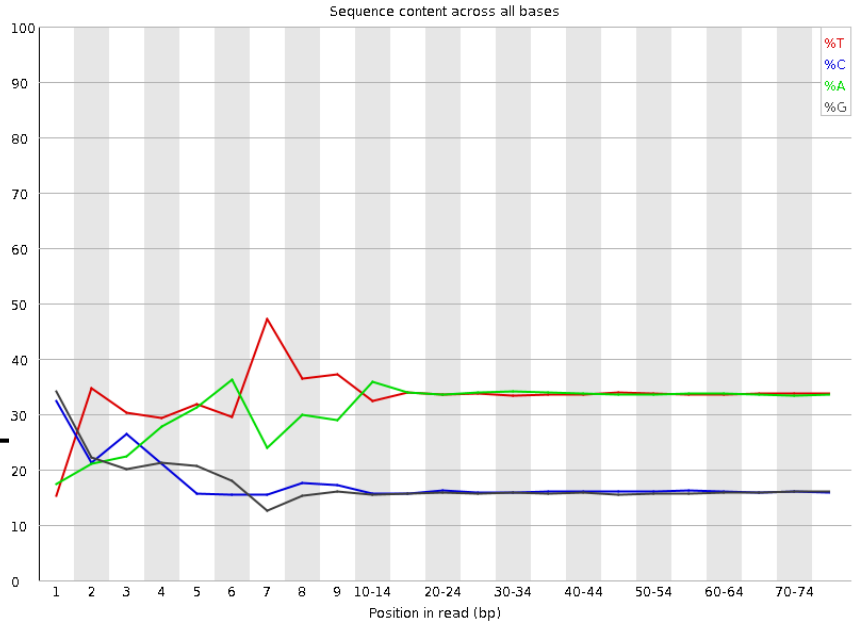
Published online 14 April 2010 *Nucleic Acids Research, 2010, Vol. 38, No. 12 e131*
doi:10.1093/nar/gkq224

Biases in Illumina transcriptome sequencing caused by random hexamer priming

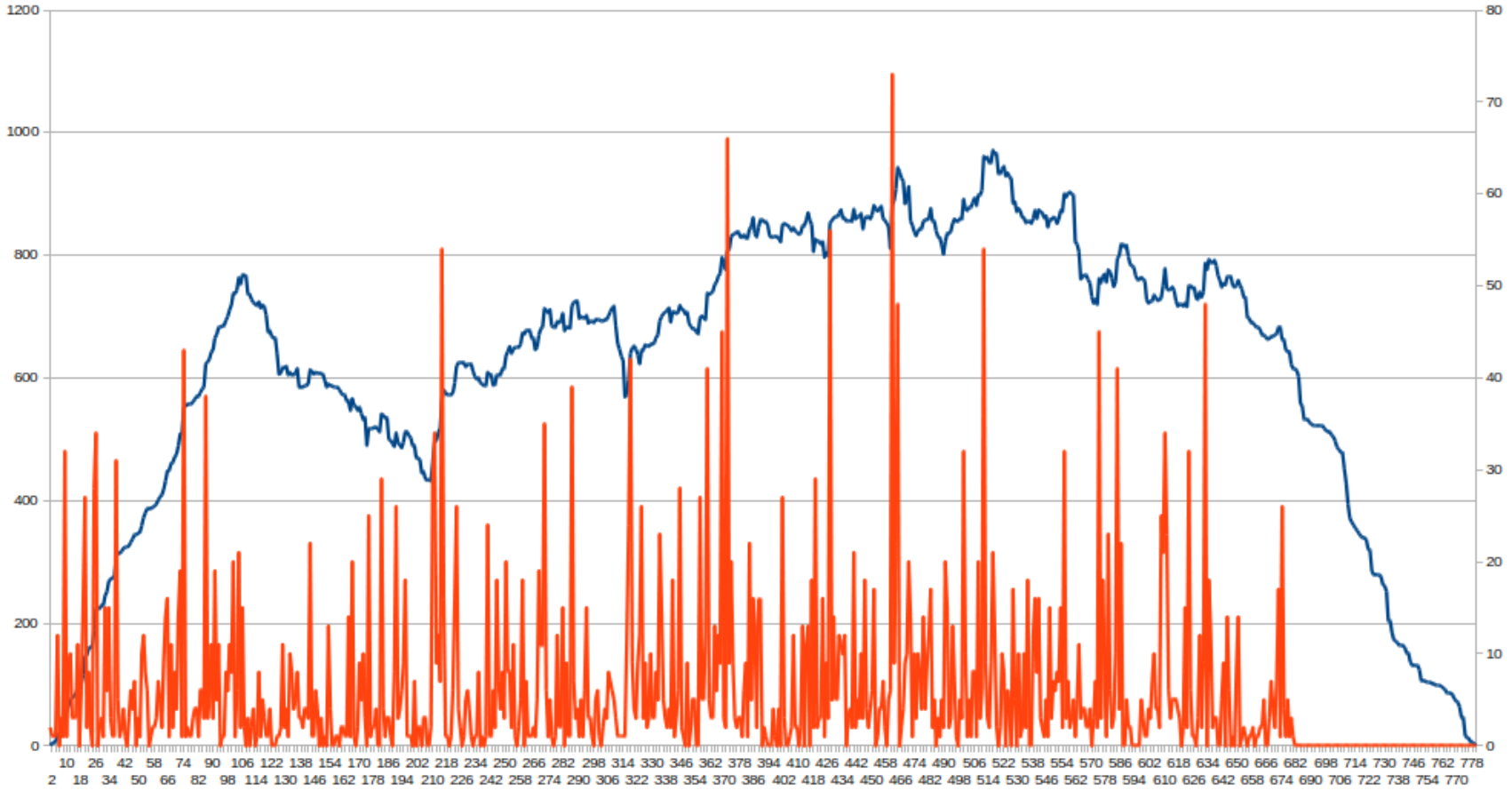
Kasper D. Hansen[1,*], Steven E. Brenner[2] and Sandrine Dudoit[1,3]

ABSTRACT

Generation of cDNA using random hexamer priming induces biases in the nucleotide composition at the beginning of transcriptome sequencing reads from the Illumina Genome Analyzer. The bias is independent of organism and laboratory and impacts the uniformity of the reads along the transcriptome. We provide a read count reweighting scheme, based on the nucleotide frequencies of the reads, that mitigates the impact of the bias.



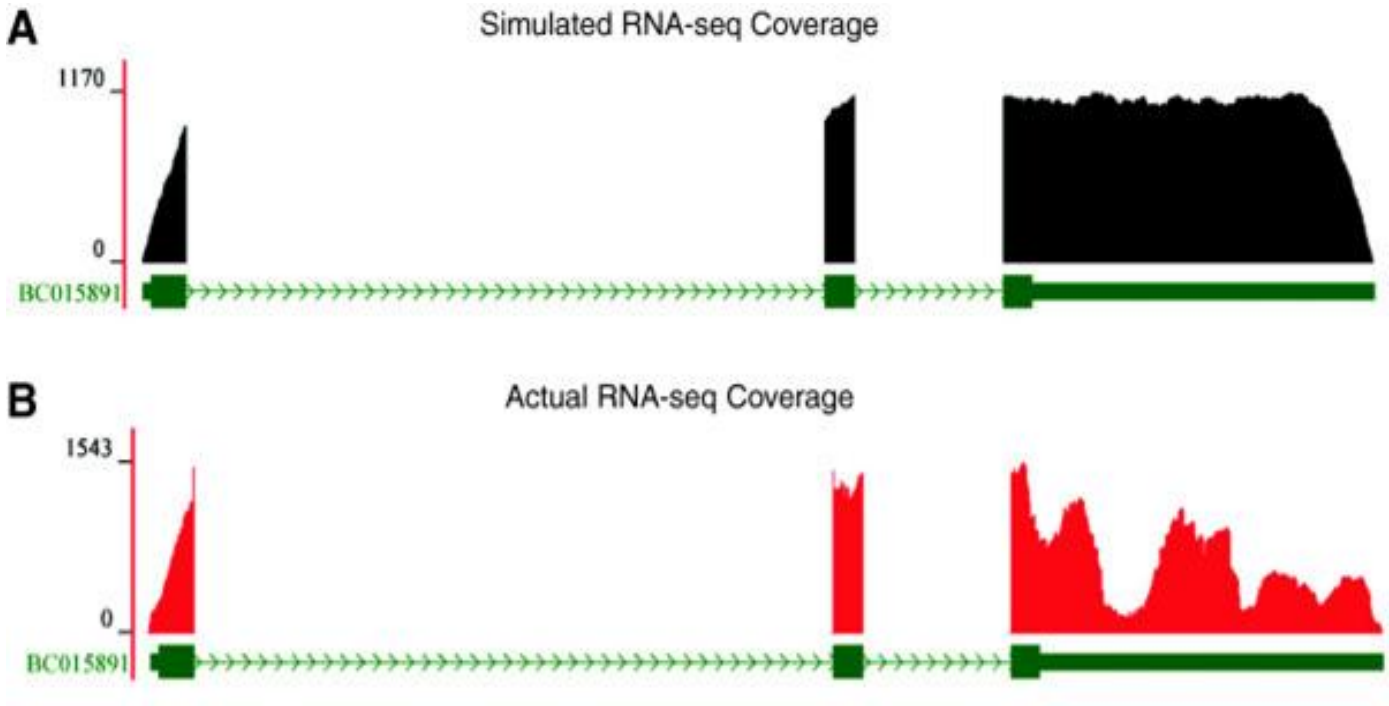
Biais : *random hexamer priming*



Orange = reads start sites

Blue = coverage

Biais : *Préparation de la librairie*



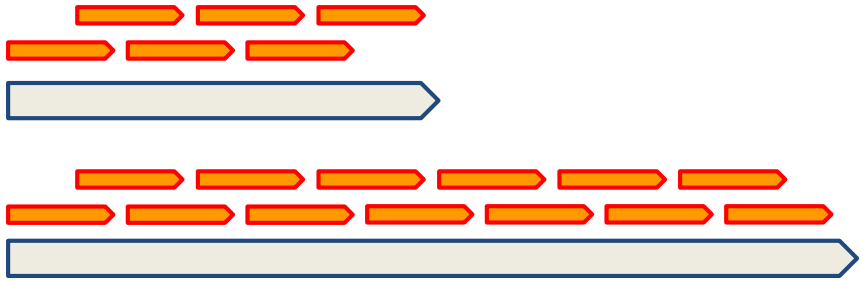
IVT-seq reveals extreme bias in RNA sequencing, Lahens et al 2014

http://genomebiology.com/2014/15/6/R86

Biais : longueur des transcrits

- La capacité, en utilisant des **comptages** obtenus par **RNA-Seq**, a observer un transcrit comme étant **différentiellement exprimé** est **directement reliée** à sa **longueur**.

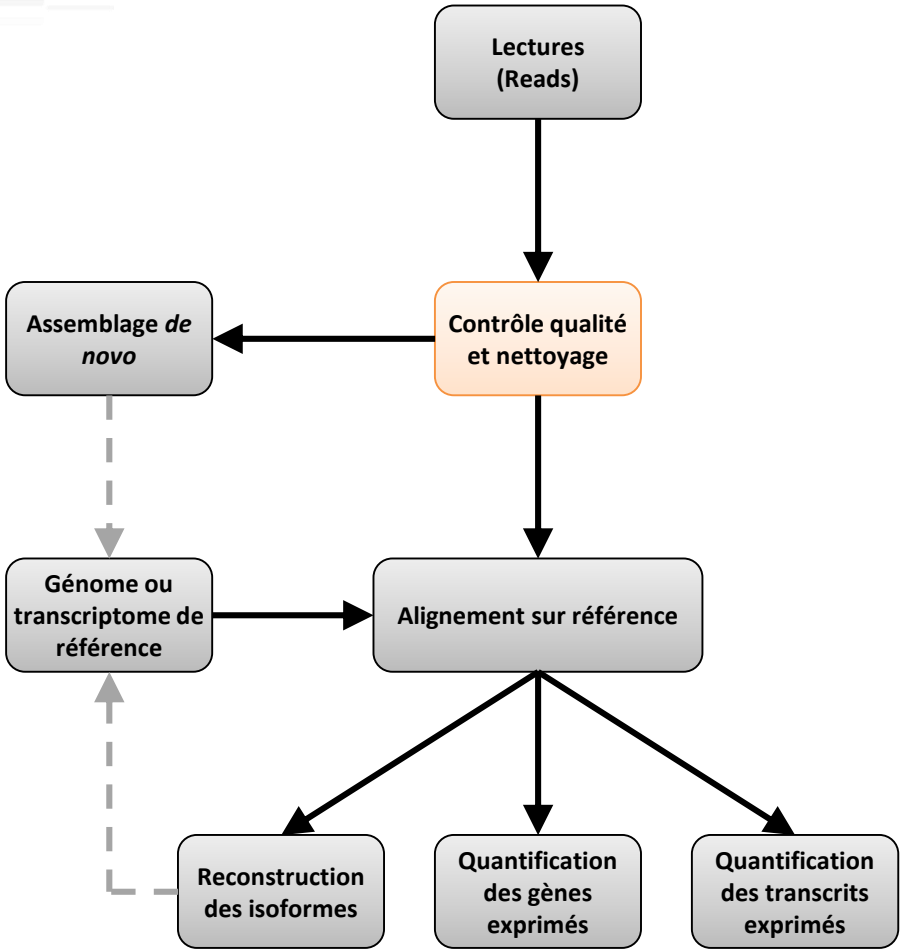
- Pour un **même gène** ayant **deux isoformes**, l'une faisant la moitié de l'autre, exprimé en **même abondance dans deux conditions différentes** :
  - L'isoforme la plus courte sera deux fois moins « comptée » que la plus longue



Biais : « mappabilité »

- Les étapes bioinformatiques peuvent être **influencées** par :
  - La **qualité** de la **référence**
    - ✓ **assemblage**
    - ✓ **finition**
  - La **composition** de la **séquence**
    - ✓ **zones répétées**
  - La **qualité** de l'**annotation**

Workflow d'analyse RNA-Seq



Contrôle qualité

Objectifs :

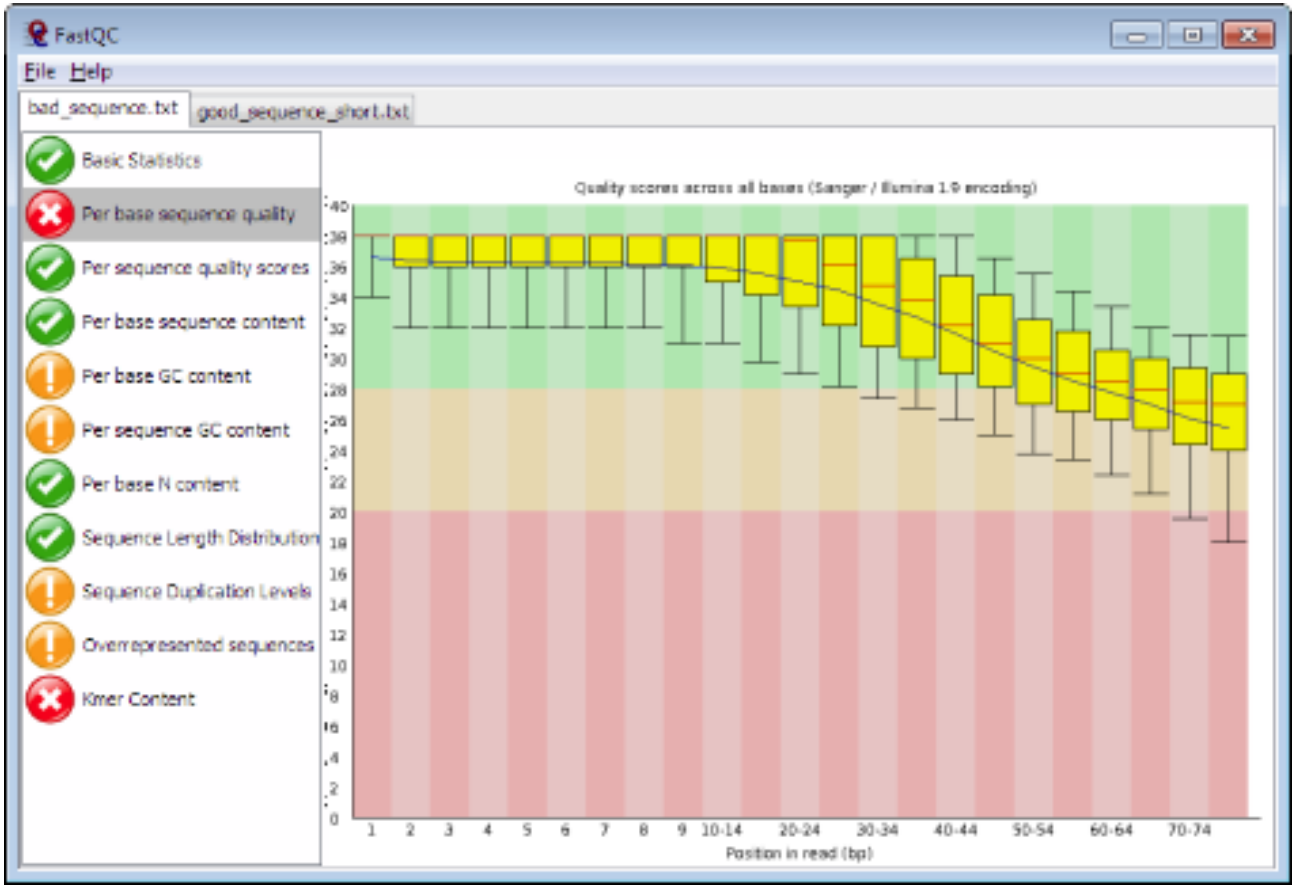
- Vérifier que les séquences sont **conformes au niveau de prestation attendu (taille, nombre, qualité,…)**

