

# TRAITEMENT BIOINFORMATIQUE DE DONNÉES RNA-Seq

[http://genoweb.toulouse.inra.fr/~formation/4\\_Galaxy\\_RNAseq/2018/](http://genoweb.toulouse.inra.fr/~formation/4_Galaxy_RNAseq/2018/)



# Formateurs



- **Sarah Maman**
- **Céline Noirot**
- **Matthias Zytnicki**

# Plan



- ❖ First day
  - 09:00 am to 12:00 am : Galaxy initiation
  - 13:30 pm to 17:00 pm : RNAseq quality control and files formats
  
- ❖ Second day :
  - 09:15 am to 12:00 am : Splicing alignment and visualisation
  - 13:30 pm to 17:00 pm : Discover new transcript and quantification
  
- ❖ Third day : 09:15 am to 12:00 am
  - Statistics analysis with SARtools

# **\_01 Initiation galaxy**

# **\_02 Rappels biologiques**

# Rappels biologiques



**Qu'est-ce qu'un gène ?**

# Rappels biologiques

## Qu'est-ce qu'un gène ?

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



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

# Rappels biologiques



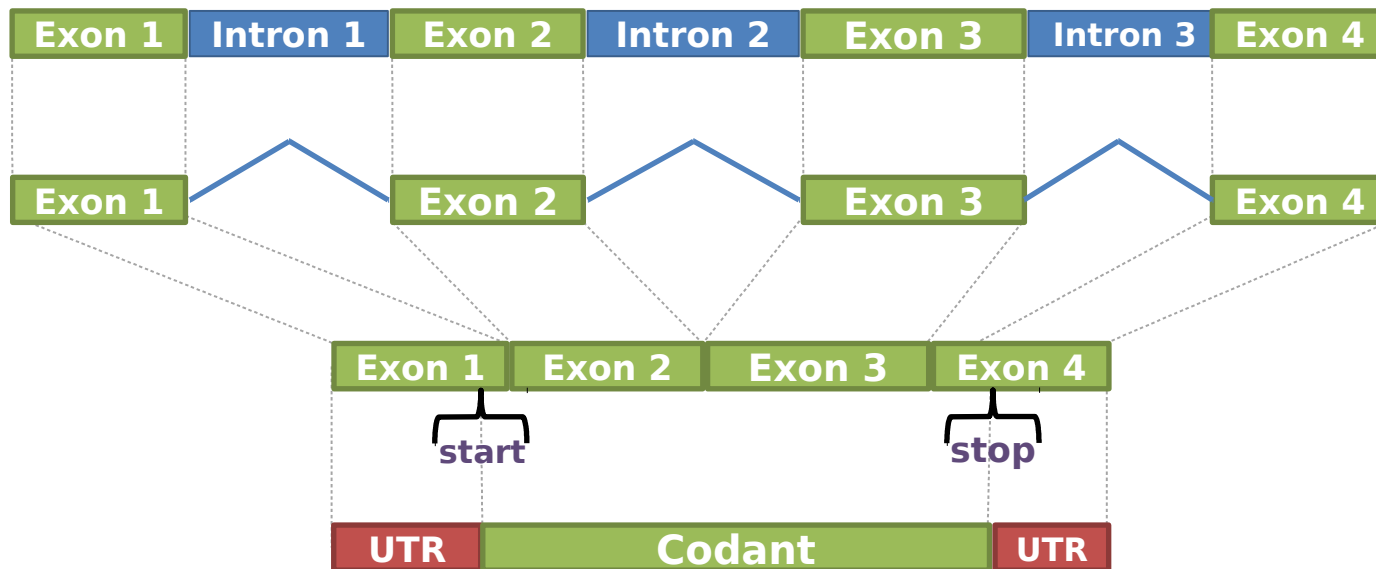
## Qu'est-ce qu'un transcrit?



# Rappels biologiques

## Qu'est-ce qu'un transcrit ?

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



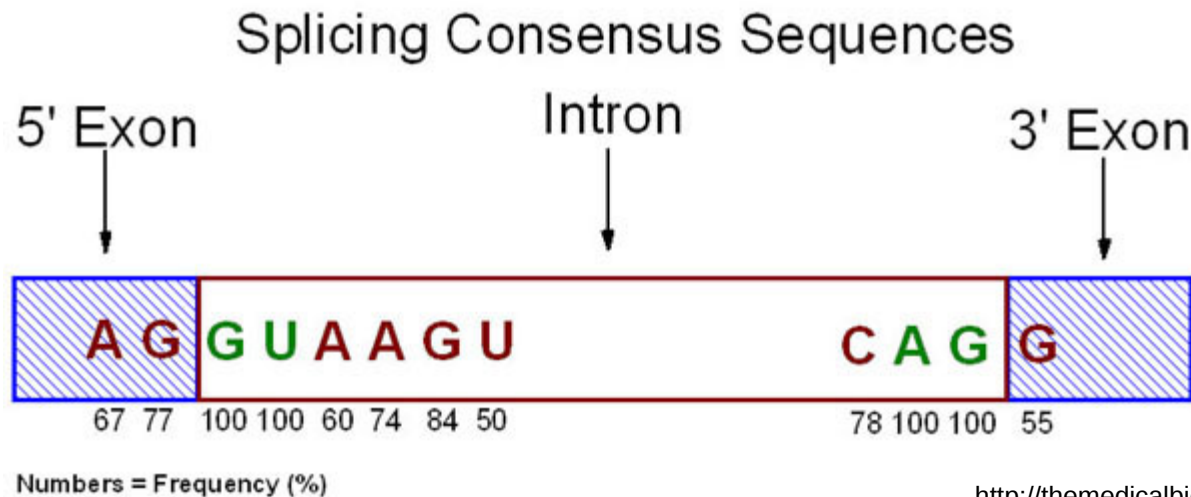
- o **Transcrit** : portion d'ADN transcrite en molécule d'ARN
- o **UTR** : région transcrite mais pas traduite