- Vérifier que les séquences peuvent **répondre au questions biologiques** posées :
  - **Biais techniques**
  - **Biais biologiques**

- Aider au paramètres pour le nettoyage des données

Contrôle qualité avec FastQC

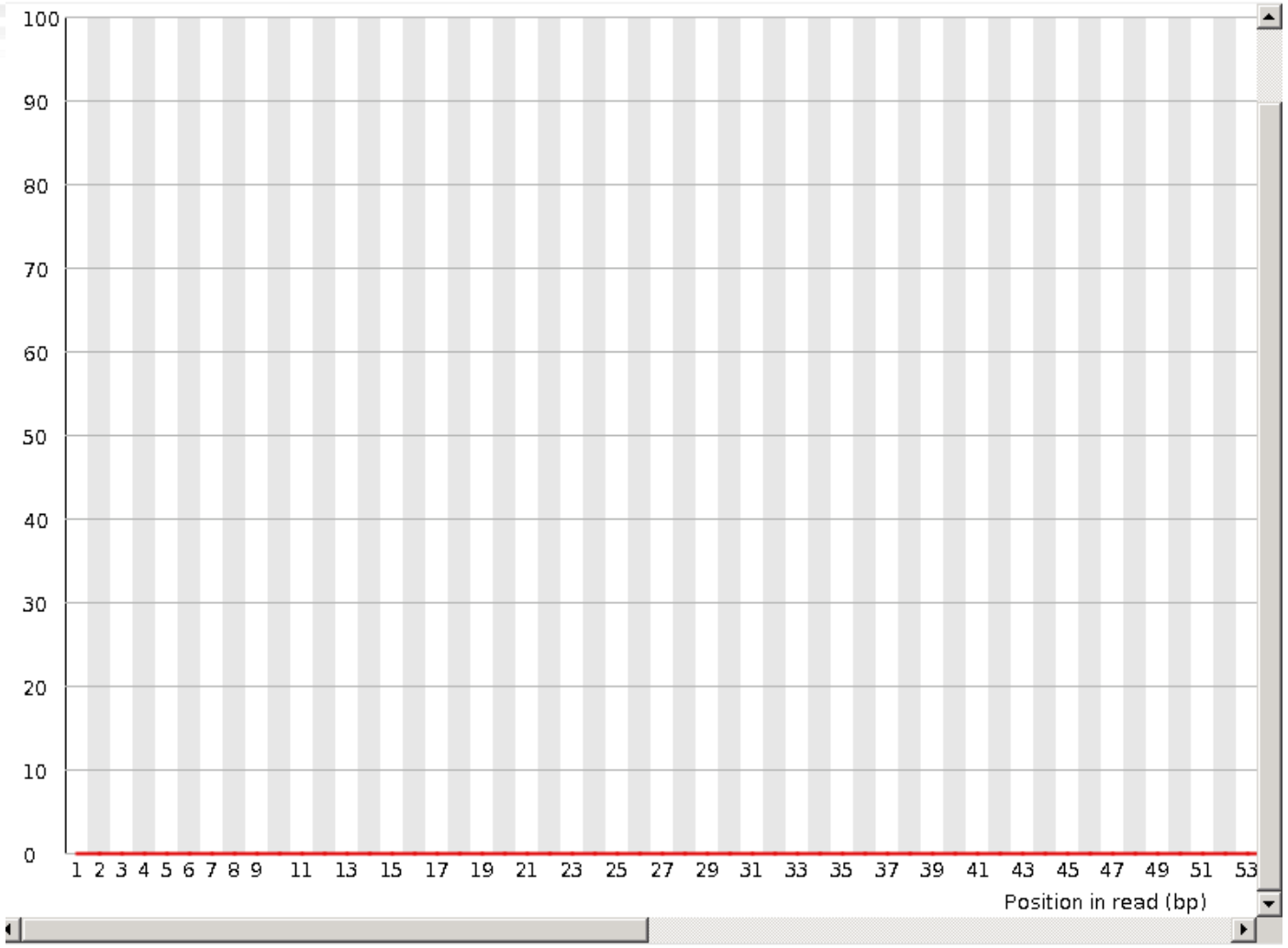
- orienté DNA-Seq



http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc

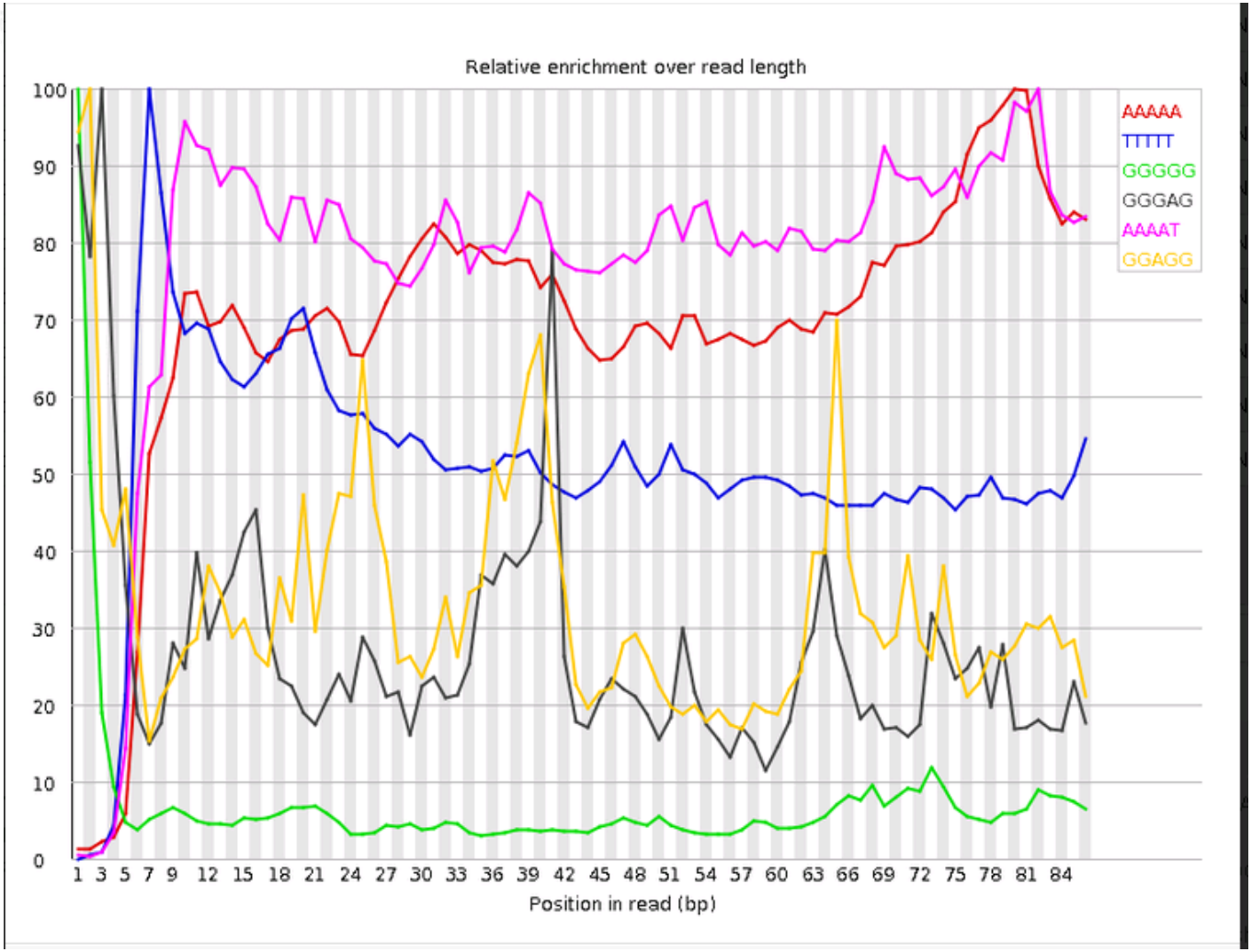
Contrôle qualité

N content



Contrôle qualité

Kmer content



Contrôle qualité

Over-expressed sequences

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	64433	0.20451959970239228	No Hit

Back to summary

- **Détecter les adaptateurs**

Nettoyage des données

- **Nettoyage « optionnel »**

- **L'alignement permettra de supprimer les lectures**
  - De mauvaise qualité
  - D'adaptateurs
  - Contaminantes

- **Les outils :**
  - Cutadapt : Nettoyage des adaptateurs & Tags
  - Prinseq : Nettoyage des lectures de mauvaise qualité
  - Sickle : Nettoyage des lectures de mauvaise qualité

Nettoyage des données

- **Principe de Sickle :**
  - Traite les paires ensemble
  - Fenêtre glissante : 10% de la taille des reads
  - Calcul de la qualité moyenne des lectures
    exemple : Longueur = 23

A	C	T	T	G	A	T	C	A	T	G	C	A	T	C	G	A	T	C	G	T	A	G
30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	25	20	18	18	10

Travaux pratiques

Présentation des objectifs

- **Aborder** les **différentes étapes** indispensables au **traitement bioinformatique** de **données RNA-Seq** à travers un **exemple** issu de **données réelles :**

- Séquençage de la tomate :

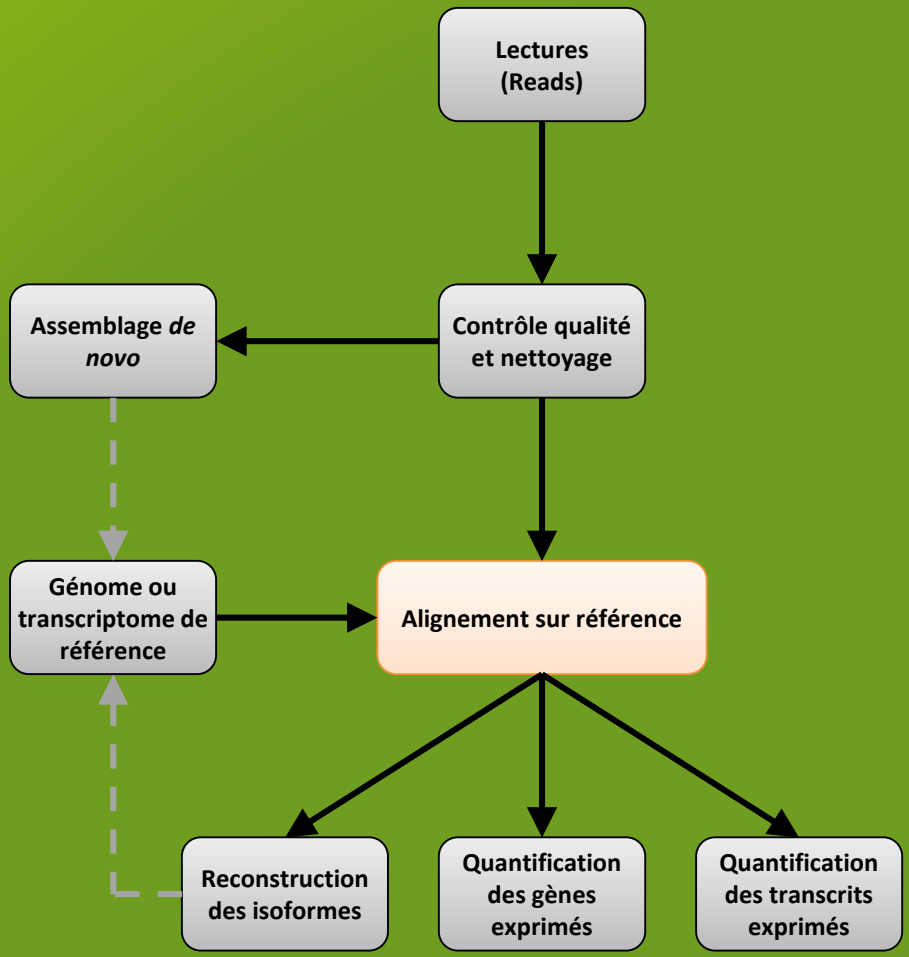
  - Wt : wild type, PAIRED

  - Mt : mutant type , PAIRED

03

MAPPING

et Visualisation



Alignement épissé

Objectifs :

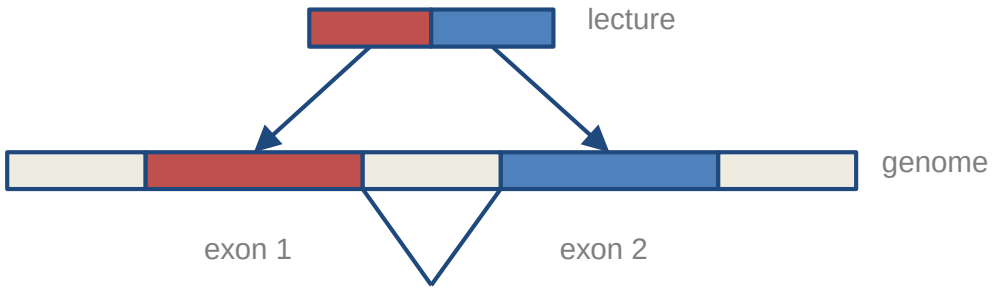
- **Aligner** les **lectures** issues du séquençage de **dscDNA** (transcrits) sur le **génome**, en tenant compte de l'**épissage alternatif**

- Être capable d'**exploiter** les listes des **jonctions exons-exons connues**, mais également d'en **détecter** de **nouvelles**

- Tout cela dans un **temps raisonnable**…

Introduction

Définition



Le *mapping* est la *prédiction* du *locus* dont est originaire la lecture.

- **Prédiction :** chaque outil propose un/plusieurs locus. Ils peuvent ne pas être les bons.
- **Locus :** le résultat est un ensemble de positions génomiques (ex.: chr1:100..150)

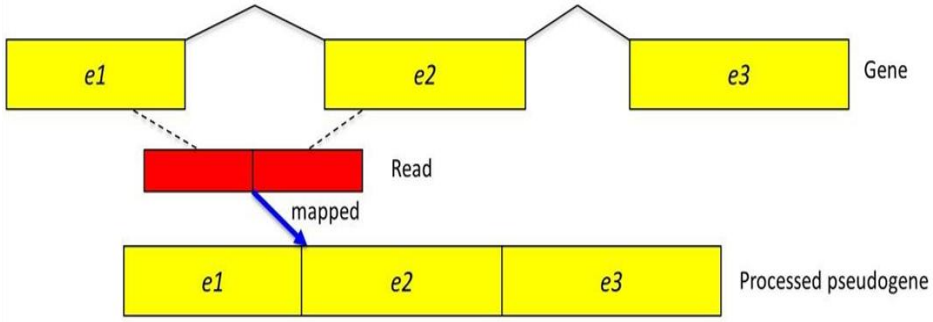
- Mapping ARN ≠ Mapping ADN

- Mapping ≠ Alignement

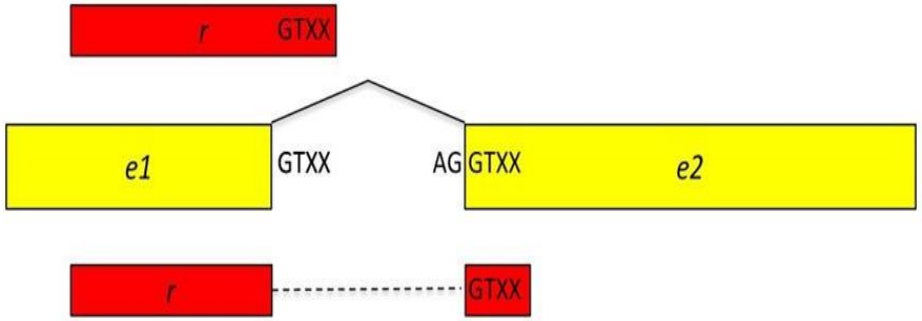
Les outils de mapping font de mauvais alignements (sauf aux jonctions).

Cas difficiles

- Beaucoup de différences : erreurs de séquençage ou locus muté
- Séquence répétée
- Lecture sur 3+ exons
- (Variante :) Gène ou pseudo-gène ?



- Fin de la lecture sur un exon propre



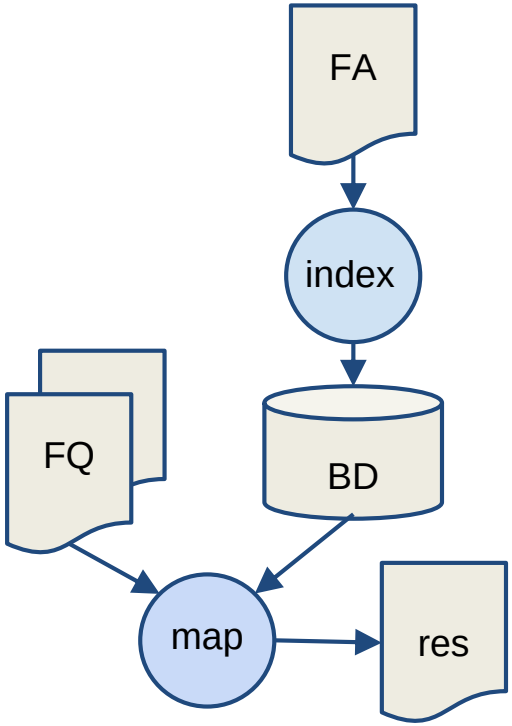
(Kim et al, Genome Biology, 2013)

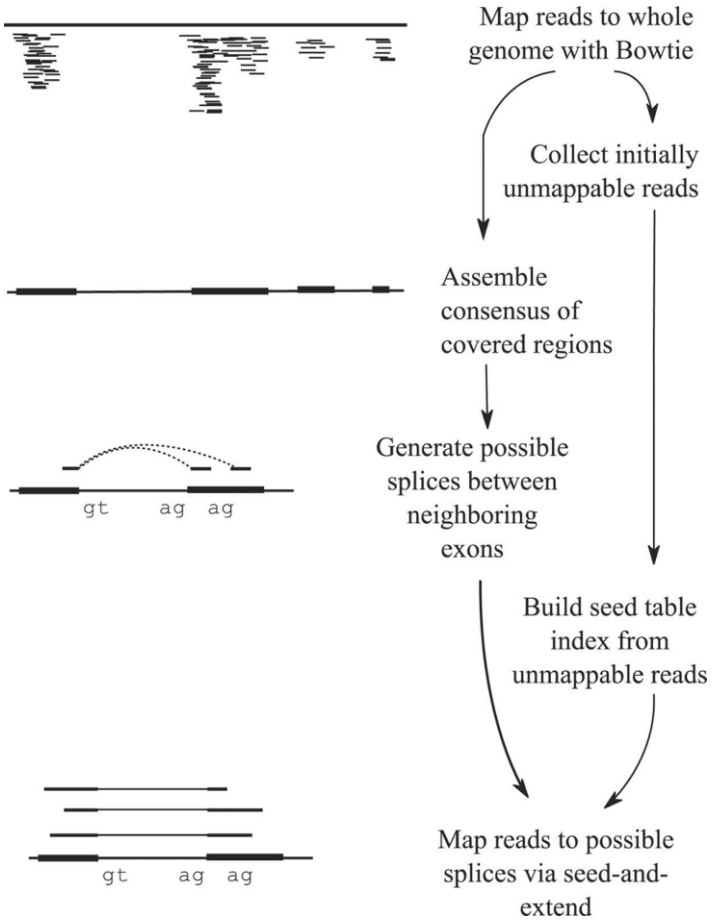
- Lecture sur une jonction non-connue d'un gène peu exprimé

Tophat1

Étapes de mapping :

- *Indexation du génome 1 fois pour toutes*
- *Mapping des lectures utilise l'index*



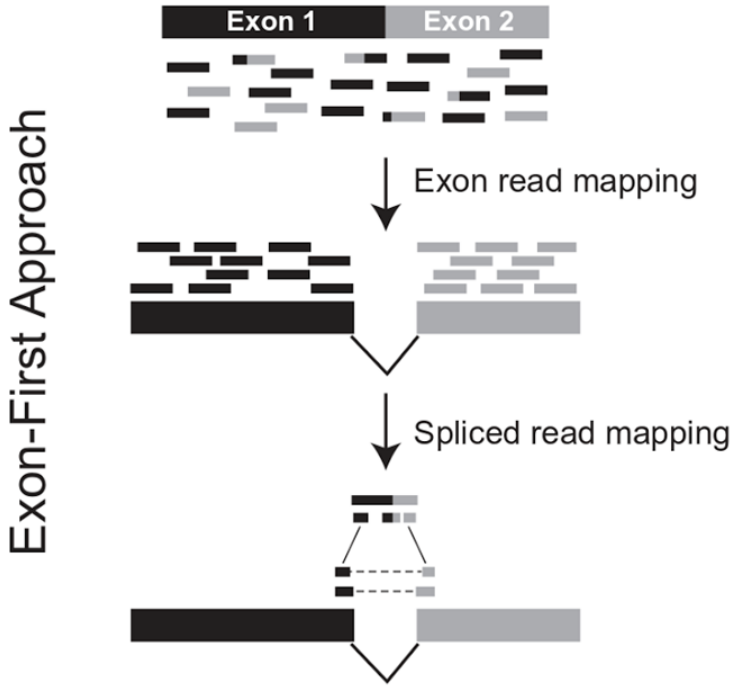


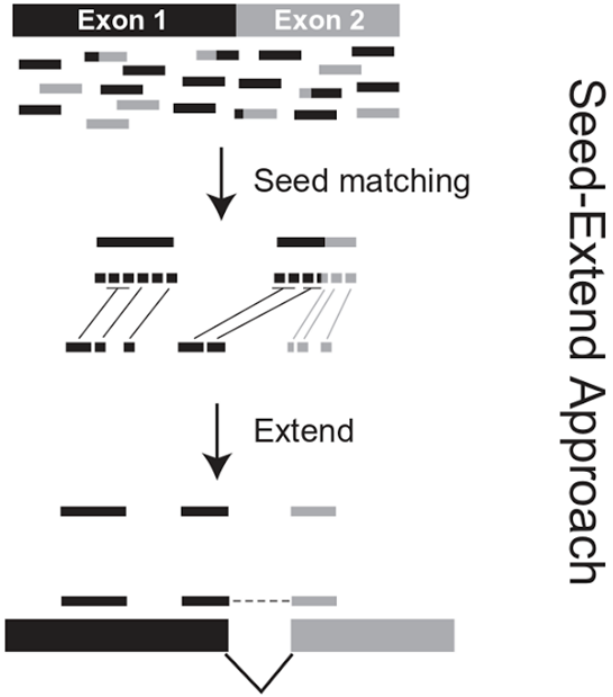
Utilise Bowtie pour mapper les lectures sur le génome.

Problème pour les pseudo-gènes !

(Trapnell et al, Bioinformatics, 2009)

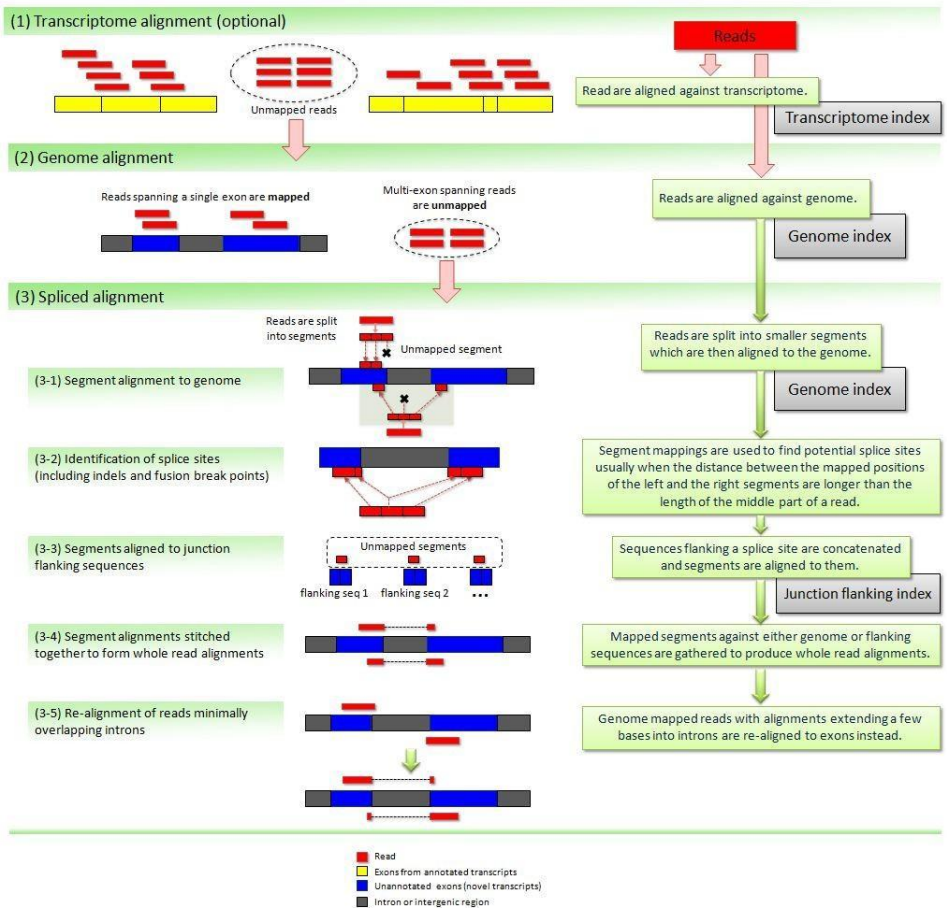
Tophat1 : 2 types d'algo





(Garber et al, Nature Methods, 2011)

Tophat2

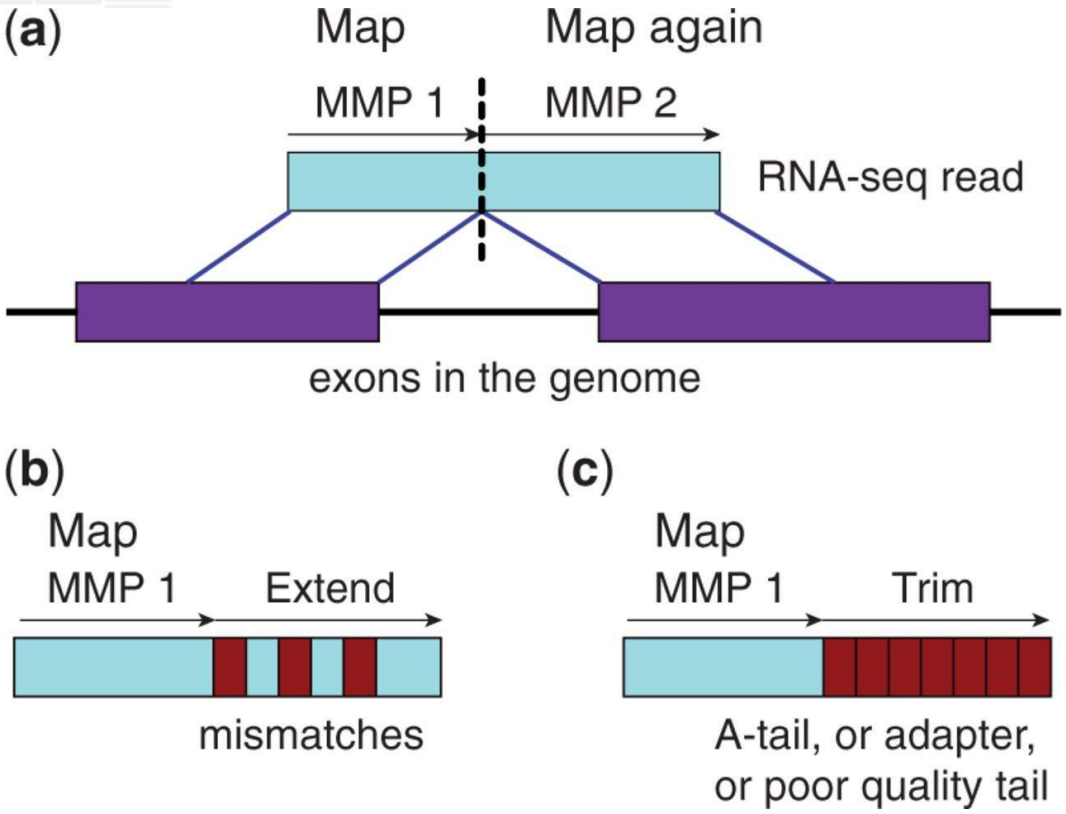


Tophat2 est constitué de beaucoup d'étape pour résoudre chaque cas difficile.

Chaque étape contient des heuristiques dont les paramètres sont à fixer.

(Kim et al, Genome Biology, 2013)

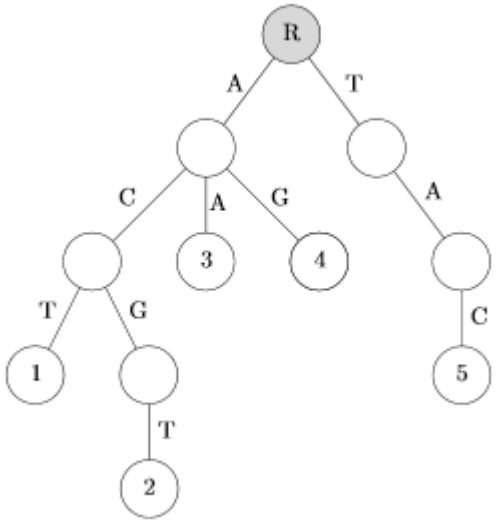
Star



(Dobin et al, Bioinformatics, 2011)

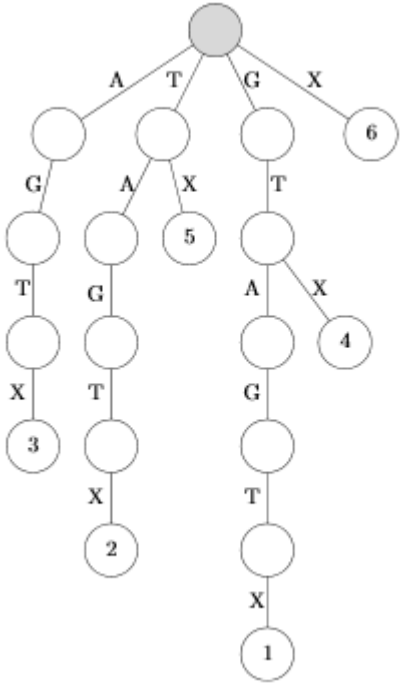
(RNA-)STAR

Aligneurs index BWT (BWA, Bowtie, SOAP)



ACT, *ACGT*, *AA*, *AG*, et *TAC*.

STAR



GTAGT.

Outils existants

- Tophat2 (le plus utilisé, le plus suivi)
- Star (runner-up)
- Crac (français !)
- GSNAP
- RUM
- MapSplice
- Gem
- …

Outils existants

La plupart des outils

- utilise des sites de jonctions donnés par l'utilisateur pour "s'aider"
- suppose des sites canoniques GT-AG

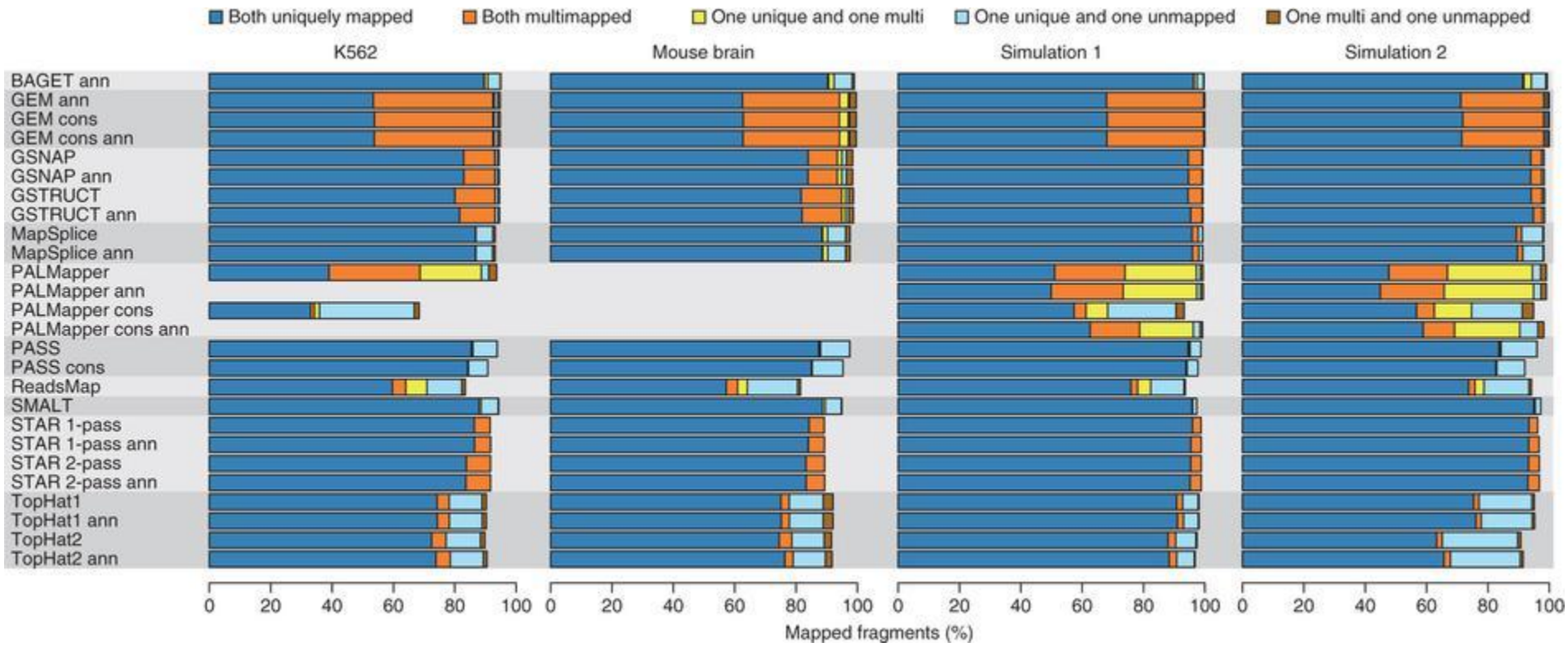
Comment évaluer un outil ?

- Sensibilité (mappe le plus de lectures)
- Spécificité (ne se trompe pas)
- … sur les lectures et sur les jonctions
- Temps
- Mémoire

En général, les critères sont contradictoires.