# Rappels biologiques

## Qu'est-ce qu'un site d'épissage?

### o Site d'épissage canonique :

- plus de **99%** de **GT** et **AG** comme sites **donneurs** et **accepteurs**

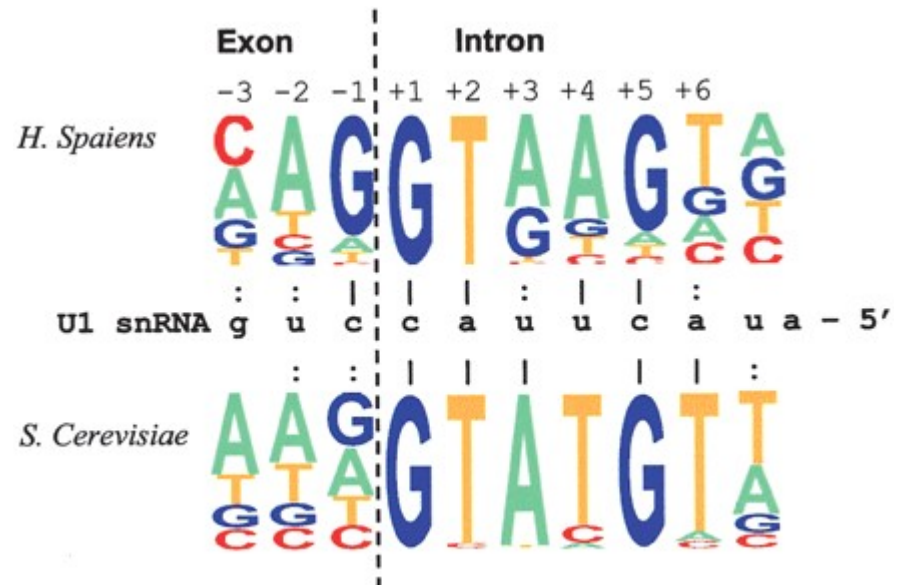


<http://themedicalbiochemistrypage.org/rna.php>

# 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 :**

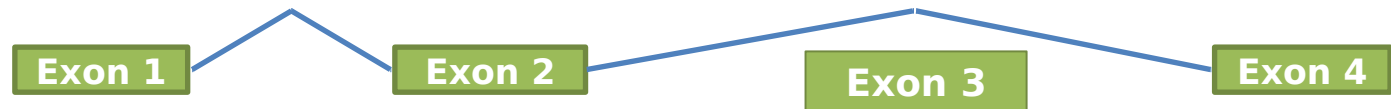


<http://rnajournal.cshlp.org/content/10/5/828.full>

# Rappels biologiques

## Epissage alternatif et isoformes

o Excision d'exon



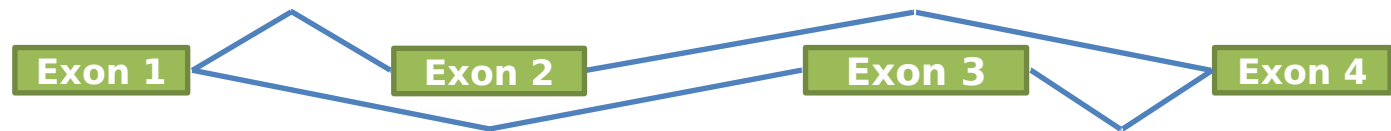
o Rétention d'intron



o TSS alternatif



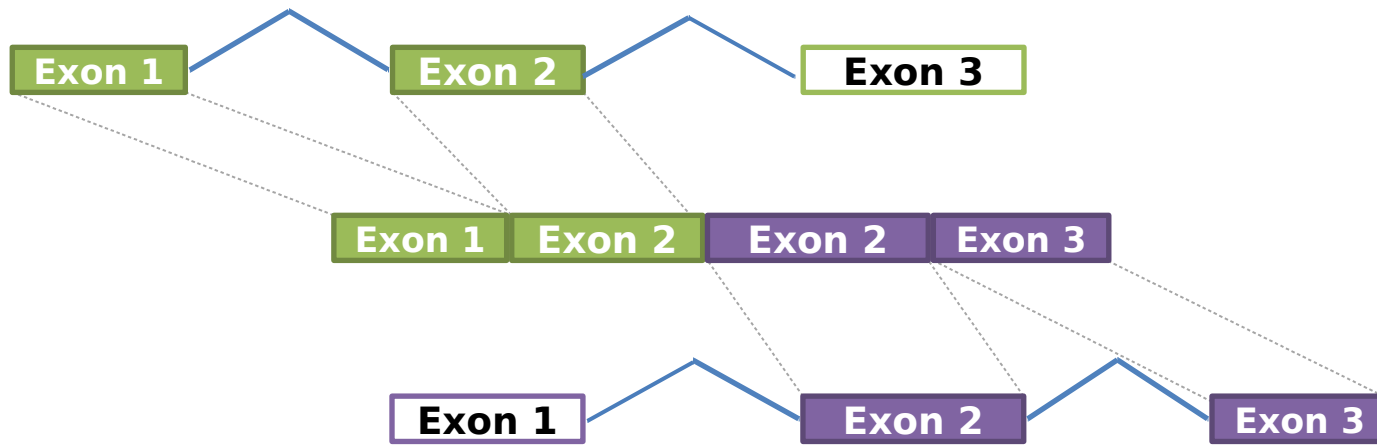
o Exons exclusifs



# Rappels biologiques

## Et plus encore ?

### o Fusion de gènes ou Trans-épissage

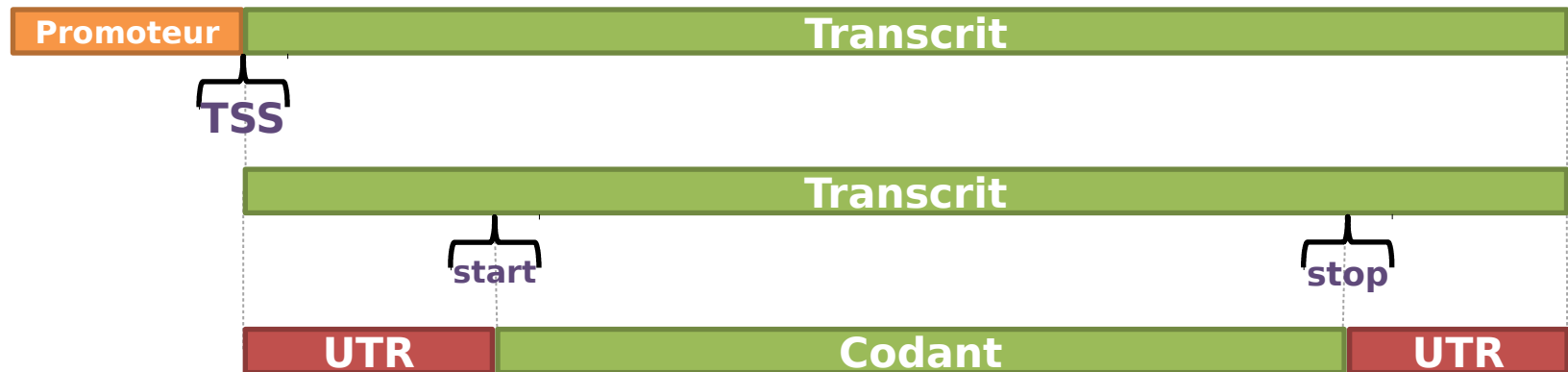


### o Chimère biologique

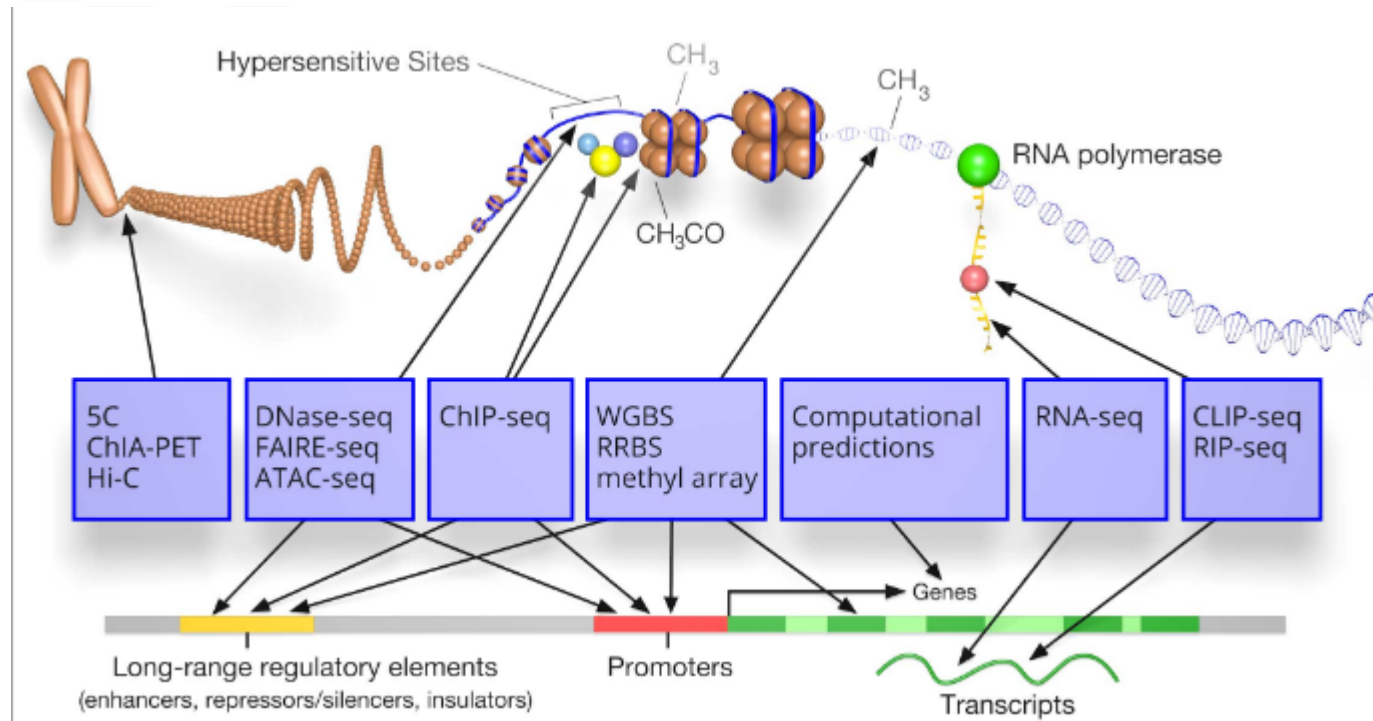
# Rappels biologiques

## Gène procaryote / gène eucaryote

o Pas d'intron chez les procaryotes



# Etude des éléments fonctionnels du génome



<https://www.encodeproject.org/>

Référence de l'ensemble des protocoles : <http://enseqlopedia.com/enseqlopedia/>



# 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 allélique, variant de structure
- ❖ Découverte de nouveaux : transcrits, isoformes, (ncRNA), structures (fusion...)
- ❖ Détection possible of SNPs, ...

# Les séquenceurs :

Platform	Reads x run: (M)	Read length: (paired-end*, Half of data in reads**)	Run time: (d)	Yield: (Gb)	Rate: (Gb/d)	Reagents: (\$K)	per-Gb: (\$)	hg-30x: (\$)	Machine: (\$)
ISeq 100 1fcell	4	150*	0.77	1.2	1.56	0.625	521	62500	19.9K
MiniSeq 1fcell	25	150*	1	7.5	7.5	1.75	233	28000	49.5K
MISeq 1fcell	25	300*	2	15	7.5	1	66	8000	99K
Next Seq 550 1fcell	400	150*	1.2	120	100	5	50	5000	250K
HiSeq 2500 RR 2fcells	600	100*	1.125	120	106.6	6.145	51.2	6144	740K
HiSeq 2500 V3 2fcells	3000	100*	11	600	55	23.47	39.1	4692	690K
HiSeq 2500 V4 2fcells	4000	125*	6	1000	166	29.9	31.7	3804	690K
HiSeq 4000 2fcells	5000	150*	3.5	1500	400	--	20.5	2460	900K
HiSeq X 2fcells	6000	150*	3	1800	600	--	7.08	849.6	1M
Nova Seq S1 2018 2fcells	3300	150*	1.66	1000	600	--	18	1800	999K
Nova Seq S2 2fcells	6600	150*	1.66	2000	1200	--	15	1564	999K
Nova Seq S4 2fcells	20000	150*	1.66	6000	3600	--	5.8	700	999K
5500 XL	1400	60	7	180	30	10.5	58.33	7000	595K
Ion S5 510 1chip	2 - 3	200 400	0.21	1	4.8	0.95	950	114000	65K
Ion S5 520 1chip	3 - 6	200 400 600	0.23	1	4.3	1	500	60000	65K
Ion S5 530 1chip	20	200 400 600	0.29	4	13.8	1.2	150	18000	65K
Ion S5 540 1chip	80	200	0.42	15	35.7	1.4	93.3	11196	65k
Ion S5 550 1chip	130	200	0.5	25	50	1.67	66.8	8016	65k
RSII P6-C4 16cells	0.88	20K**	4.3	12	2.8	2.4	200	24000	695K
Sequel 16cells 2018	6.4	33K**	6.6	160	24.2	--	80	9600	350K
R&D end 2018	--	32K**	--	192	--	1	6.6	1000	350K
Smidg ION RnD	--	--	--	4	--	--	--	--	--
Mini ION R9.5 1fcell	--	--	2	10-20	5-10	0.5 - 0.9	--	--	--
Grid ION X5 5fcells	--	--	2	50 - 100	25-50	1.5 - 4.5	--	--	125K
Prome thION RnD 48fcells	--	--	2	2400	1200	32.64	20	2400	75K
Prome thION R&D 48fcells	--	--	5	5760	1152	--	5	600	75K
QiaGen Gene Reader	400	--	--	80	--	0.5	--	--	--
BGI SEQ 500	1600	100*	7	260	37.1	--	--	600?	500K
BGI SEQ 50	1600	50*	0.4	8	20	--	--	--	--
MGI SEQ 2000	--	100*	2	600	300	4.8	8	960	310K
MIG SEQ 200	--	100*	--	60	--	--	--	--	150K

# Les protocoles NGS d'analyse du transcriptome

- ❖ RNA-seq : short-read on illumina

Encode directives: <https://www.encodeproject.org/rna-seq/small-rnas/>

- ❖ ISO-seq : long-read on pacbio

- ❖ ONT RNA-seq : long-read on MinION :

“Short-read RNAseq is limited in its ability to resolve complex isoforms because it fails to sequence full-length cDNA copies of RNA molecules. Here, we investigate whether RNAseq using the long-read single-molecule Oxford Nanopore MinION sequencer is able to identify and quantify complex isoforms without sacrificing accurate gene expression quantification.”

Received 24 Apr 2017 | Accepted 23 May 2017 | Published 19 Jul 2017

DOI: [10.1038/ncomms16027](https://doi.org/10.1038/ncomms16027)

OPEN

Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells

Ashley Byrne<sup>1,2</sup>, Anna E. Beaudin<sup>3,†</sup>, Hugh E. Olsen<sup>2,3</sup>, Miten Jain<sup>2,3</sup>, Charles Cole<sup>2,3</sup>, Theron Palmer<sup>3</sup>, Rebecca M. DuBois<sup>3</sup>, E. Camilla Forsberg<sup>3,4</sup>, Mark Akeson<sup>2,3</sup> & Christopher Vollmers<sup>2,3</sup>

Encode directives: <https://www.encodeproject.org/rna-seq/long-rnas/>

# Les données publiques

- ❖ Archive de short-read :
  - Données brutes de séquences
  - SRA <https://www.ncbi.nlm.nih.gov/sra>
  - ENA <https://www.ebi.ac.uk/ena>
- ❖ Gene Expression Omnibus
  - Données analysées (bed, peak, bigwig ...),
  - Lien vers SRA
  - <https://www.ncbi.nlm.nih.gov/geo/>
- ❖ Expression Atlas
  - Interface d'exploration de données publiques
  - <https://www.ebi.ac.uk/gxa/home>
- ❖ Genomes browser (Ensembl, UCSC, ...):
  - Offre la visualisation de données RNAseq publiques via options.