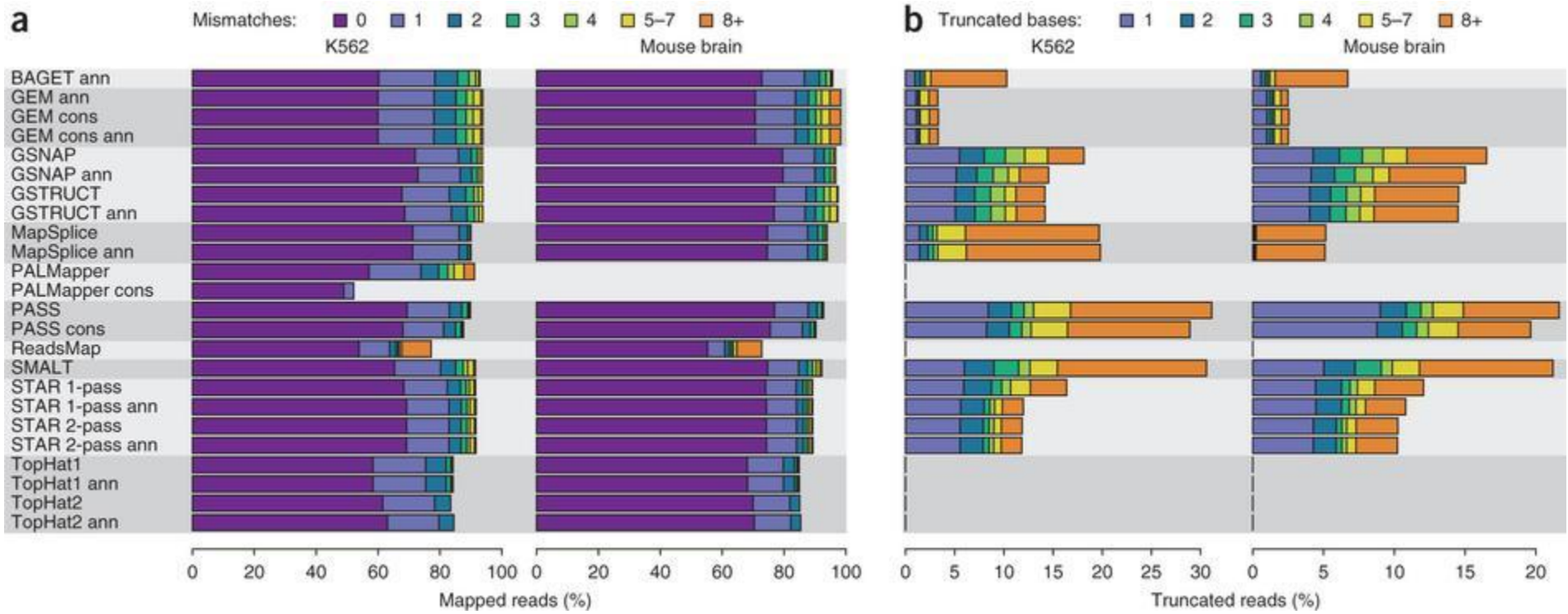
RGASP 3

The RNA-seq Genome Annotation Assessment Project
(Engström et al., Nature Methods, 2013)



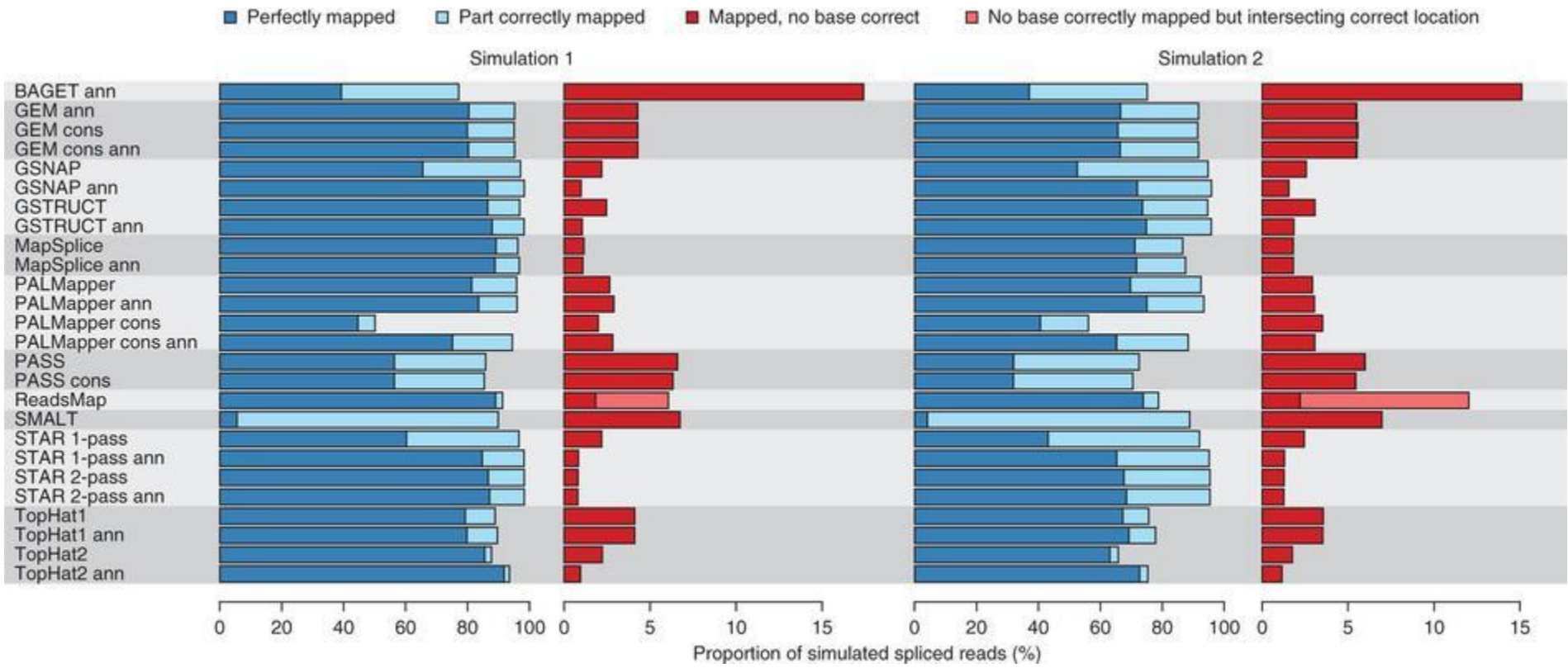
RGASP 3

The RNA-seq Genome Annotation Assessment Project
(Engström et al., Nature Methods, 2013)



RGASP 3

The RNA-seq Genome Annotation Assessment Project
(Engström et al., Nature Methods, 2013)



RGASP 3

The RNA-seq Genome Annotation Assessment Project
(Engström et al., Nature Methods, 2013)

Les phrases clés

« Mapping properties are largely dependent on software algorithms even when the genome and transcriptome are virtually identical »

« Exon detection results based on K562 data were similar for GEM, GSNAP, GSTRUCT, MapSplice, STAR and TopHat »

RGASP 3

The RNA-seq Genome Annotation Assessment Project
(Engström et al., Nature Methods, 2013)

STAR	vs	TopHat
+	# de lectures alignées	-
-	# de lectures correctement alignées	+
-	sensibilité aux variations	+
-	sensibilité aux annotations	+

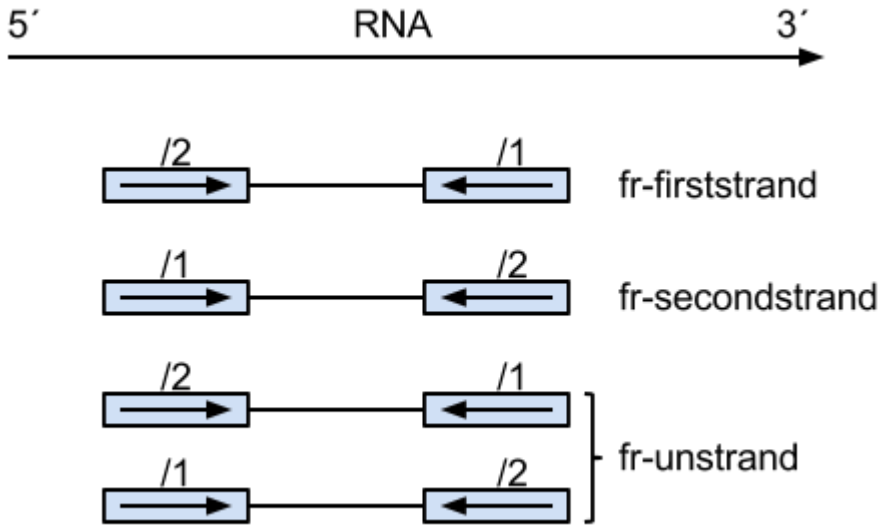
Alignement : données initiales

- **Lectures (brutes / nettoyées ?)**
- **Génome de référence éventuellement annoté :**
  - Séquence nucléique (fasta)
  - Annotation structurale (GTF)
- **Ou trouver un génome et un transcriptome de référence ?**
  - Ensembl
  - NCBI
- **Exo : trouver votre génome préféré et son annotation.**

Tophat

- En **entrée** :
  - **lectures** (**.fastq**)
  - **index bowtie2** de la **référence**
  - **annotation structurale** du **génome** (**.gtf**) [optionnel]
  - ***jonction*** *(**.bed**) [optionnel]*
  - ***insertions / délétions*** *(**.bed**) [optionnel]*

- Les options :
  - si paired, type de librairie.



Tophat

Attention aux paramètres par défaut !

- **Dans le manuel :**

> Please Note TopHat has a number of parameters and options, and their default values are tuned for processing mammalian RNA-Seq reads.
>
> If you would like to use TopHat for another class of organism, we recommend setting some of the parameters with more strict, conservative values than their defaults.
>
> Usually, setting the maximum intron size to 4 or 5 Kb is sufficient to discover most junctions while keeping the number of false positives low.

http://tophat.cbcb.umd.edu/manual.shtml

Format GTF : Gene Transfert Format

- **Dérivé** du format généraliste GFF (General Feature Format)
- Contient l'**annotation structurale** du **génome** (gène, transcrits)

<seqname> <source> **<feature>** <start> **<end>** <score> **<strand>** <frame> **[attributes]** [comments]

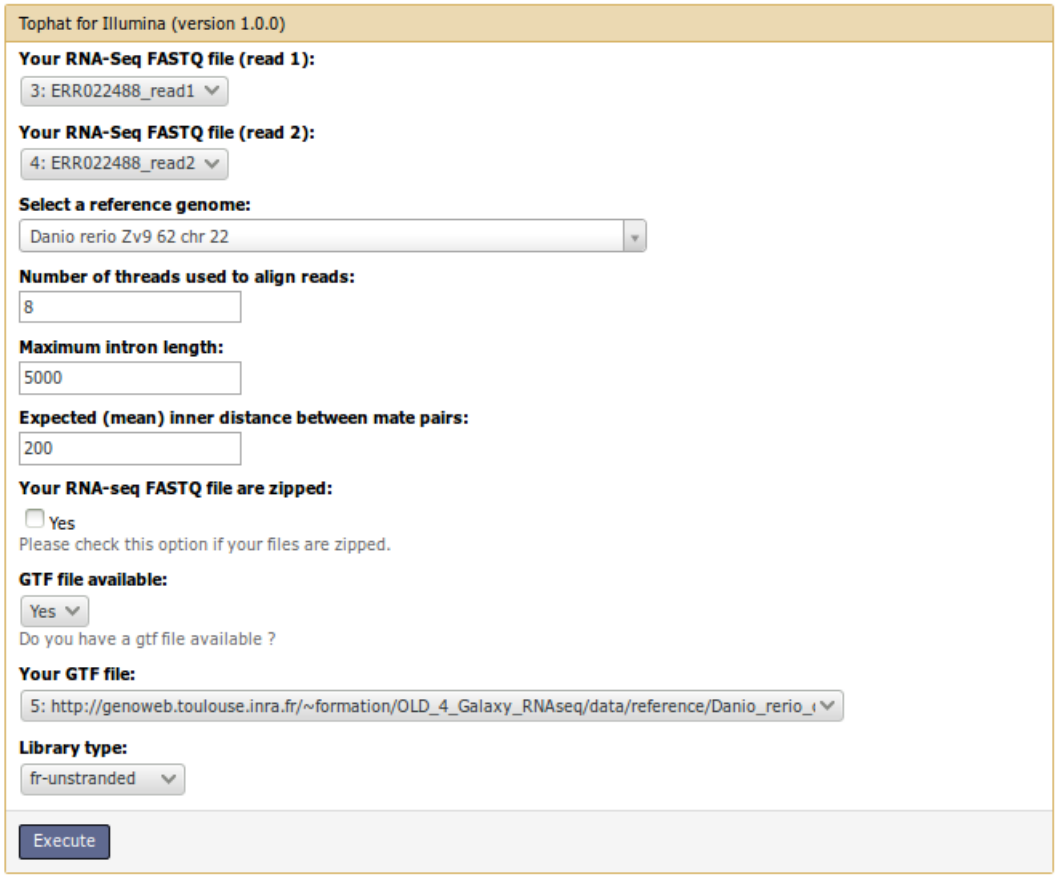
3R protein_coding **exon** 380 **509** . **+** . **gene_id "FBgn0037213"; transcript_id "FBtr0078961";**
exon_number "1"; gene_name "CG12581"; transcript_name "CG12581-RB";

- **Le champ attribut doit :**
  - **Commencer** par le ***gene_id*** : identifiant **unique** du gène
  - **Être suivi** par ***transcript_id*** : identifiant **unique** du transcrit prédit

- Les identifiants du chromosome (**Fasta** et **1ère colonne** du **GTF**) doivent être les **mêmes**

http://genome.ucsc.edu/FAQ/FAQformat.html#format4

TP : Assemblage avec Tophat



- ***Lancer Tophat maintenant ! Avant de continuer la présentation.***

http://tophat.cbcb.umd.edu/manual.shtml

Alignement : Format SAM/BAM

- Le partage des données est un problème majeur dans le projets “1000 génomes”
- Capturez toute l'information critique sur les données de NGS dans un seul fichier indexé et comprimé
- Alignement format générique
- Prise en charge reads de taille variable (454 - Solexa - Solid … PacBio)
- Flexible dans le style , de taille compacte , efficace en accès aléatoire

Website :
http://samtools.sourceforge.net
Paper :
Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: 19505943]

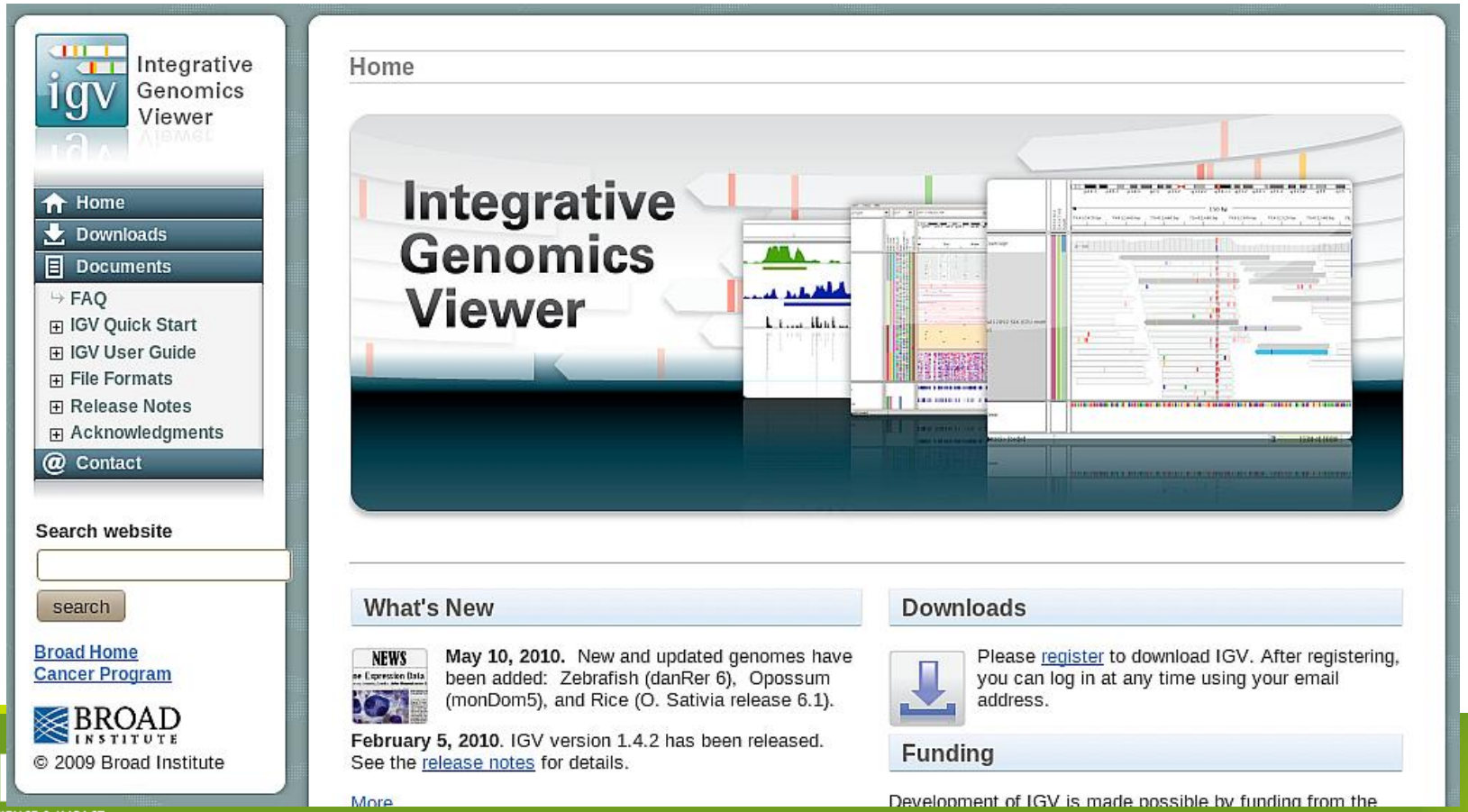
Alignement : Format SAM

Quelles informations doivent être stockées dans un fichier d'alignement SAM ?