# A quelles questions biologiques PEUT répondre le RNA-seq ?

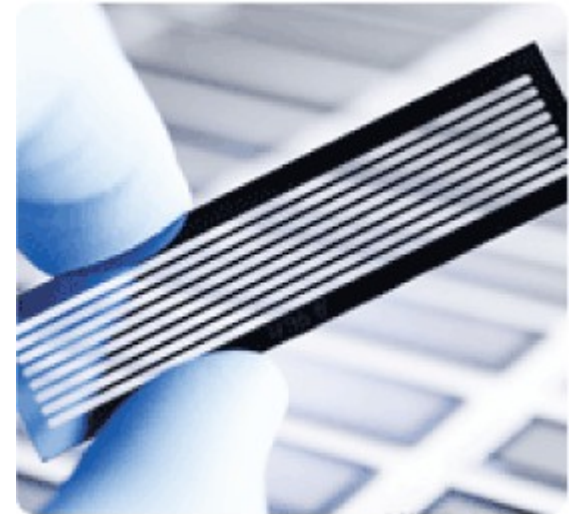
- ❖ **L'analyse d'expression différentielle** (différence d'expression) au niveau du transcriptome
- ❖ **L'étude de l'épissage alternatif** (isoformes) et recherche de **nouveaux transcrits**
  - amélioration des annotations structurales existantes
  - **L'analyse de l'épissage différentiel**
- ❖ La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

# 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 2500 : séquençage possible de 2 flowcells en parallèle.



# Le protocole RNAseq

## Préparation des Echantillons biologiques pour le RNAseq

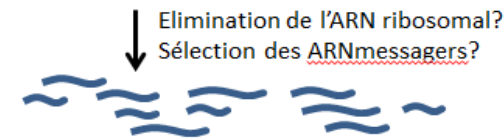
1. ARN messager ou ARN total



2. Elimination de l'ADN contaminant



3. Fragmentation de l'ARN



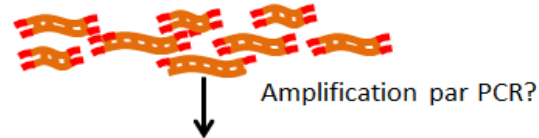
4. Retro-transcription de l'ARN en cDNA, hybride d'ADN/ARN



5. Synthèse du second brin d'ADN et ligation d'adaptateurs



6. Sélection des fragments par la taille



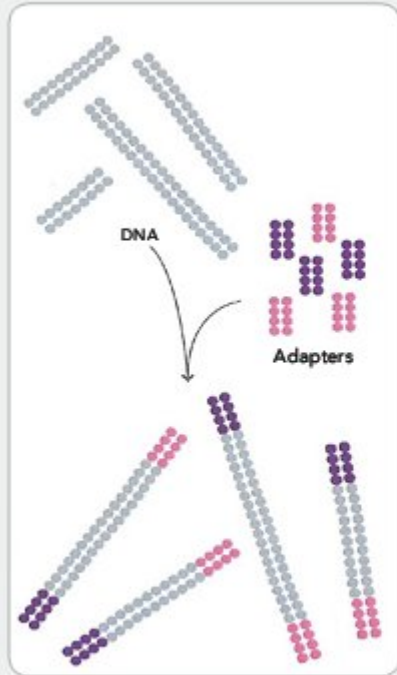
7. Séquençage des extrémités et production de « reads »





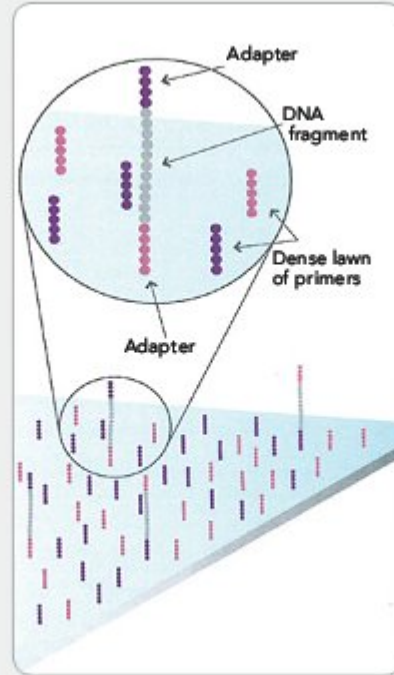
# Séquençage illumina

## 1. PREPARE GENOMIC DNA SAMPLE



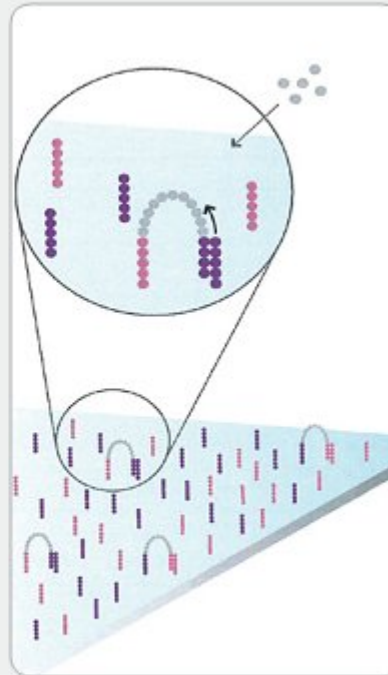
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

## 2. ATTACH DNA TO SURFACE



Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

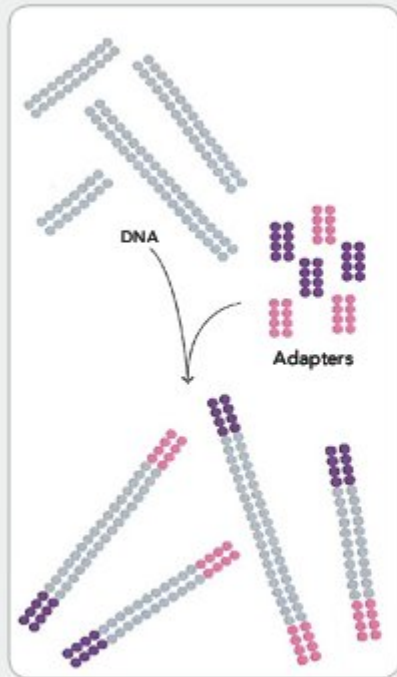
## 3. BRIDGE AMPLIFICATION



Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

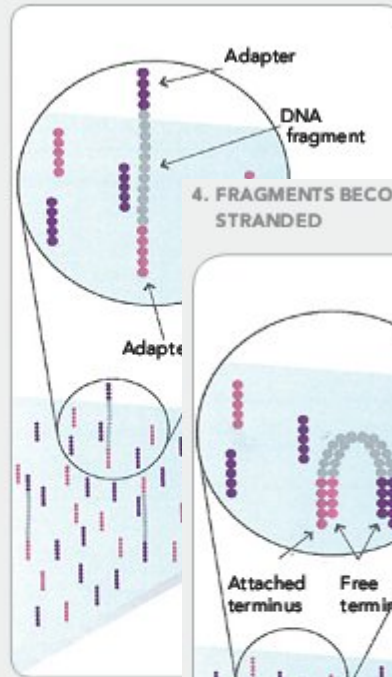
# Séquençage illumina

## 1. PREPARE GENOMIC DNA SAMPLE



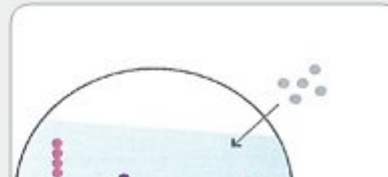
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

## 2. ATTACH DNA TO SURFACE

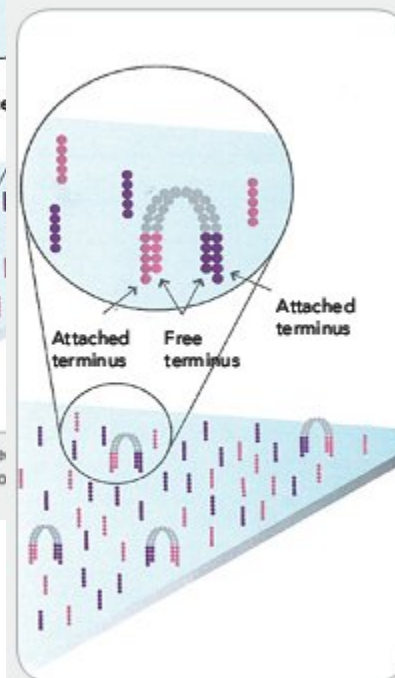


Bind single-stranded the inside surface of

## 3. BRIDGE AMPLIFICATION

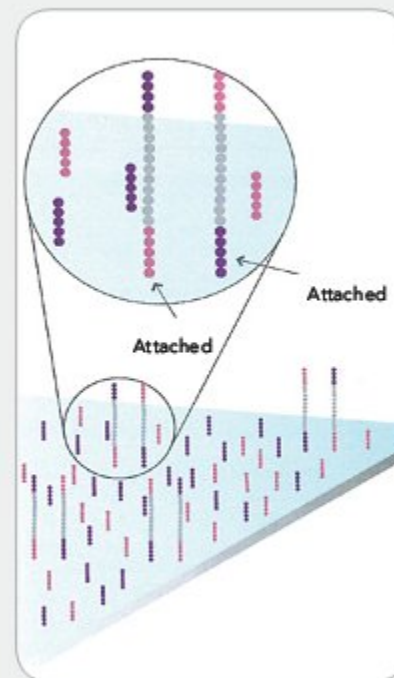


## 4. FRAGMENTS BECOME DOUBLE STRANDED



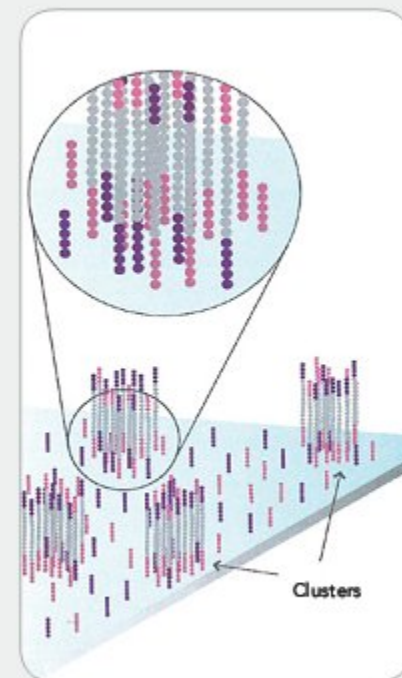
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

## 5. DENATURE THE DOUBLE-STRANDED MOLECULES



Denaturation leaves single-stranded templates anchored to the substrate.

## 6. COMPLETE AMPLIFICATION



Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.