http://genoweb.toulouse.inra.fr/~formation/2_Galaxy_SGS-SNP/.SAMformat/sam.html

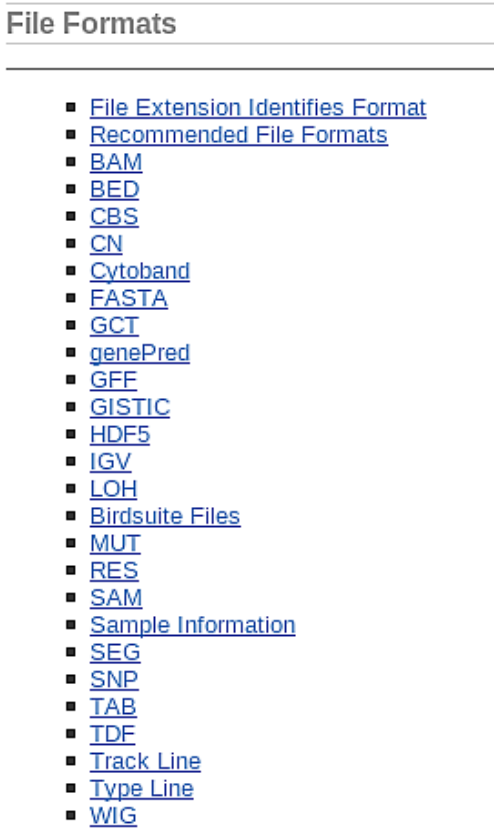
Visualisation des alignements avec IGV

- IGV : Integrative Genomics Viewer
- Website : http://www.broadinstitute.org/igv



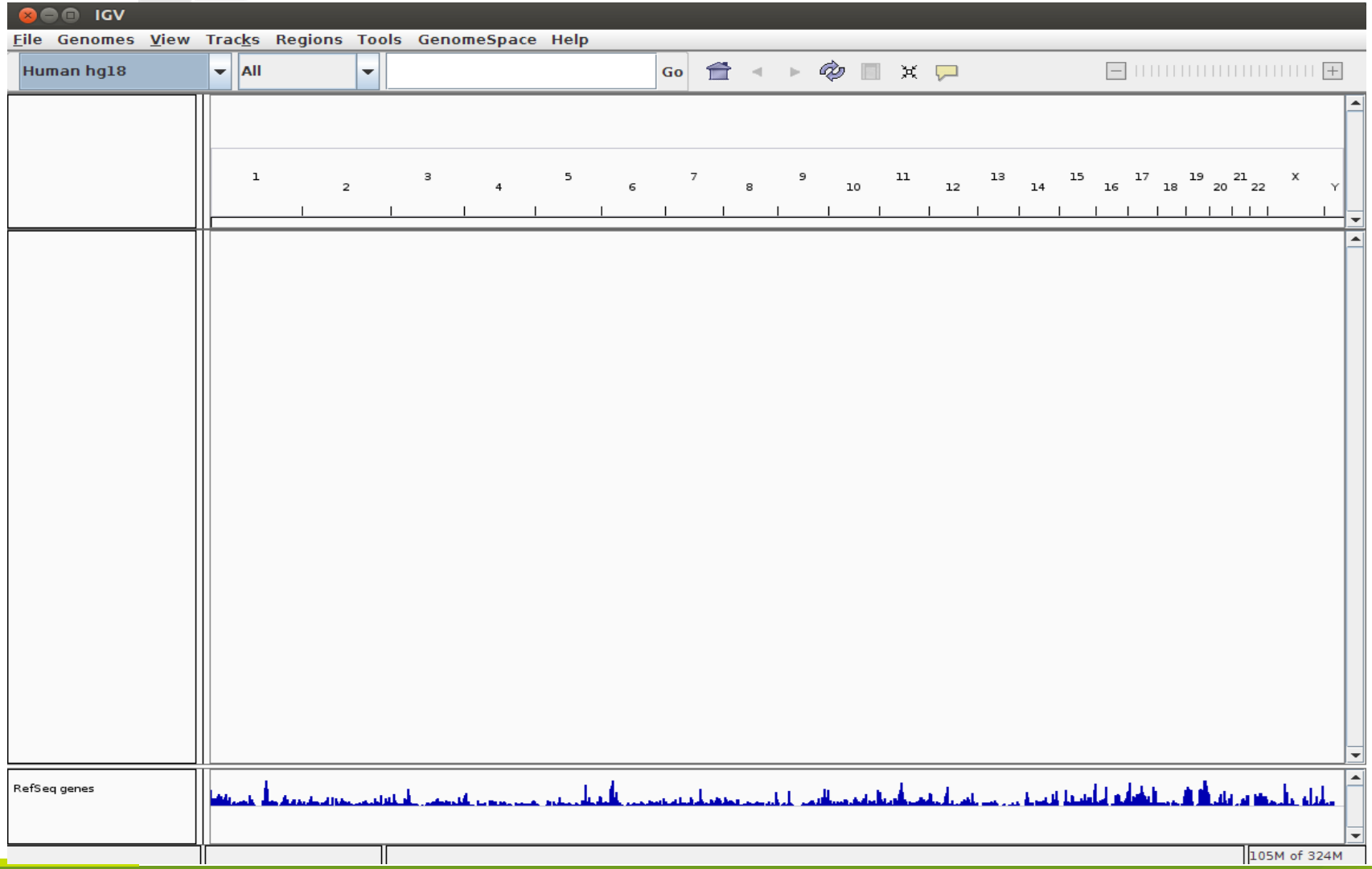
Visualisation des alignements avec IGV

- High-performance visualization tool
- Interactive exploration of large, integrated datasets
- Supports a wide variety of data types
- Documentations
- Developed at the Broad Institute of MIT and Harvard

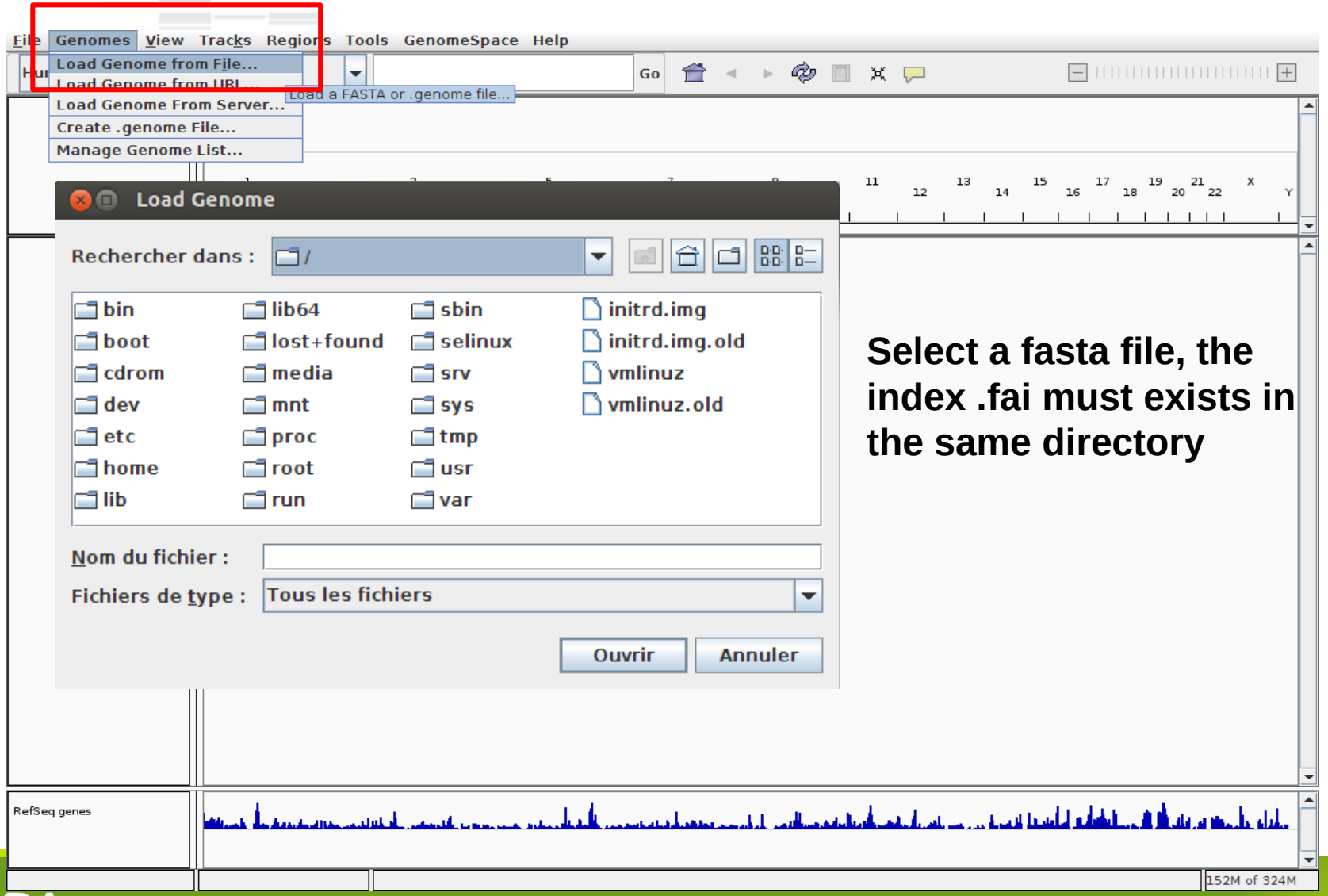

File Formats

- File Extension Identifies Format
- Recommended File Formats
- BAM
- BED
- CBS
- CN
- Cytoband
- FASTA
- GCT
- genePred
- GFF
- GISTIC
- HDF5
- IGV
- LOH
- Birdsuite Files
- MUT
- RES
- SAM
- Sample Information
- SEG
- SNP
- TAB
- TDF
- Track Line
- Type Line
- WIG

Visualisation des alignements avec IGV

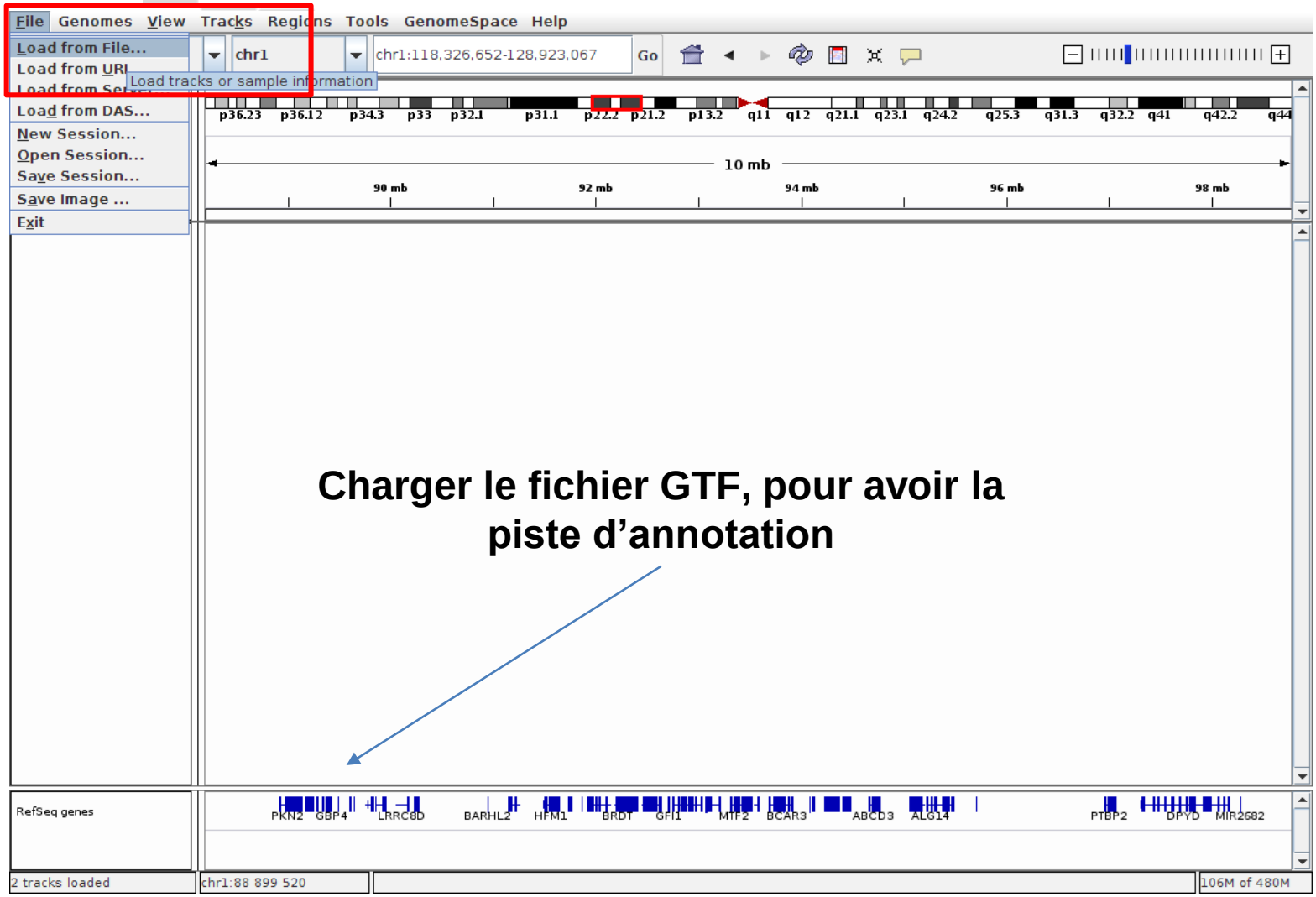


IGV : Chargement de la référence

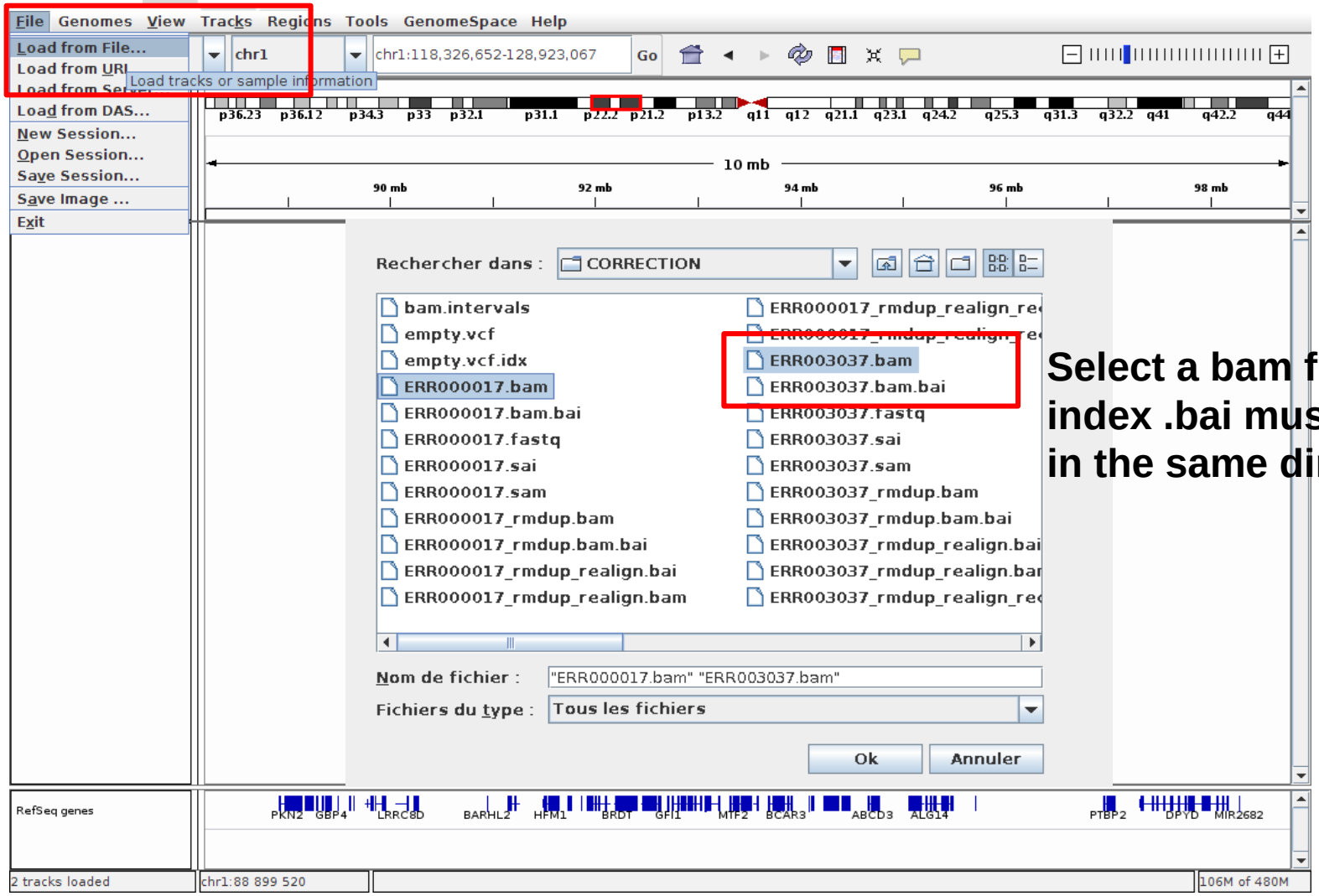


Select a fasta file, the index .fai must exists in the same directory

IGV : Chargement de l'annotation

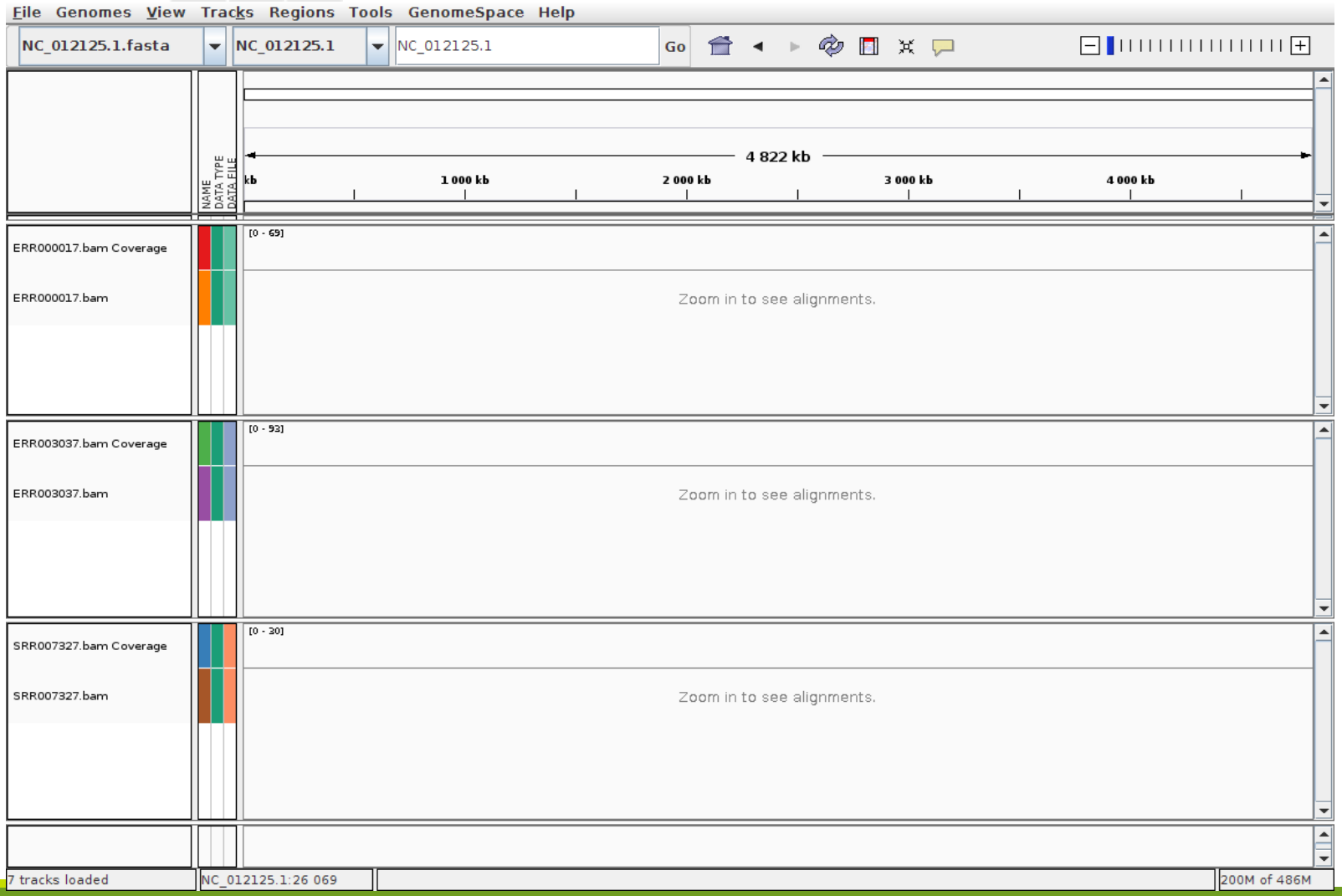


IGV : Chargement des alignements

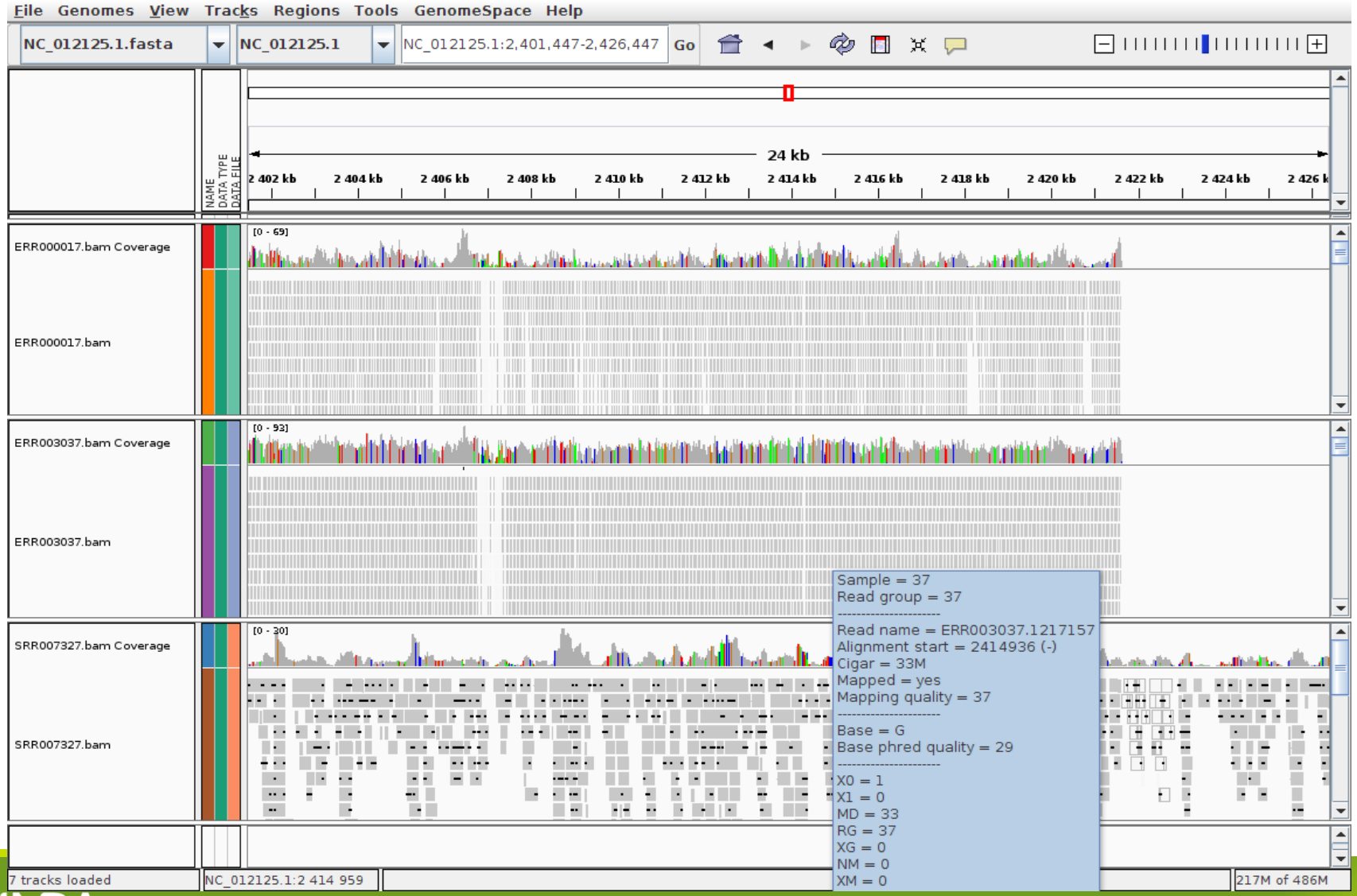


Select a bam file, the index .bai must exists in the same directory

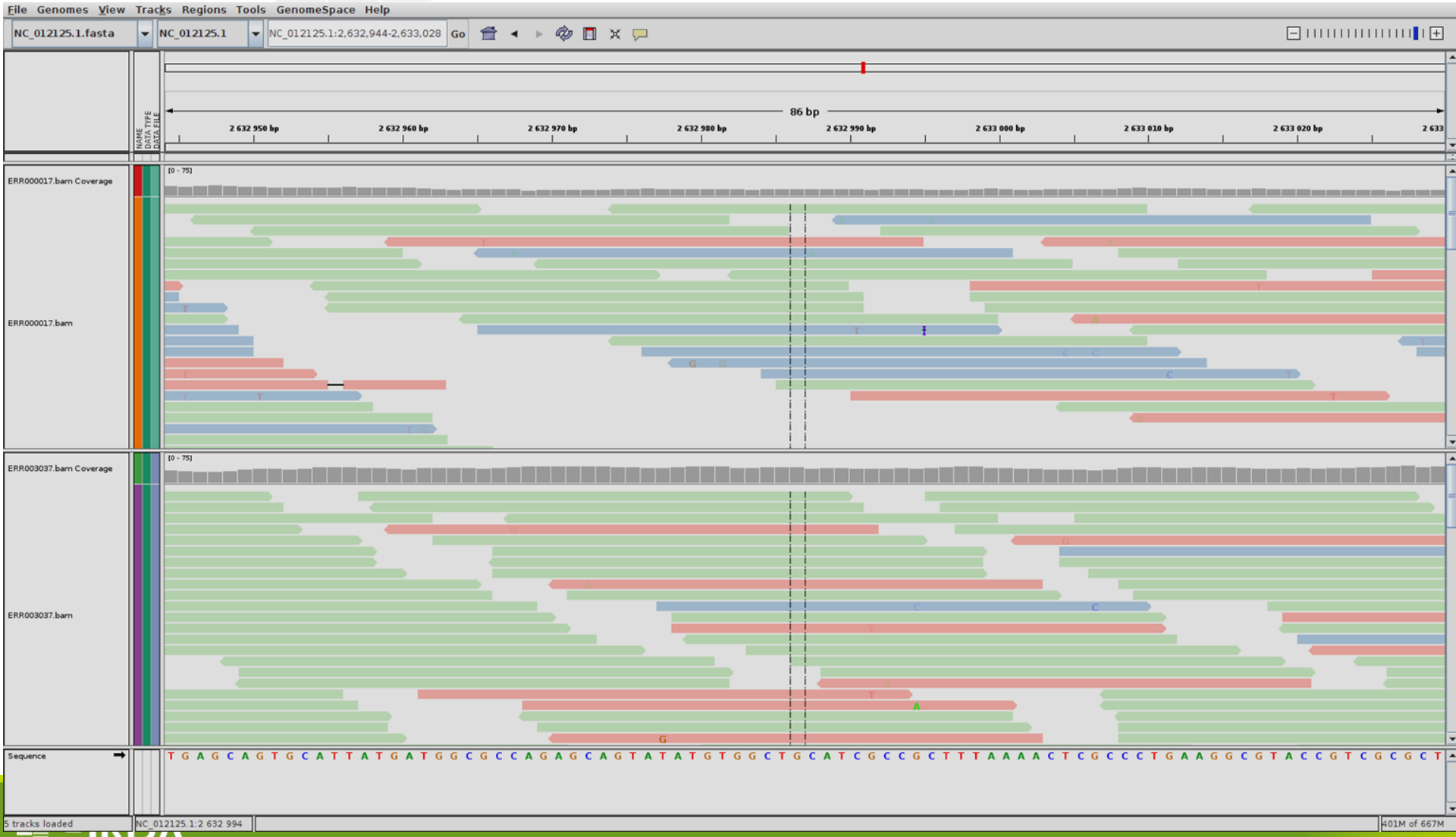
IGV : Chargement des alignements



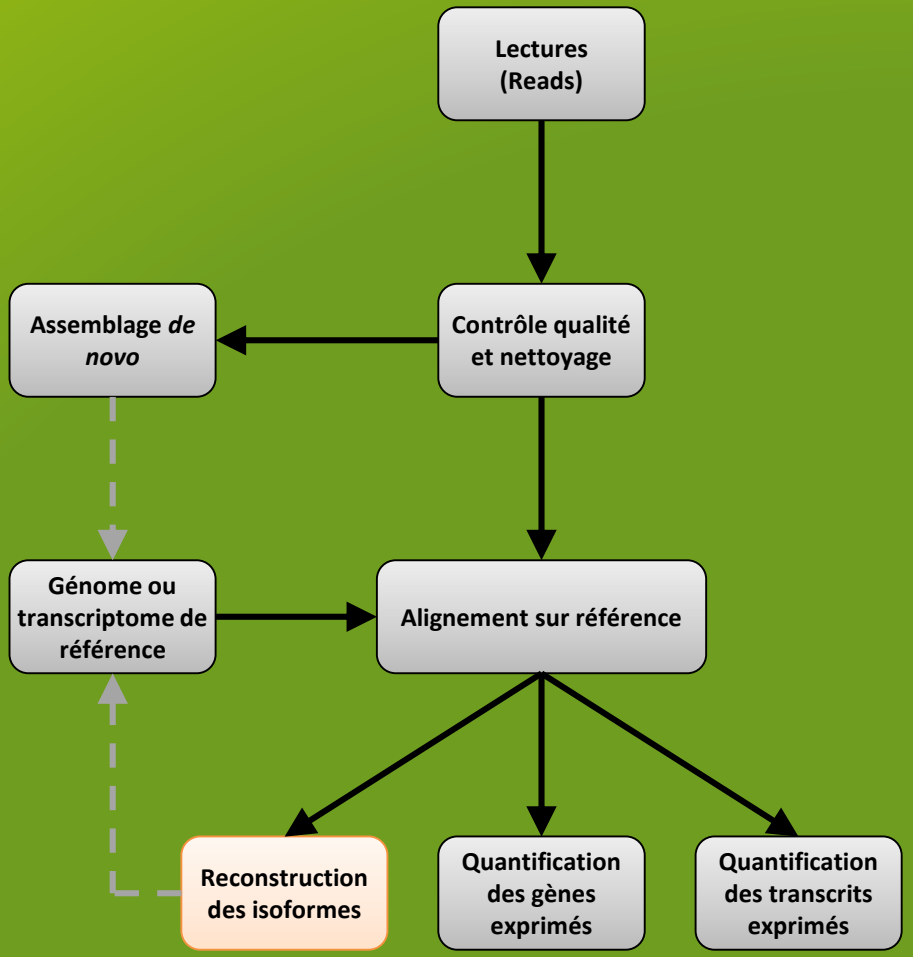
IGV : Chargement des alignements



IGV : Chargement des alignements



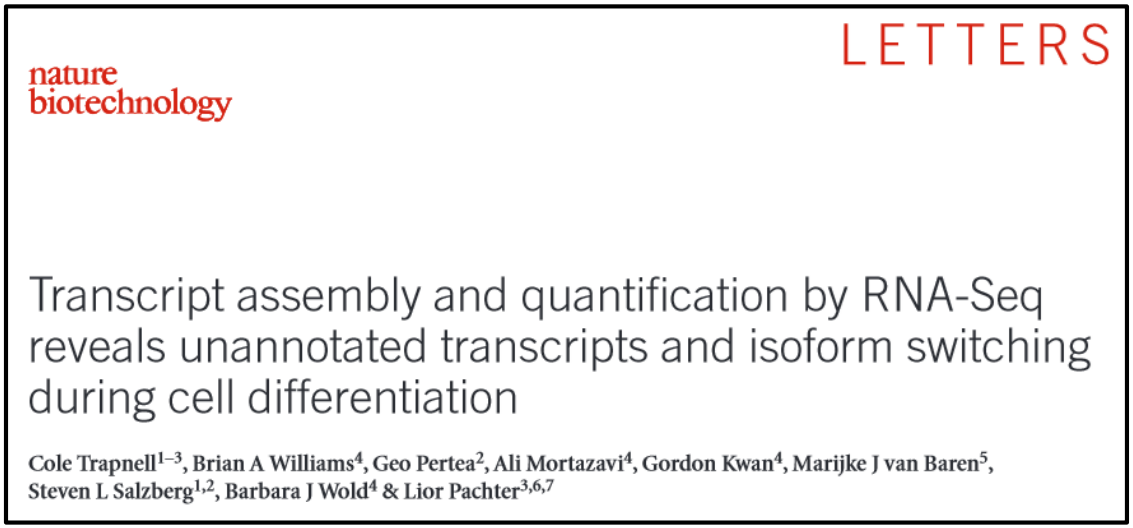
TP : Visualisation avec IGV



_04 Reconstruction de transcrit

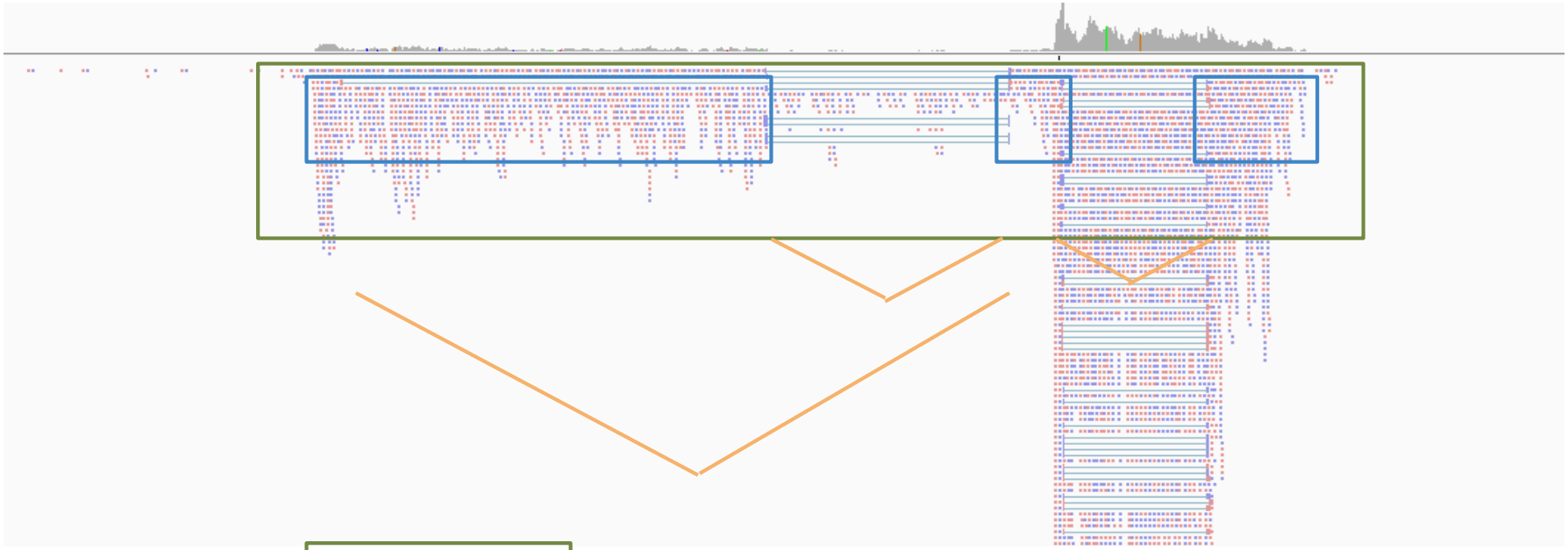
Cufflinks

- **Pipeline / suite logiciel** de **traitement RNA-Seq** :
  - **assemble** les **transcrits** (cufflinks)
  - **quantifie** l'**abondance** des transcrits (cufflinks)
  - **compare** les **annotations** des transcrits (cuffcompare)
  - **analyse** l'**expression différentielle** des transcrits (cuffdiff)



http://cufflinks.cbcb.umd.edu/

The element of the model



Gene :