# 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 ?

## ❖ Résultats semblables d'après :

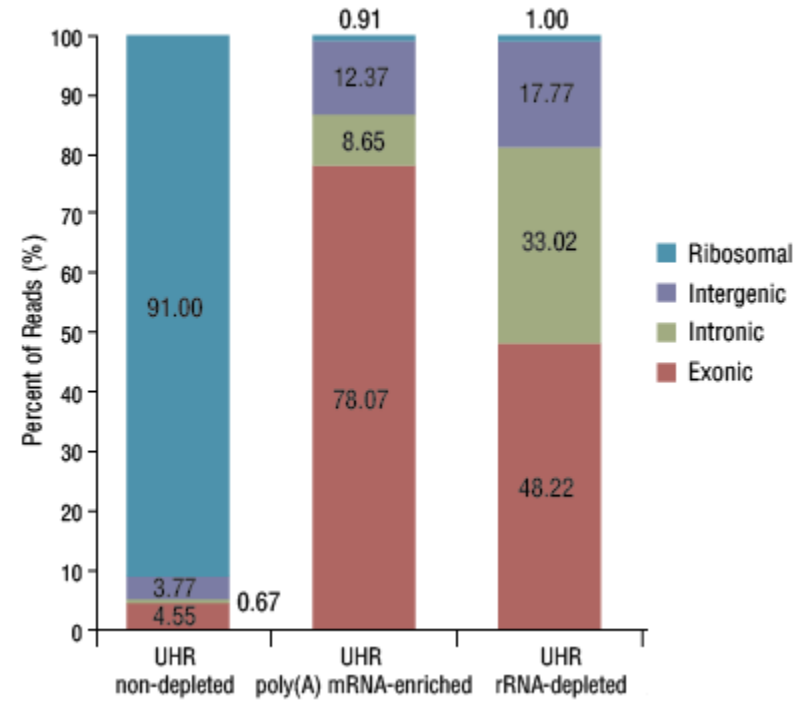
*Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling, BMC Genomics , 2014*

## ❖ Déplétion rRNA:

- For bacterial
- ARN plus varié
- Analyse des circRNA, d'ARN non-codant possible
- 

## ❖ Enrichissement polyA :

- Plus de read ds les exons
- Peu de matériel bio
- Pas de transcrits sans queue PolyA ou partiellement dégradés

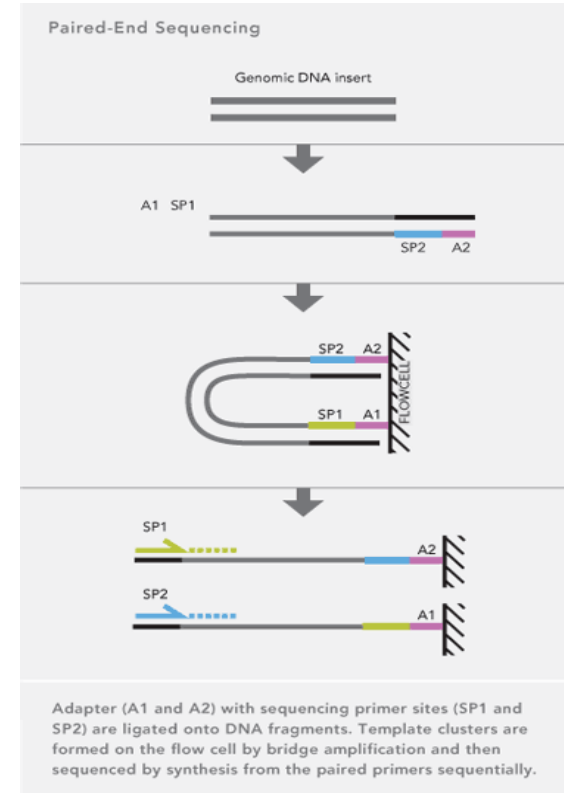


# Paired-end



Protocole différent (Adaptateurs spécifiques)

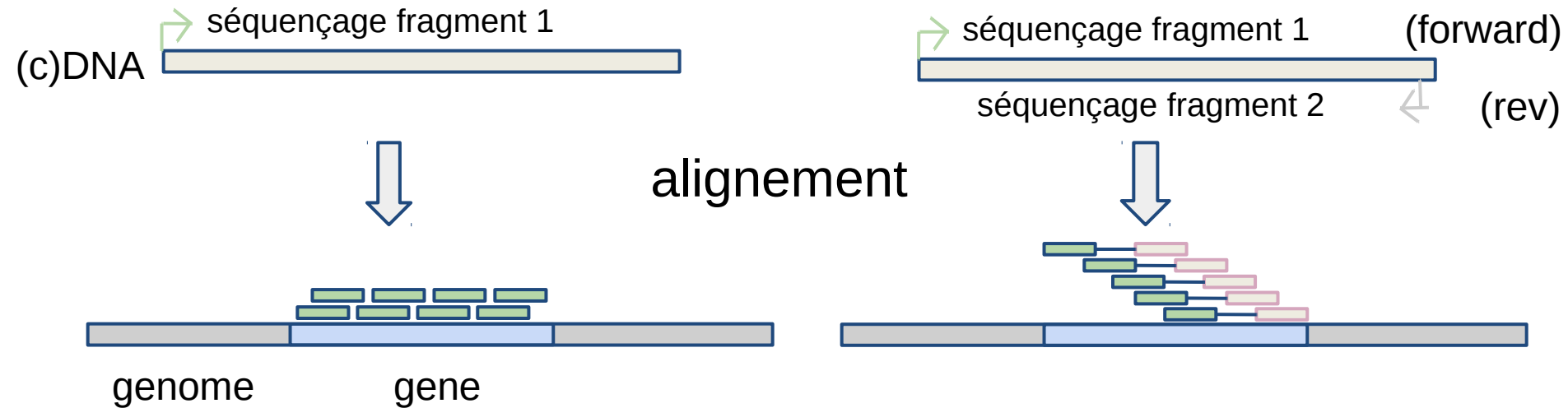
- ❖ Améliore le mapping
- ❖ Aide à la détection de variant alternatif
- ❖ Plus généralement aide à la détection de : variation structurale de génome (insertion/délétion), CNV, réarrangement génomique



# Single-end vs Paired-end

## Single-end

## Paired-end



- ❖ La taille des cDNA détermine la taille d'insert (p. ex. 200-500 pb).
- ❖ Les fragments sont habituellement en Forward-Reverse.
- ❖ Le type de librairie est demandé par les aligneurs

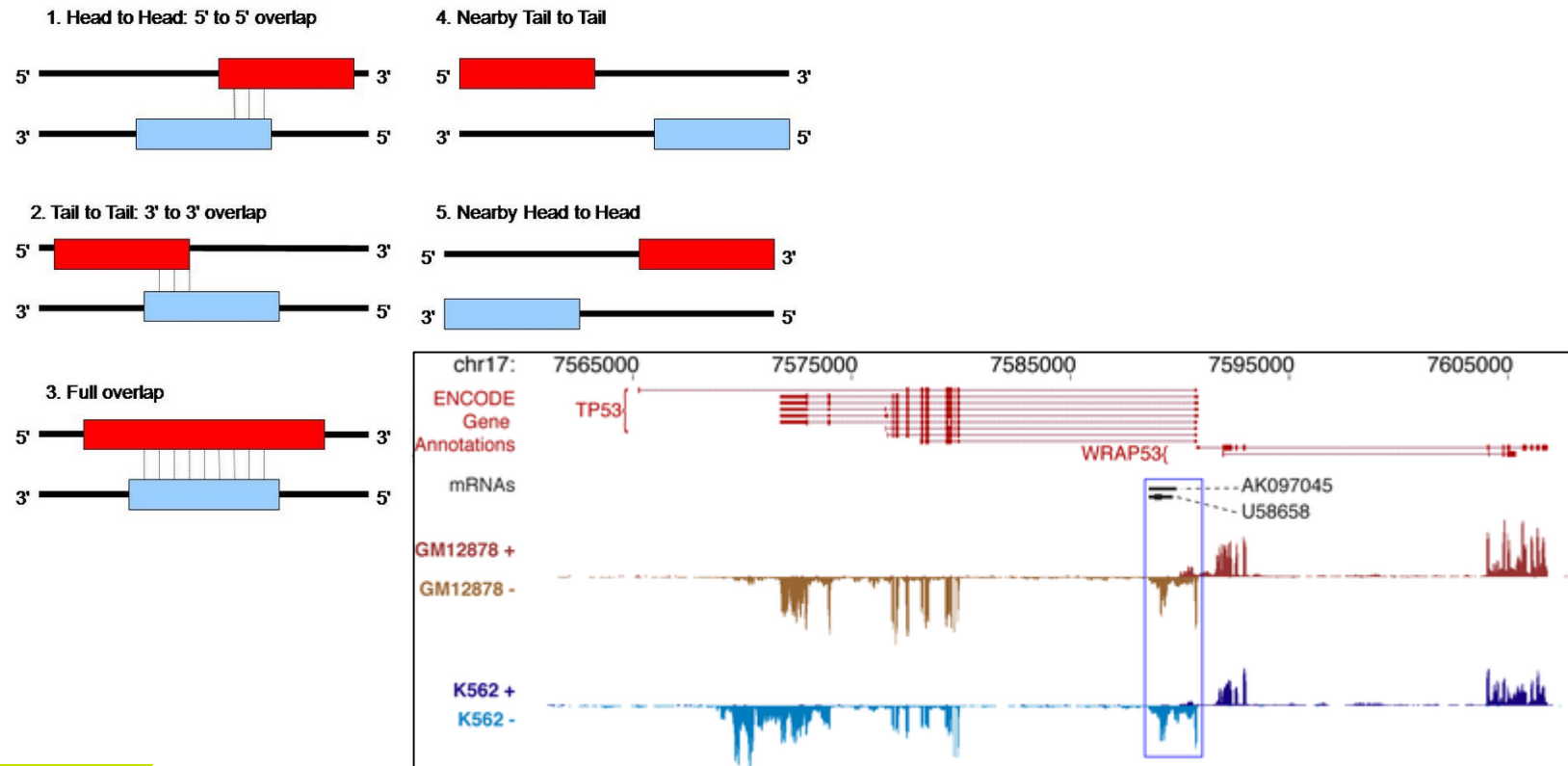
# L'intérêt des librairies brin spécifique

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



# Equilibre profondeur / répétitions ?

- ❖ directives du consortium ENCODE (Juin 2017)
  - **plus de deux répétitions biologique**
  - **Sequencing depth : “Each RNA-Seq library must have a minimum of 30 million aligned reads/mate-pairs.”**