Exons :

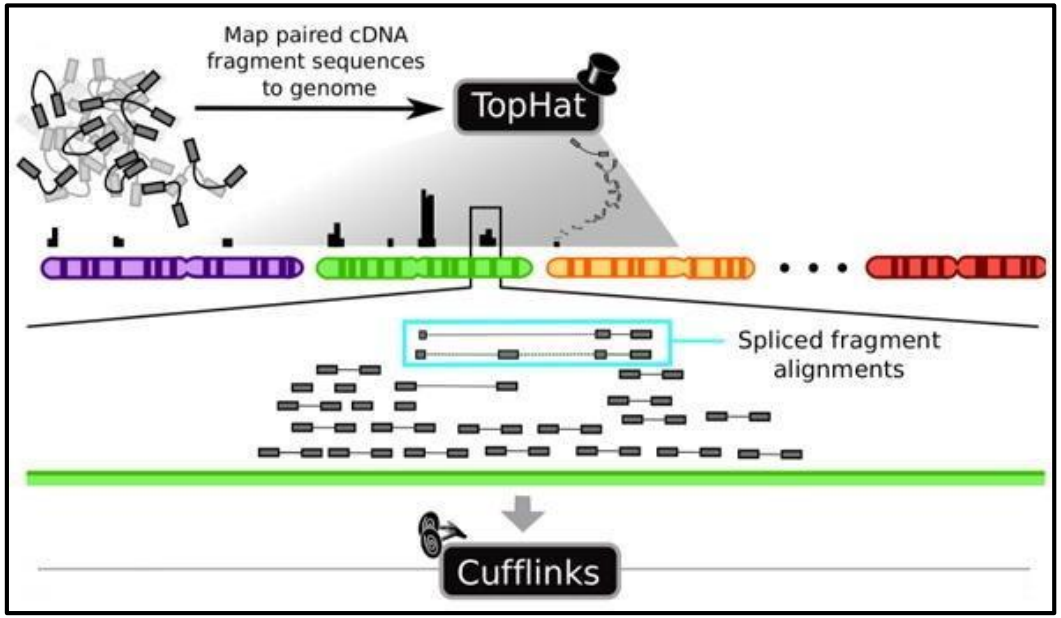
Jonctions (dans les paires & les Reads)

Cufflinks

Reconstruction de **transcrits**

- Fragments divisés en ***loci*** **non chevauchants**

- Chaque ***locus*** est **assemblé indépendamment**



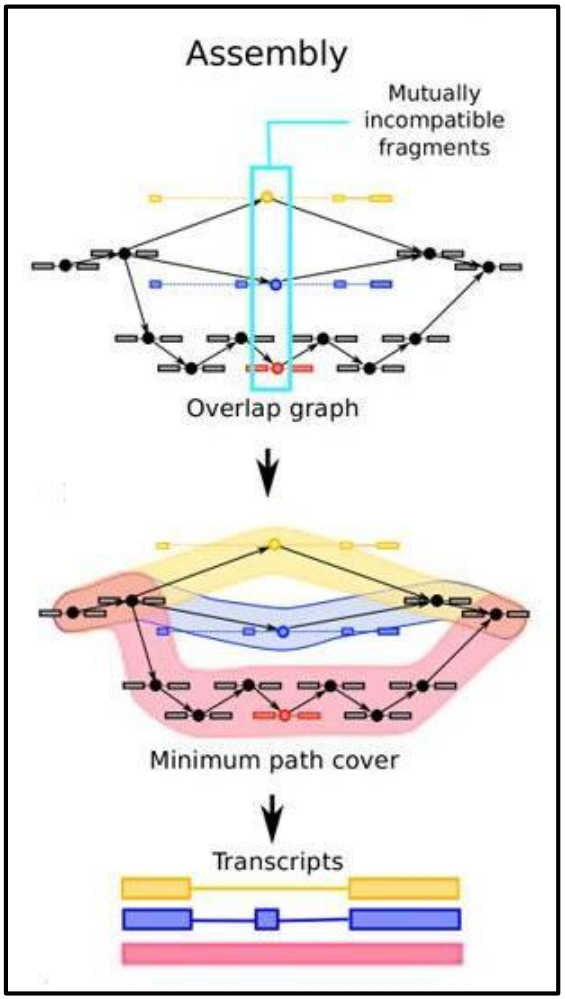
Trapnell et al. Nat Biotechnol. 2010

Cufflinks

Reconstruction de **transcrits**

- **Les différents chemins** :
  - trouver les **positions** des **gènes**
  - trouver les **exons**
  - trouver les **jonctions :**
    - **entre les paires**
    - **dans les séquences**

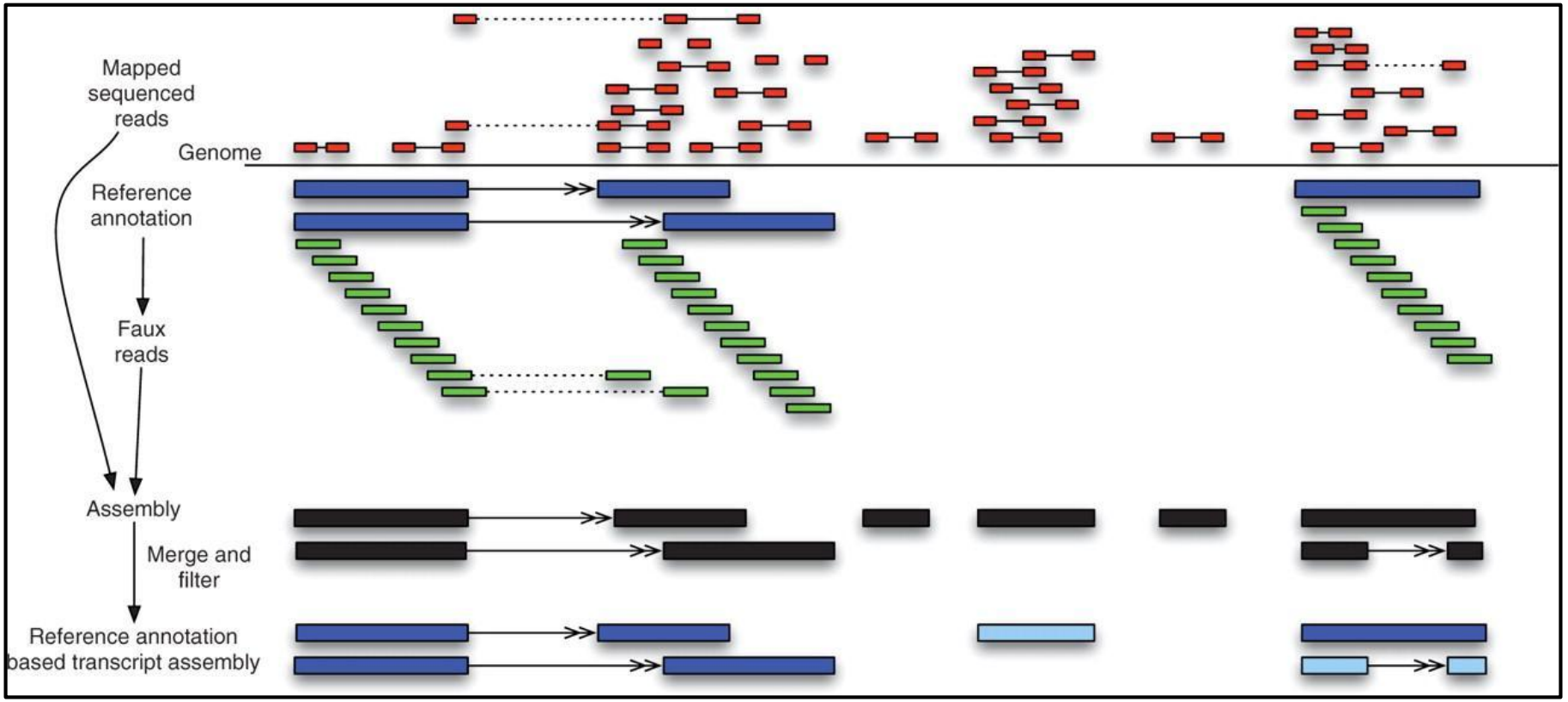
- **Stratégie** de **construction** du **modèle** :
  - trouver le **nombre minimum de modèles qui expliquent les lectures** :
    - **minimum de chemins**, théorème de Dilworth
    - **nb** de **lectures incompatibles**
      = **nb minimum** de **transcrits** nécessaires
    - **1 chemin = 1 isoforme**



Trapnell et al. Nat Biotechnol. 2010

Cufflinks

Reference Annotation Based Transcripts Assembly



Roberts et al. Bioinformatics 2011

Cufflinks

- Reference fasta (génome)
- Référence gtf (transcriptome)
- 1 bam par échantillon
- Quelles sont les stratégies possibles pour identifier le **maximum** de transcrits ?

Cufflinks

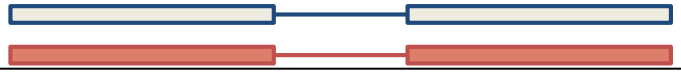
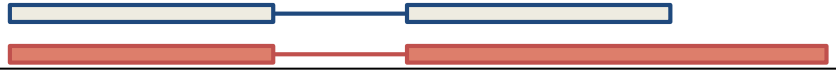
Reconstruction des **transcrits**

- En **entrée** :
  - **lectures** (**.sam/.bam**)
  - Use guide transcript assembly : **annotations** (**.gtf**)

- En **sortie** :
  - **transcrits** (**.gtf**) :
    - positionnement et quantification des isoformes
  - **gènes** (**.fpkm_tracking**) :
    - F/RPKM des gènes
  - **isoformes** (**.fpkm_tracking**) :
    - F/RPKM des isoformes

Cufflinks - Cuffcompare

Class code de cuffcompare

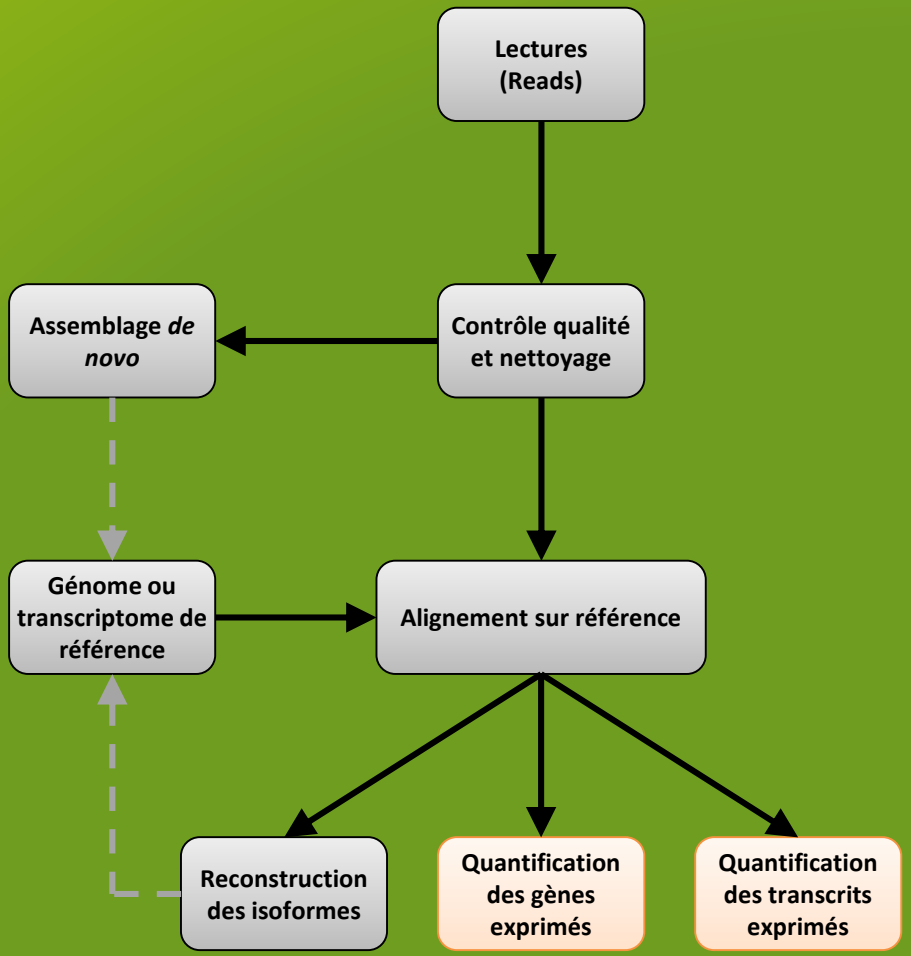
=	identité	
c	inclus	
j	nouvel isoforme	
e	exon	
i	intron	
o	chevauchant	
p	polymerase run-on	
r	répétition	
u	autre	
x	exon antisens	
s	intron antisens	

http://cufflinks.cbcb.umd.edu/manual.html#class_codes

TP - Découverte de transcrit

- Fusionner les alignements (samtools merge)
- Supprimer les duplicats (samtools rmdup)
- Détecter les nouveaux transcripts (cufflinks)
- Ouvrir le nouveau transcriptome dans IGV

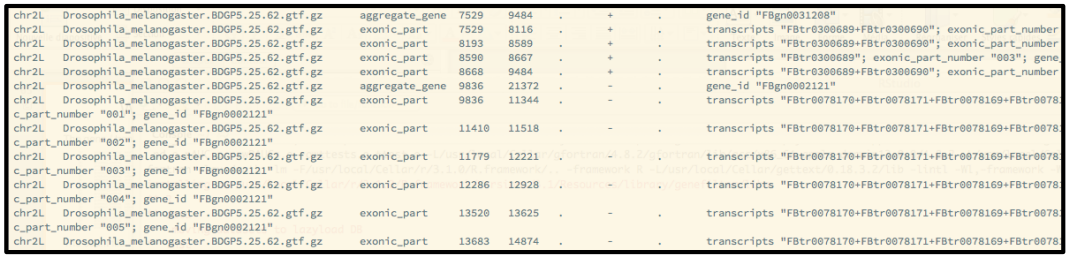
05 Quantification



Quantification

Que cherche-t-on à compter ?

- Quel *feature* compter ?
  - **gènes**
  - **exons**
  - **transcrits**



- **Comptage brut** sur les **gènes** ou les **exons** :
  - **htseq-count**
- **Comptage brut** sur les **gènes** ou les **exons** ou les **transcrits**:
  - **featureCount**
- **Estimation** de l'**abondance** des **transcrits reconstruits** :
  - **Cufflinks**

- Dépend des données disponibles

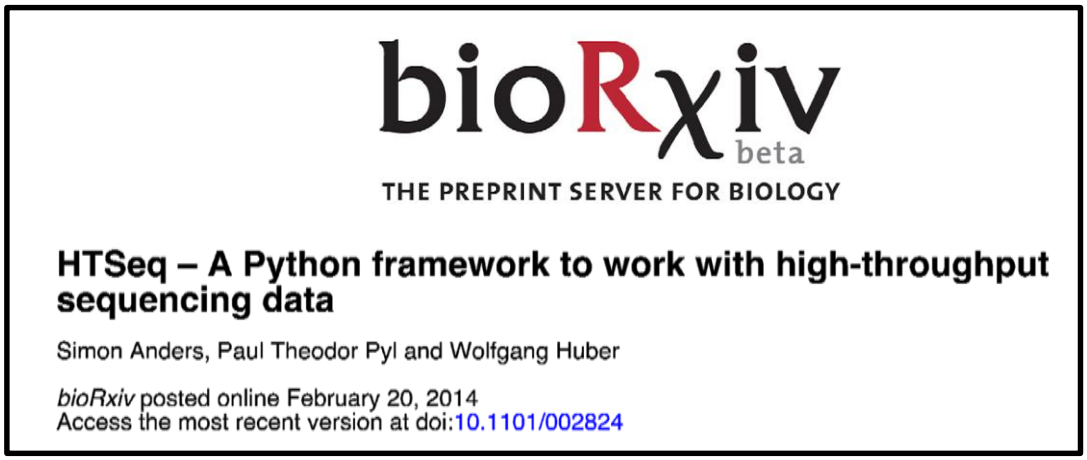
gene_id	untreated1	untreated2	untreated3	untreated4	treated1	
FBgn0000003	0	0	0	0	0	1
FBgn0000008	92	161	76	70	140	88	70
FBgn0000014	5	1	0	0	4	0	0
FBgn0000015	0	2	1	2	1	0	0
FBgn0000017	4664	8714	3564	3150	6205	3072	3334
FBgn0000018	583	761	245	310	722	299	308
FBgn0000022	0	1	0	0	0	0	0
FBgn0000024	10	11	3	3	10	7	5
FBgn0000028	0	1	0	0	0	1	1
FBgn0000032	1446	1713	615	672	1698	696	757

Comptage des gènes et exons

HTSeq-count

- **Comptage** des **lectures** s'alignant sur une *feature* donnée :
  - gène
  - exon

- **Utilise** les fichiers d'**alignement** (SAM/BAM) et une **annotation**


bioRxiv beta

THE PREPRINT SERVER FOR BIOLOGY

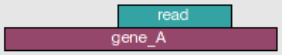
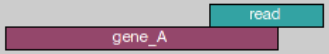
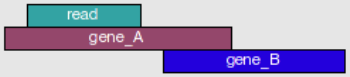
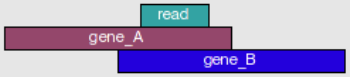
HTSeq – A Python framework to work with high-throughput sequencing data

Simon Anders, Paul Theodor Pyl and Wolfgang Huber

bioRxiv posted online February 20, 2014
Access the most recent version at doi:10.1101/002824

http://www-huber.embl.de/users/anders/HTSeq/doc/count.html#count

HTSeq-count

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

```
Usage: htseq-count [options] alignment_file gff_file

This script takes an alignment file in SAM/BAM format and a feature file in
GFF format and calculates for each feature the number of reads mapping to it.
See http://www-huber.embl.de/users/anders/HTSeq/doc/count.html for details.

Options:
  -h, --help            show this help message and exit
  -f SAMTYPE, --format=SAMTYPE
                        type of <alignment_file> data, either 'sam' or 'bam'
                        (default: sam)
  -r ORDER, --order=ORDER
                        'pos' or 'name'. Sorting order of <alignment_file>
                        (default: name). Paired-end sequencing data must be
                        sorted either by position or by read name, and the
                        sorting order must be specified. Ignored for single-
                        end data.
  -s STRANDED, --stranded=STRANDED
                        whether the data is from a strand-specific assay.
                        Specify 'yes', 'no', or 'reverse' (default: yes).
                        'reverse' means 'yes' with reversed strand
                        interpretation
  -a MINAQUAL, --minaqual=MINAQUAL
                        skip all reads with alignment quality lower than the
                        given minimum value (default: 10)
  -t FEATURETYPE, --type=FEATURETYPE
                        feature type (3rd column in GFF file) to be used, all
                        features of other type are ignored (default, suitable
                        for Ensembl GTF files: exon)
  -i IDATTR, --idattr=IDATTR
                        GFF attribute to be used as feature ID (default,
                        suitable for Ensembl GTF files: gene_id)
  -m MODE, --mode=MODE  mode to handle reads overlapping more than one feature
                        (choices: union, intersection-strict, intersection-
                        nonempty; default: union)
  -o SAMOUT, --samout=SAMOUT
                        write out all SAM alignment records into an output SAM
                        file called SAMOUT, annotating each line with its
                        feature assignment (as an optional field with tag
                        'XF')
  -q, --quiet           suppress progress report
```

HTSeq-count

fichiers de sortie

- Une **table de comptage** pour chaque *feature* ainsi qu'un **résumé**
  - **__no_feature** : lectures non assignées
  - **__ambiguous** : lectures assignables à plus d'un feature, non comptées
  - **__too_low_aQual** : lectures filtrées sur la qualité d'alignement (-a)
  - **__not_aligned** : lectures non alignées du fichier d'entrée
  - **__alignement_not_unique** : lectures avec alignement multiple (BAM)

```
FBgn0264694             246
FBgn0264706             0
FBgn0264712             251
__no_feature            65531
__ambiguous             11617
__too_low_aQual         0
__not_aligned           0
__alignment_not_unique          1967339
```

featureCounts

- htseq-count ++.
- Niveau exon, gène, transcrit.
- 1 read peut être attribué à plusieurs Feature.
- Reads avec alignement multiples peuvent être pris en compte.
- Brin-spécifique très bien géré.
- 2 Notions :

  - *feature* (e.g. exon)

  - *meta-feature* : agrégation de feature (e.g. gene)

featureCounts options

Feature Counts (version 1.0.0)

Your annotation file (gtf file):

39: Cufflinks on merged: assembled transcripts

Give the name of the annotation file. The program assumes that the provided annotation file is in GTF format. Use -F option to specify other annotation formats.

First SAM/BAM file:

29: {WT_rep1_1_Ch6.fastq}-Tophat_mapped.bam

Give the names of input read files that include the read mapping results. Format of input files is automatically determined (SAM or BAM). Paired-end reads will be automatically re
each other. Multiple files can be provided at the same time.

Add another BAM/SAM datasets

Add another BAM/SAM dataset 1

Other SAM/BAM files:

32: {MT_rep1_1_Ch6.fastq}-Tophat_mapped.bam

Remove Add another BAM/SAM dataset 1

Add new Add another BAM/SAM dataset

Specify feature type:

exon

Only rows which have the matched matched feature type in the provided GTF annotation file will be included for read counting. `exon' by default

Specify the attribute type used to group features (eg. exons) into meta-features (eg. genes), when GTF annotation is provided:

gene_id

Reads will be allowed to be assigned to more than one matched meta-feature:

Yes

Indicate if strand-specific read counting should be performed:

unstranded

Multi-mapping reads/fragments will be counted:

Yes

Only primary alignments will be counted:

Yes

inimum number of overlapped bases required to assign a read to a feature:

30

Negative values are permitted, indicating a gap being allowed between a read and a feature.

Optional paired-end parameters:

Paired-end reads

featureCounts : options

Multi-mapping reads/fragments will be counted:

Yes

Only primary alignments will be counted:

Yes

inimum number of overlapped bases required to assign a read to a feature:

30

Negative values are permitted, indicating a gap being allowed between a read and a feature.

Optional paired-end parameters:

Paired-end reads

Fragments (or templates) will be counted instead of reads. The two reads from the same fragment must be adjacent to each other in the provided SAM/BAM file:

Fragments NOT counted instead of reads

Paired-end distance will be checked when assigning fragments to meta-features or features:

Paired-end distance will NOT be checked.

Minimum fragment/template length:

50

Minimum fragment/template length, 50 by default.

Maximum fragment/template length:

600

Maximum fragment/template length, 600 by default.

If specified, only fragments that have both ends successfully aligned will be considered for summarization:

Not only fragments with both ends successfully aligned

If specified, the chimeric fragments (those fragments that have their two ends aligned to different chromosomes) will NOT be included for summarization:

The chimeric fragments will NOT be included

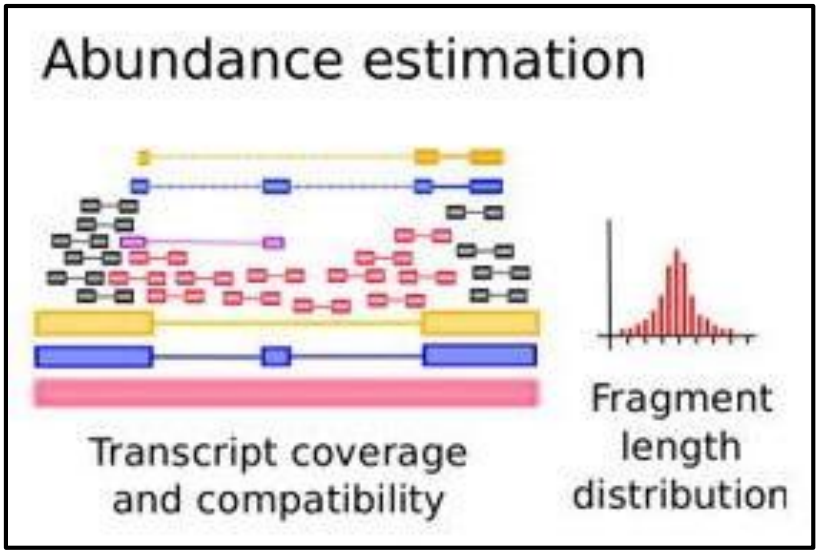
Execute

Cufflinks

Principes

- **Assignation des lectures** à un transcrit

- **Estimation** de l'**abondance** de **chaque transcrit** mesurée en :
  - **RPKM** (*single reads*)
  - **FPKM** (*paired-end reads*)



Trapnell et al. Nat Biotechnol. 2010

Cufflinks

RPKM / FPKM

- Permet de corriger les **biais de longueur** des transcrits

- **RPKM** :
  - **R**eads **P**er **K**ilobase of exon per **M**illion fragments mapped :

  R = Nombre de read mappés

  N = Nombre total de read de la librairie

  L = taille des exons du gène en bp

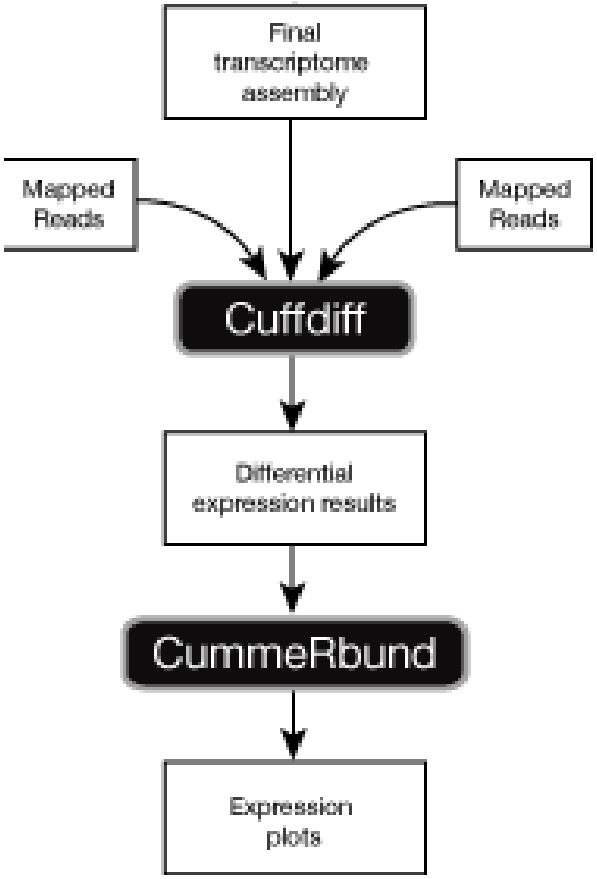
$$RPKM = \frac{10^9 \times R}{N \times L}$$

- **FPKM :**
  - **F**ragments **P**er **K**ilobase of exon per **M**illion fragments mapped
  - **1 paire de lecture = 1 fragment**

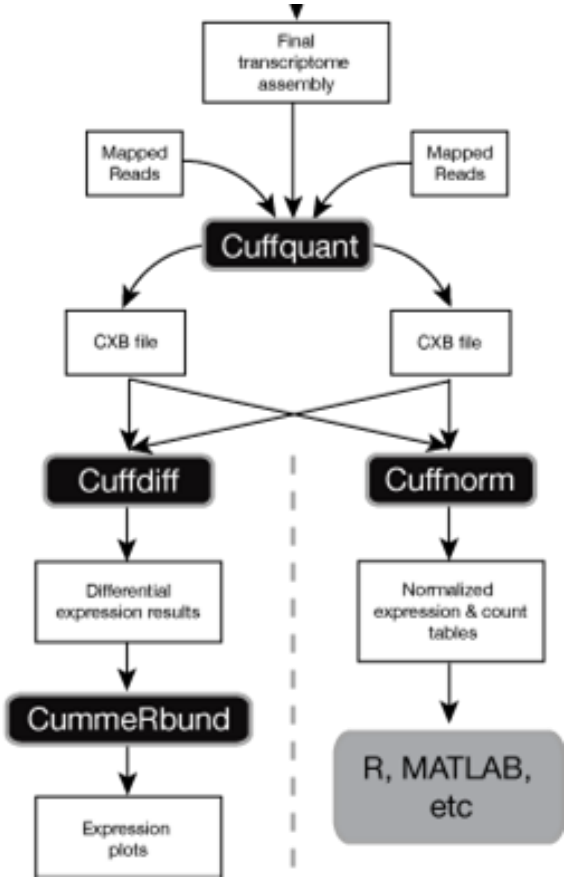
Mortazavi et al. Nature Methods 2008

Cufflinks - Estimation de l'abondance

<2.2.0



>=2.2.0



Cufflinks

- En **entrée** :
  - transcriptome de référence
  - bam d'un échantillon
  - mode : assemble ou quantifie ?

- En **sortie** :
  - **assembled transcripts** (**.gtf**) :
    - positionnement et quantification des isoformes
  - **gene expression** (**.fpkm_tracking**) :
    - F/RPKM des gènes
  - **transcript expression** (**.fpkm_tracking**) :
    - F/RPKM des isoformes

Cufflinks outputs

Description du format GTF

- **transcripts.gtf** :
  - **coordonnées** et **abondance** des **isoformes**
  - **annotations** (**.gtf**)

- **Score :**
  - **le plus abondant = 1000**
  - **moins abondant : ratio = minor FPKM / major FPKM**

3R	Cufflinks	transcript	30207	41011	901	-	.
3R	Cufflinks	exon	30207	31507	901	-	.
3R	Cufflinks	exon	40845	41011	901	-	.

gene_id "FBgn0037217"; transcript_id "FBtr0111206"; FPKM "3.7298141180"; frac "0.003967"; co
gene_id "FBgn0037217"; transcript_id "FBtr0111206"; exon_number "1"; FPKM "3.7298141180"; f
gene_id "FBgn0037217"; transcript_id "FBtr0111206"; exon_number "2"; FPKM "3.7298141180"; f

"; conf_lo "0.000000"; conf_hi "7.855077"; cov "0.330100";
30"; frac "0.003967"; conf_lo "0.000000"; conf_hi "7.855077"; cov "0.330100";
30"; frac "0.003967"; conf_lo "0.000000"; conf_hi "7.855077"; cov "0.330100";

Cufflinks outputs

- Isoform estimation
- Gene estimation
- Format :

tracking_id	class_code	nearest_ref_id	gene_id	gene_short_name	tss_id	locus
CUFF.1.1	-	-	CUFF.1	-	-	2L:5374-5556
CUFF.2.1	-	-	CUFF.2	-	-	2L:5382-5555
CUFF.3.1	-	-	CUFF.3	-	-	2L:5707-5856

length	coverage	FPKM	FPKM_conf_lo	FPKM_conf_hi	FPKM_status
182	351.21	12.2989	8.54337	16.0544	OK
173	121.069	4.23967	1.69117	6.78817	OK
149	620.192	21.7182	13.0174	30.419	OK

TP : Quantification

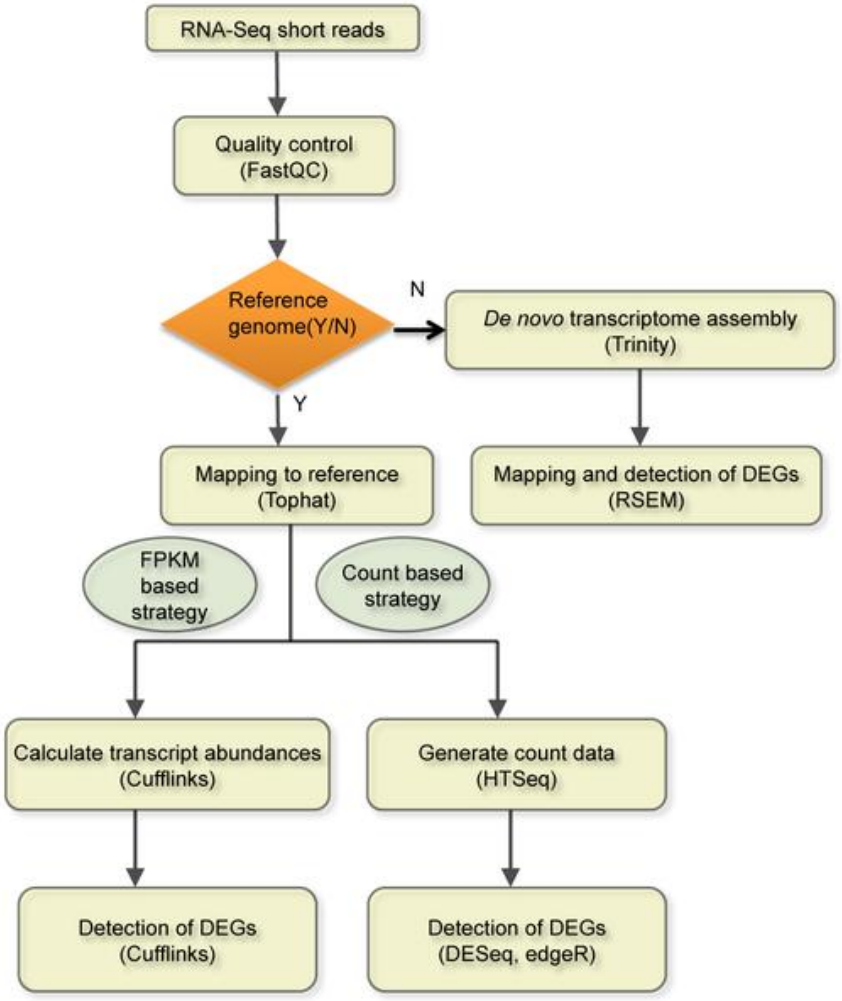
06

Conclusion

Conclusion générale

- Workflow galaxy à construire
- Choix des outils dépendent des données disponibles et de la question biologique
- Tous les outils sont dispo sur Migale et Galaxy
- Et maintenant en avant pour les stats !

Figure 1. The workflow of differential expression analysis for RNA-Seq data.



Zhang ZH, Jhaveri DJ, Marshall VM, Bauer DC, Edson J, et al. (2014) A Comparative Study of Techniques for Differential Expression Analysis on RNA-Seq Data. PLoS ONE 9(8): e103207. doi:10.1371/journal.pone.0103207
http://127.0.0.1:8081/plosone/article?id=info:doi/10.1371/journal.pone.0103207

Liens utiles

- **Seqanswer** : *http://seqanswers.com/*
- **Biostar** : *https://www.biostars.org/*
- **RNA-Seq blog** : *http://rna-seqblog.com/*

Remerciements

- Le groupe de travail **« Planification d'experiences et RNA-seq »** du **PEPI IBIS**

Satisfaction form :
http://bioinfo.genotoul.fr/index.php?id=79