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

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



# Equilibre 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*)

# Equilibre profondeur / répétitions ?

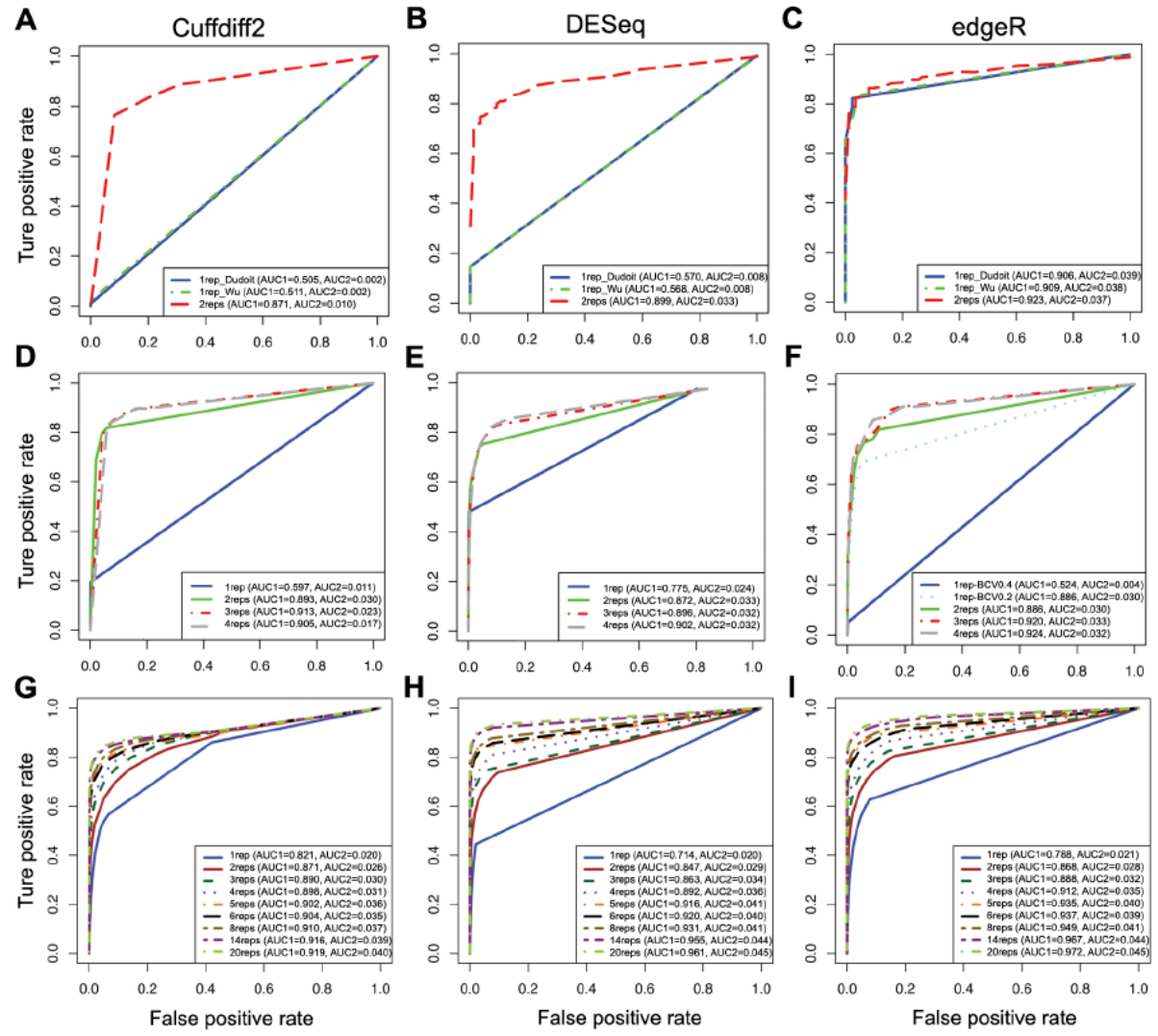
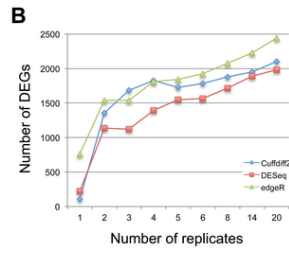
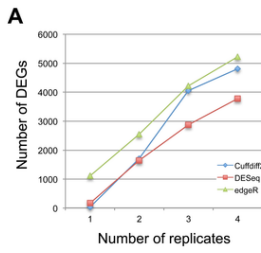
L'effet du nombre de réplicats sur le taux de vrai positifs et de faux positifs



K\_N

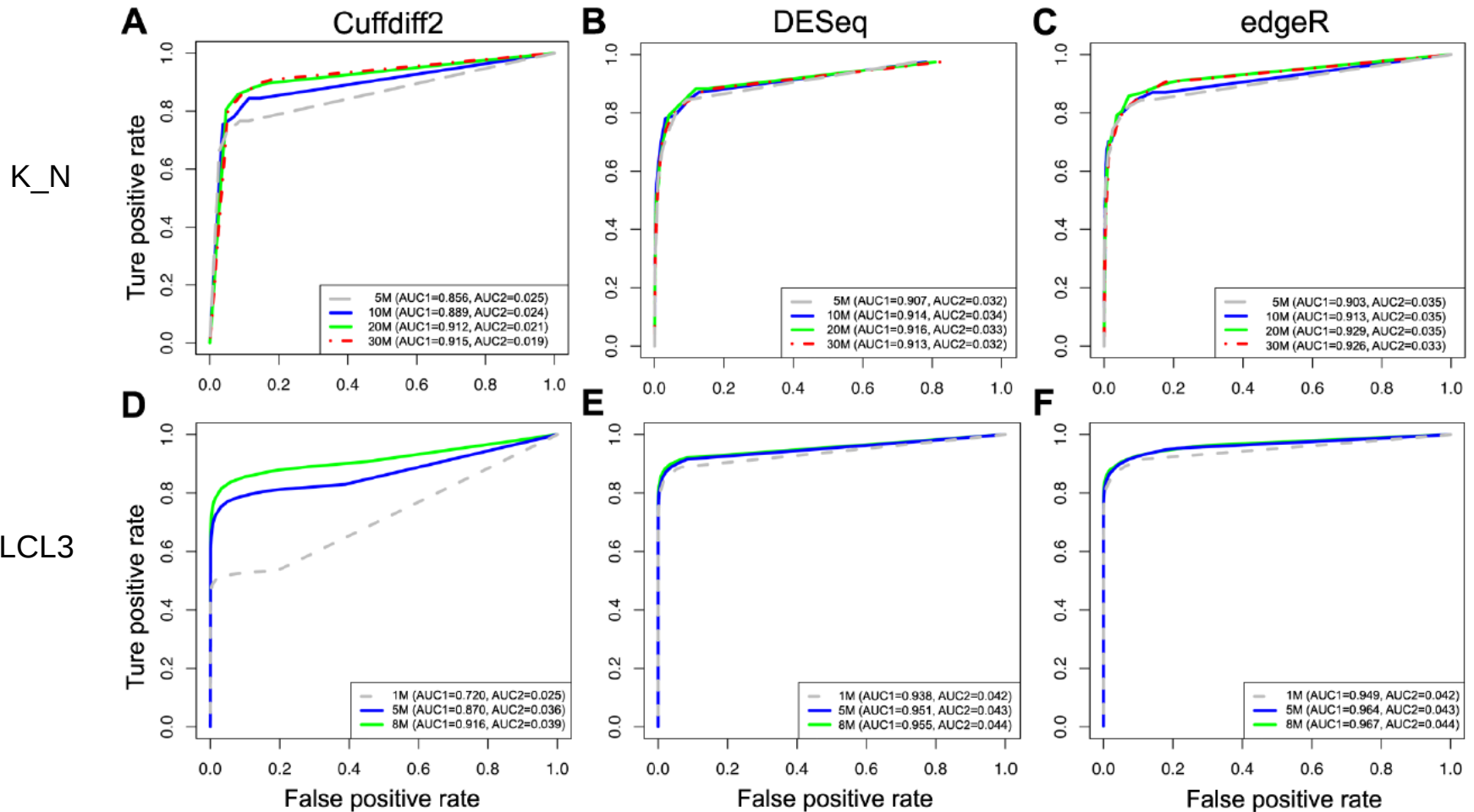
K\_N

LCL2



# Equilibre profondeur / répétitions ?

L'effet de la profondeur.



# Equilibre 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

# 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 référence

- Dépendant de la qualité de l'annotation structurale
- Peu coûteux

## ❖ Génome de référence

- Permet une approche combinée :
  - sur **transcriptome**
  - recherche de **nouveaux transcrits**
- Ressources ++
- Alignement épissé

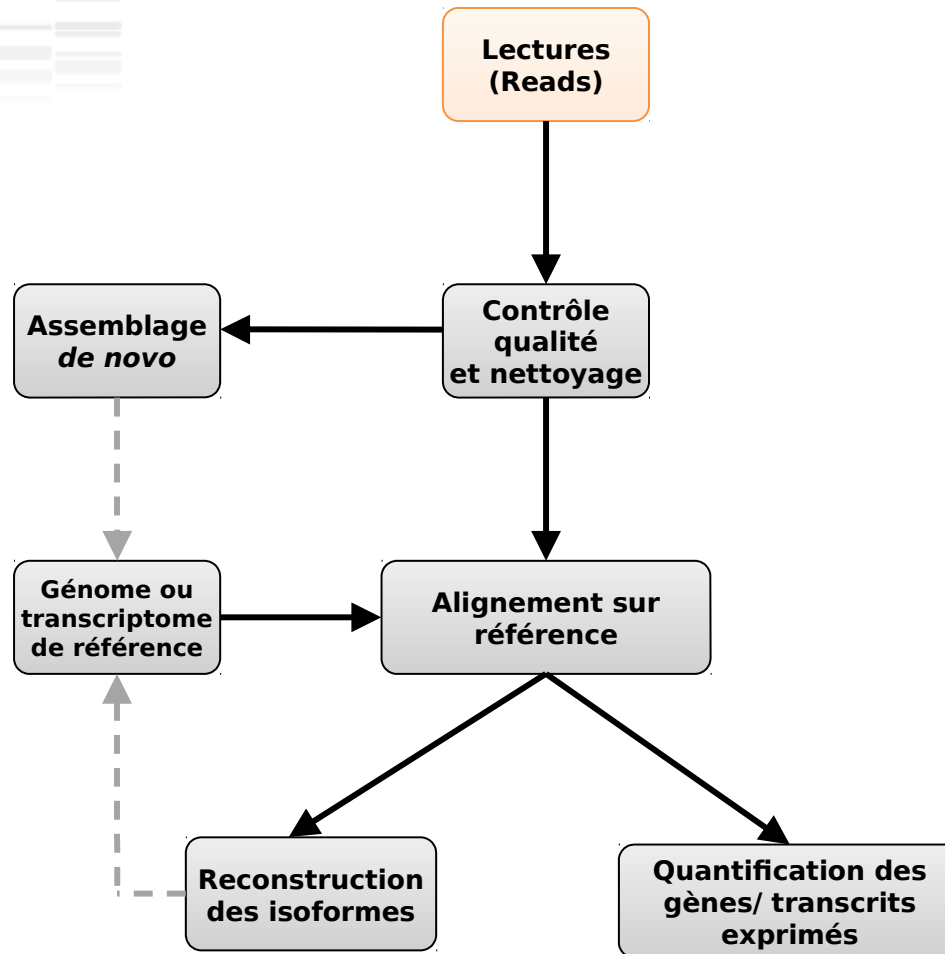


# Pipeline d'analyse RNA-Seq : avec référence

- ❖ **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
- ❖ **Reconstruction de nouveaux isoformes**
- ❖ **Comptage** des gènes / transcrits

# **\_03** **Obtenir des séquences de qualité**

# Workflow d'analyse RNA-Seq







# Plan : Données brutes et qualité

- ❖ **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
GATTTGGGGTTCAAAGCAGTATCGATCAAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((( (**+))%%%+)(%%%) .1***-+*'')**55CCF>>>>>CCCCCCC65
```

- 1 ère ligne = identifiant de la séquence

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

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



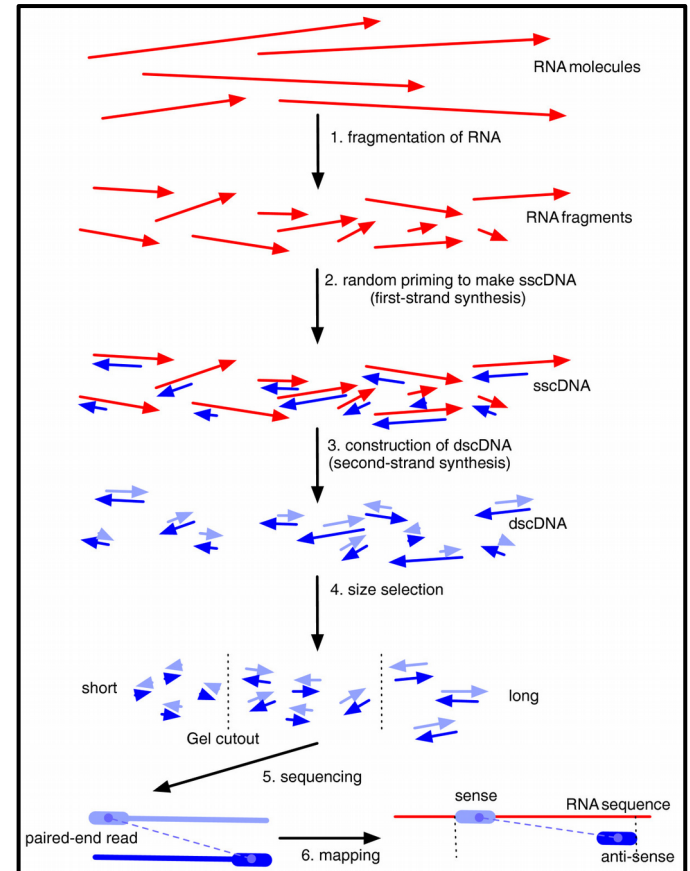


# Biais spécifiques au RNA-Seq

- ❖ Influence du mode de préparation de la banque
  - amplification hexamérique 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

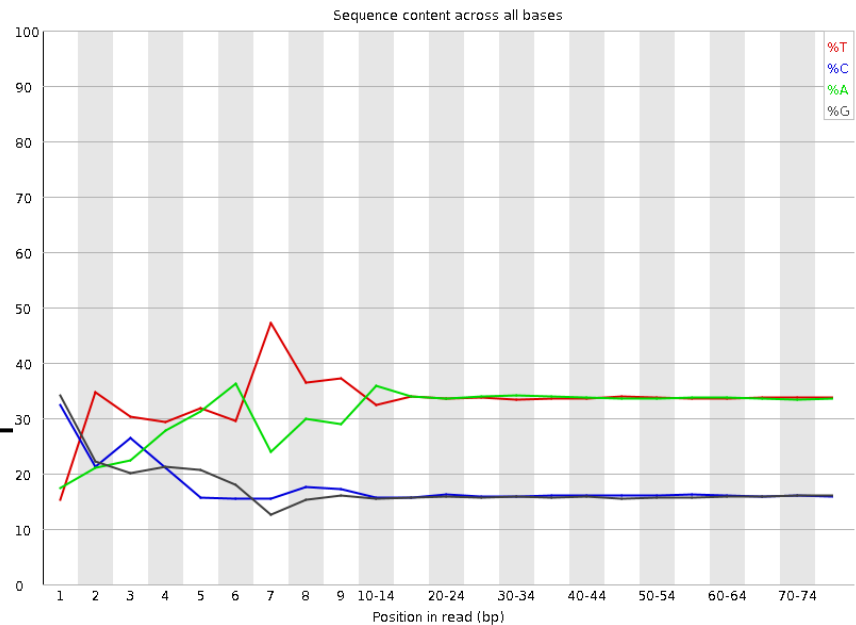
Nucleic Acids Research, 2010, Vol. 38, No. 12 e131  
doi:10.1093/nar/gkq224

## Biases in Illumina transcriptome sequencing caused by random hexamer priming

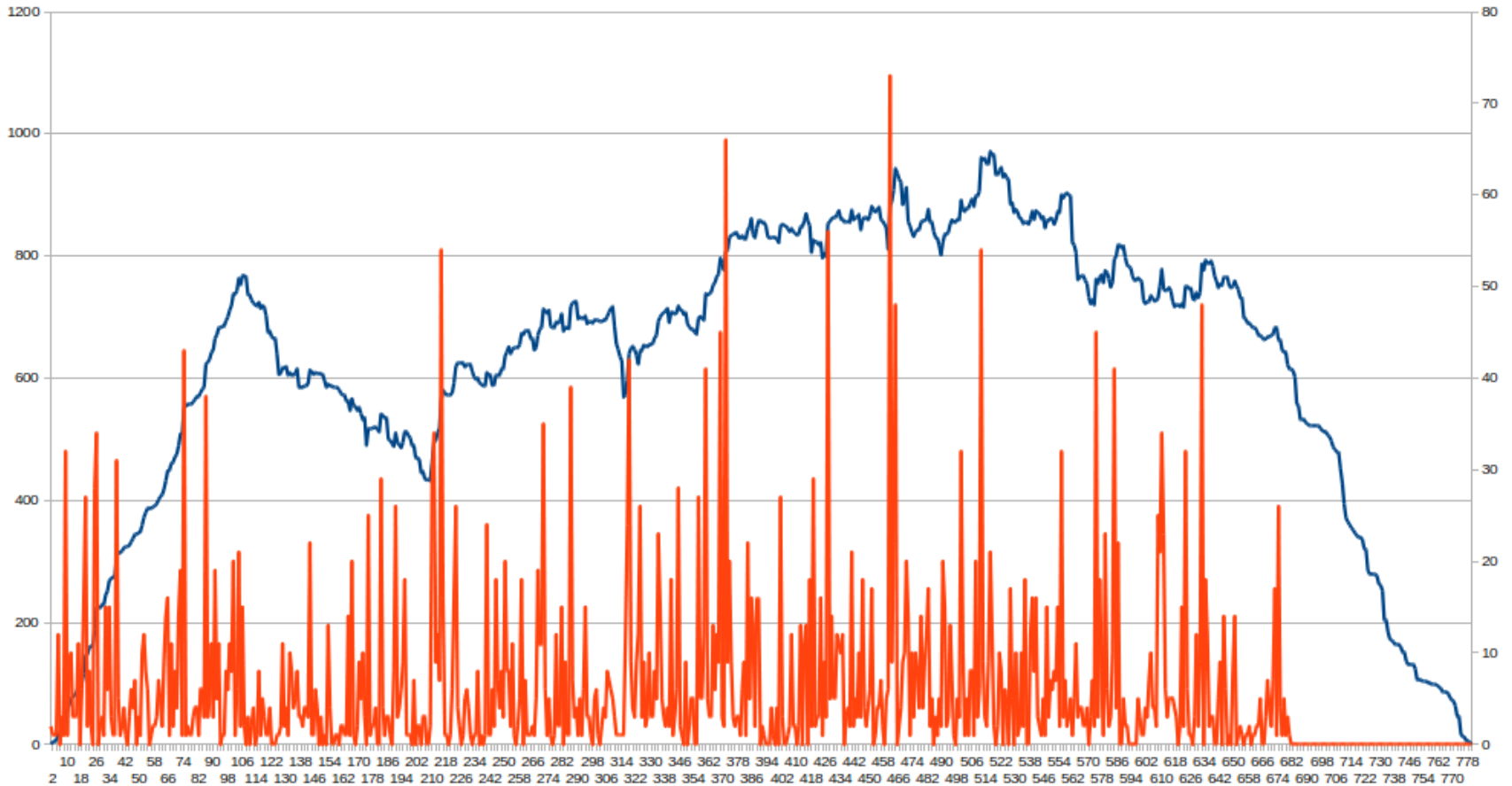
Kasper D. Hansen<sup>1,\*</sup>, Steven E. Brenner<sup>2</sup> and Sandrine Dudoit<sup>1,3</sup>

### 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*

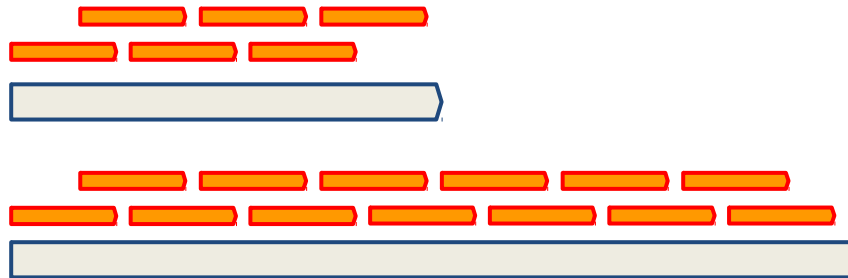


Orange = reads start sites  
Blue = coverage



# Biais : longueur des transcrits

- La capacité, en utilisant des **comptages** obtenus par **RNA-Seq**, à 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

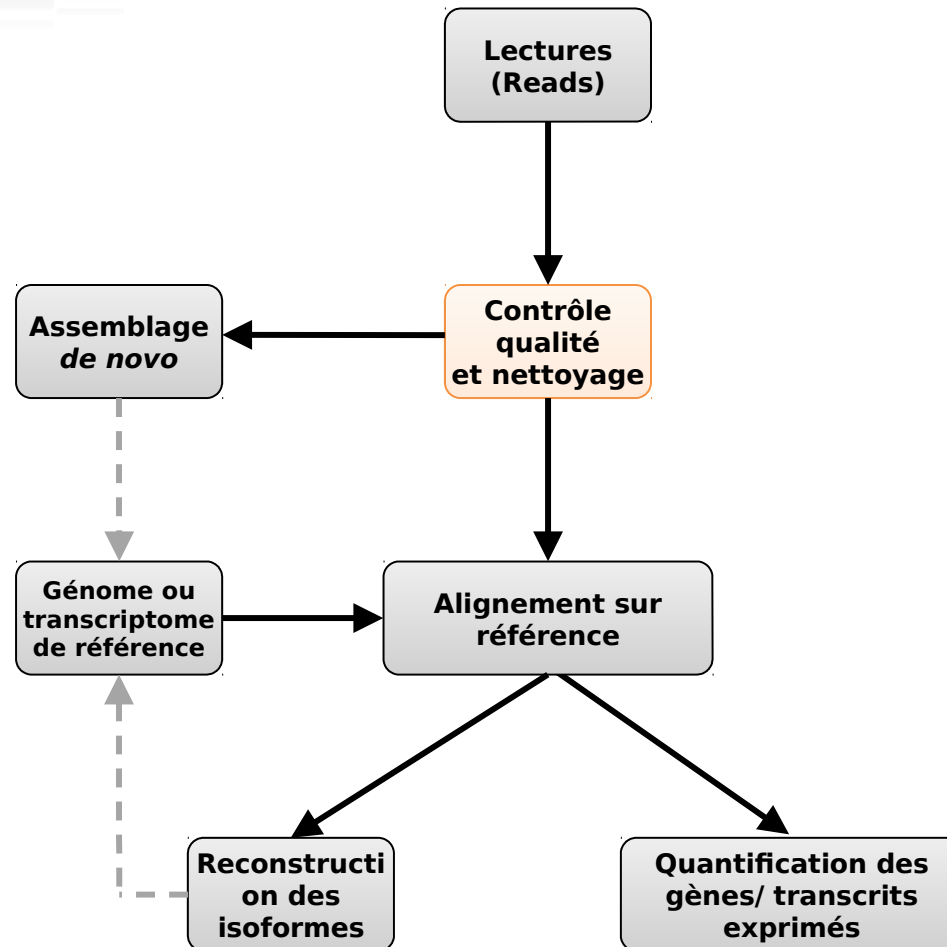




## Biais : « mappabilité »

- 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



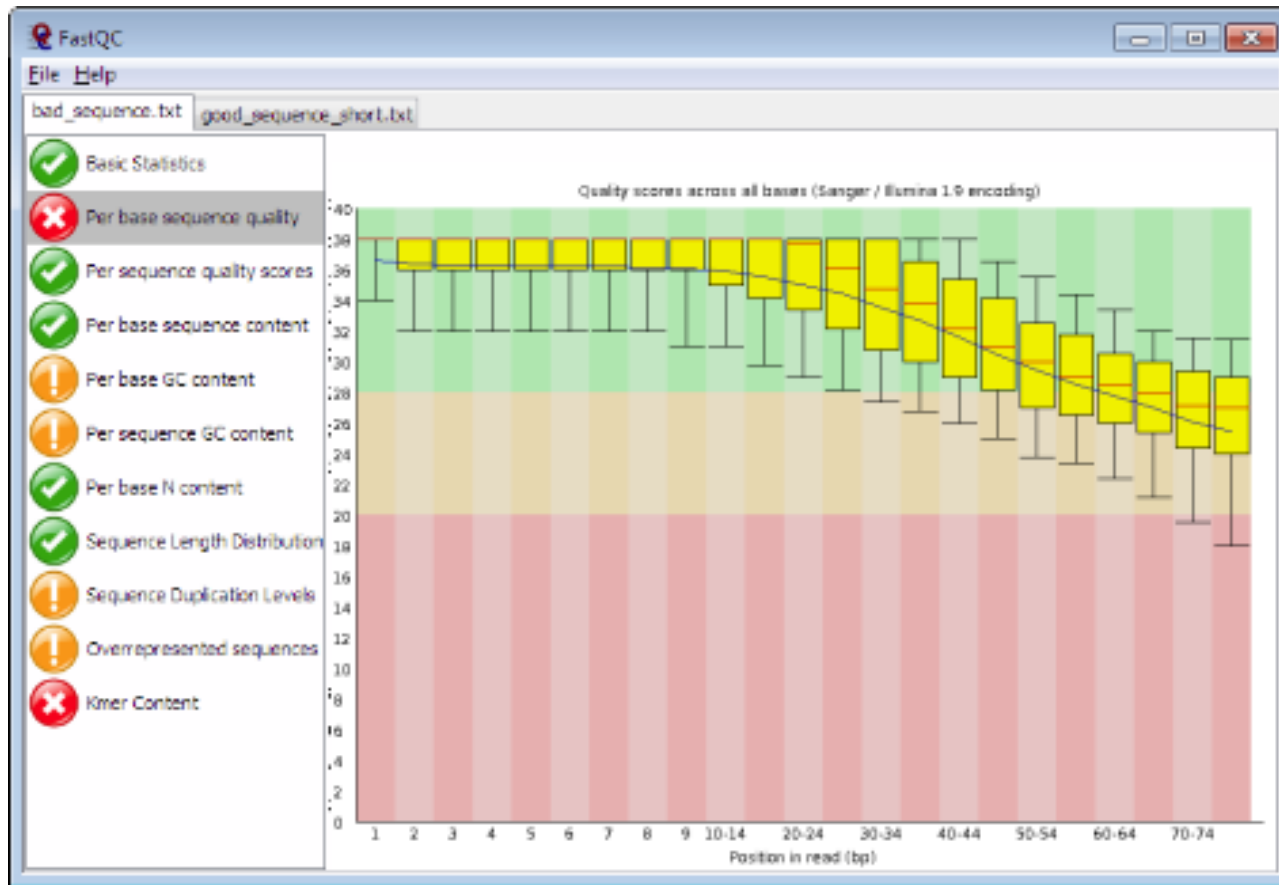
# Contrôle qualité

## 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 choix des paramètres pour le nettoyage des données

# Contrôle qualité avec FastQC

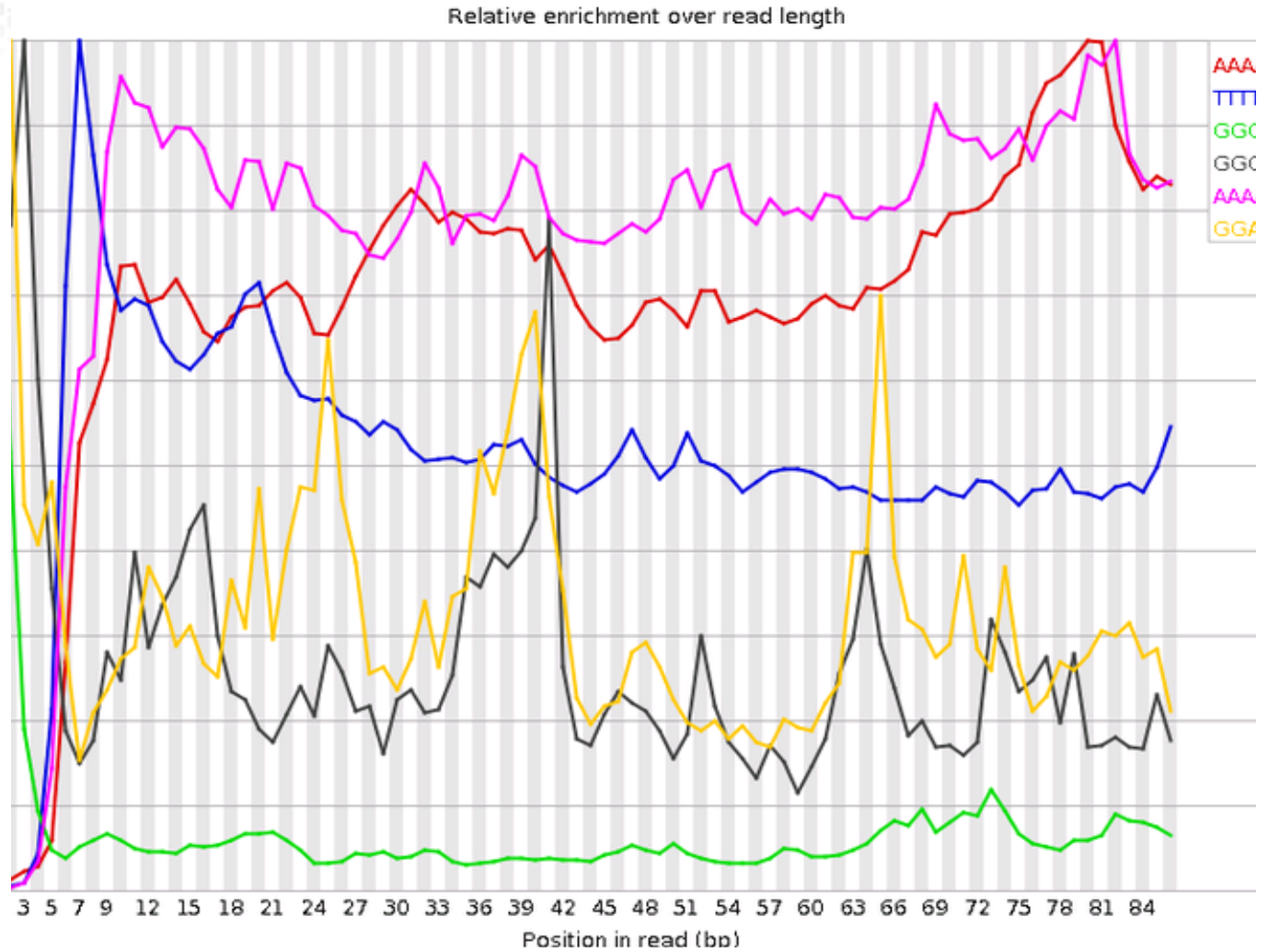
❖ orienté DNA-Seq



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

# Contrôle qualité

Kmer  
content





# 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



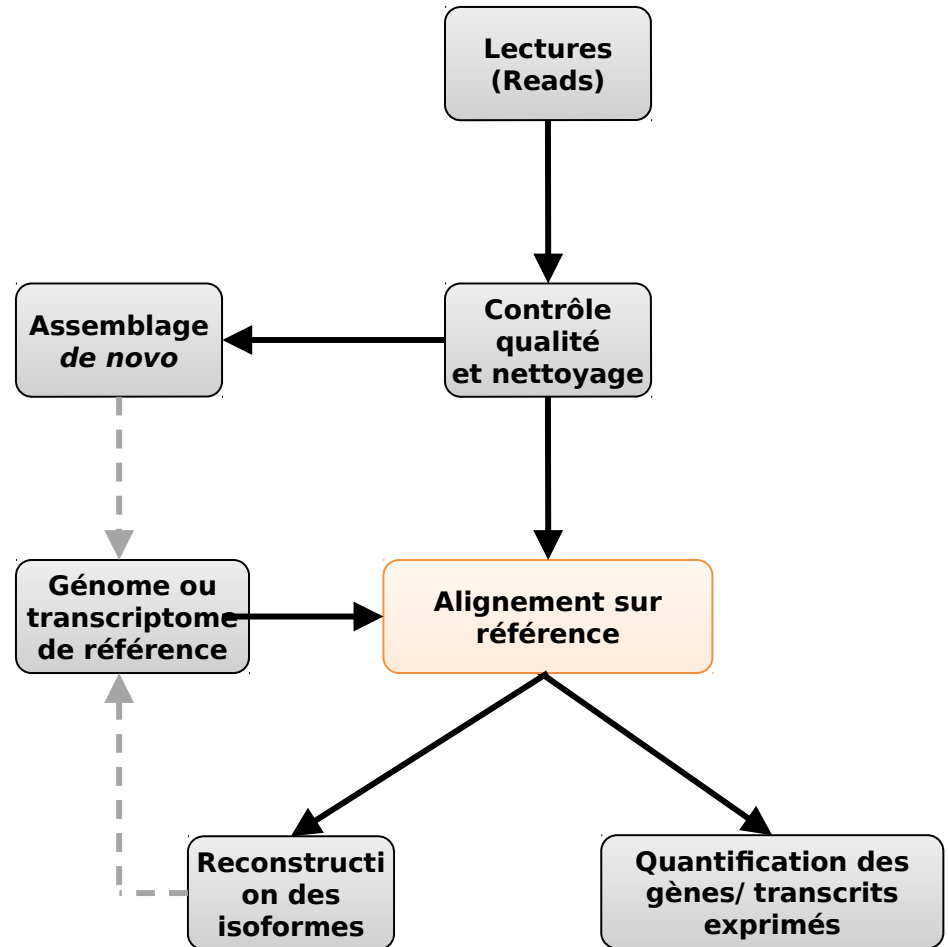
# Travaux pratiques

## 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

# 04

## MAPPING et Visualisation



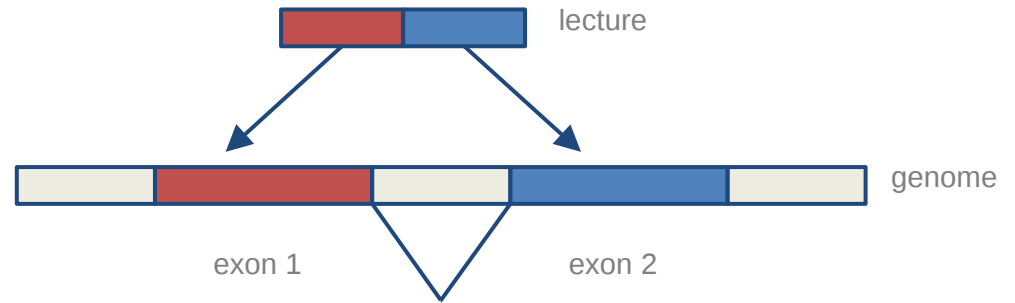
# Alignement épissé

## Objectifs :

- ❖ **Aligner** les **lectures** issues du séquençage de **dscDNA** (transcrits) sur le **génom**e, 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



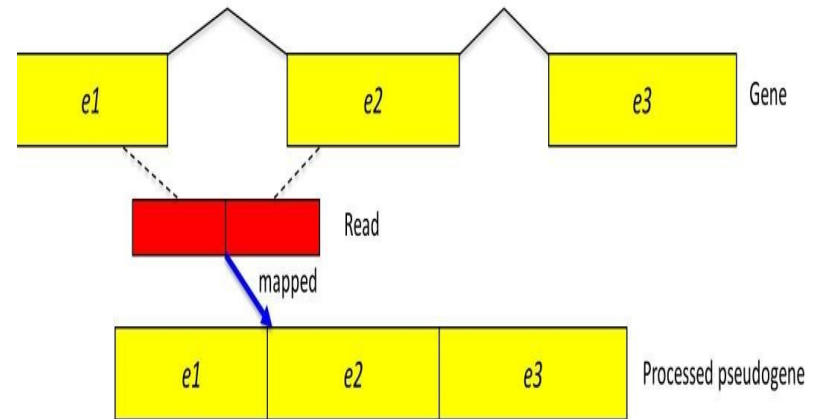
**Le *mapping* est la *prédiction* du *locus* dont est originaire la lecture.**

- **Prédiction** : chaque outil propose un/plusieurs locus.
- **Locus** : le résultat est un ensemble de positions génomiques (ex.: chr1:100..150)
- Mapping ARN  $\neq$  Mapping ADN
- Mapping  $\neq$  Alignement

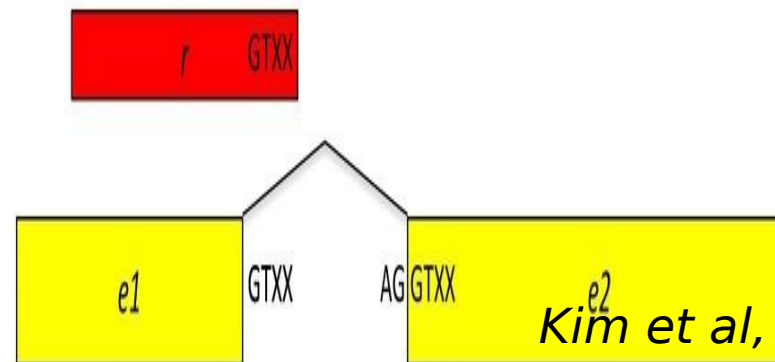
Les outils de mapping font de mauvais alignements (sauf aux jonctions).

# Cas difficiles

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



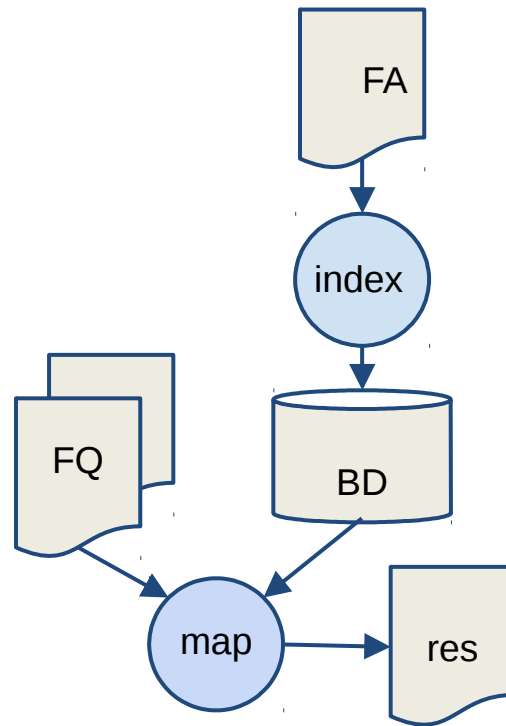
- Fin de la lecture sur un exon propre
- Lecture sur une jonction non-connue d'un gène peu exprimé



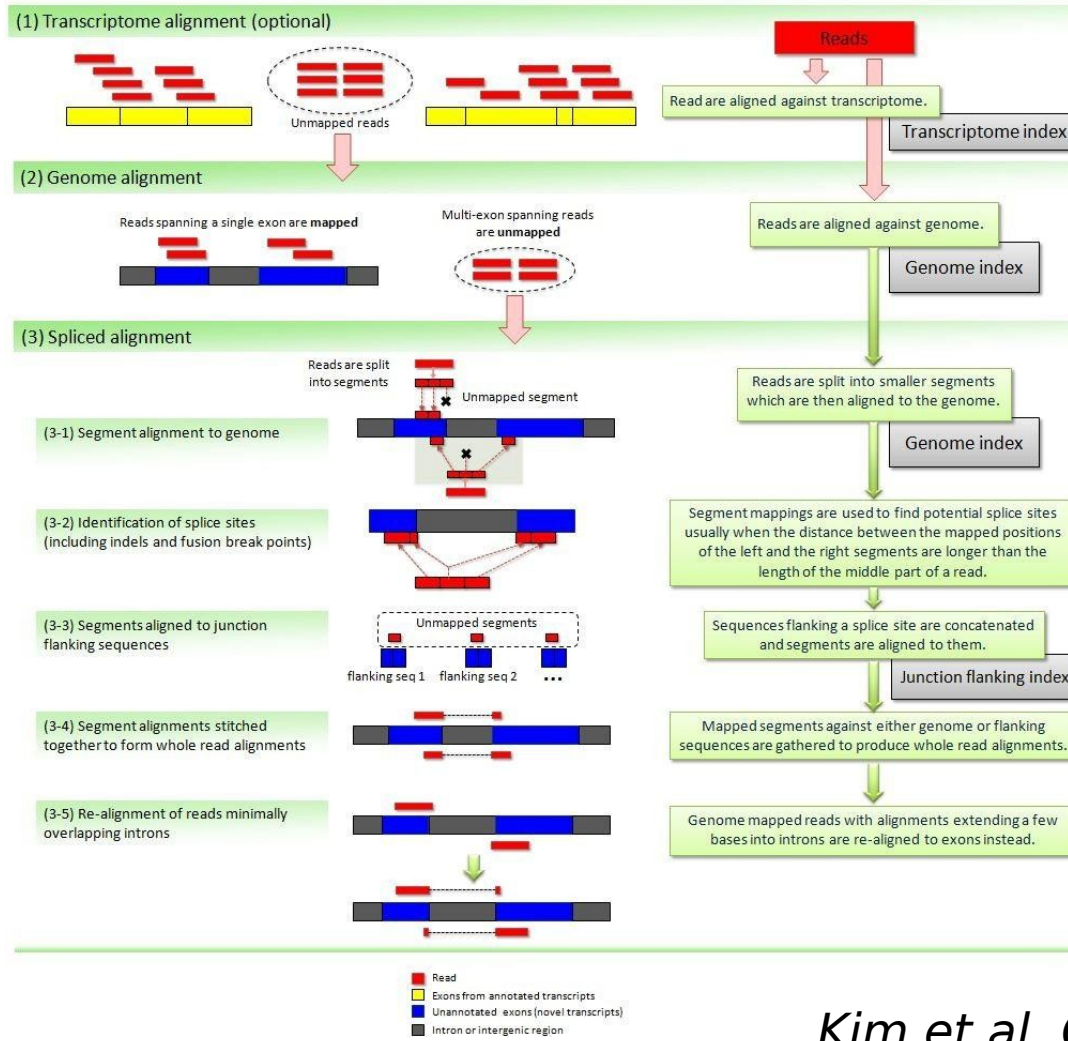
*Kim et al, Genome Biology, 2013*

# Étapes de mapping

- ❖ *Indexation du génome une fois pour toutes*
- ❖ *Mapping des lectures en utilisant l'index*



# Tophat2



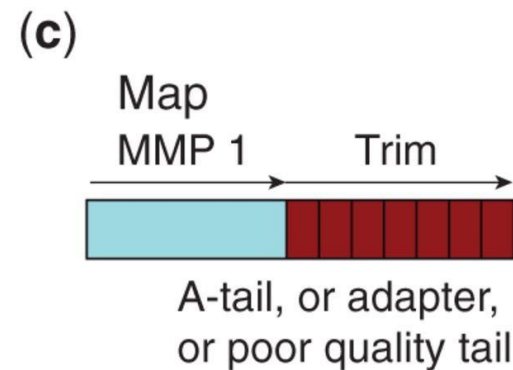
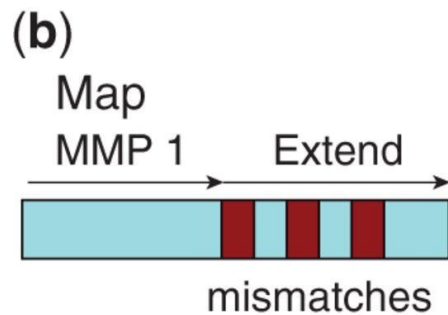
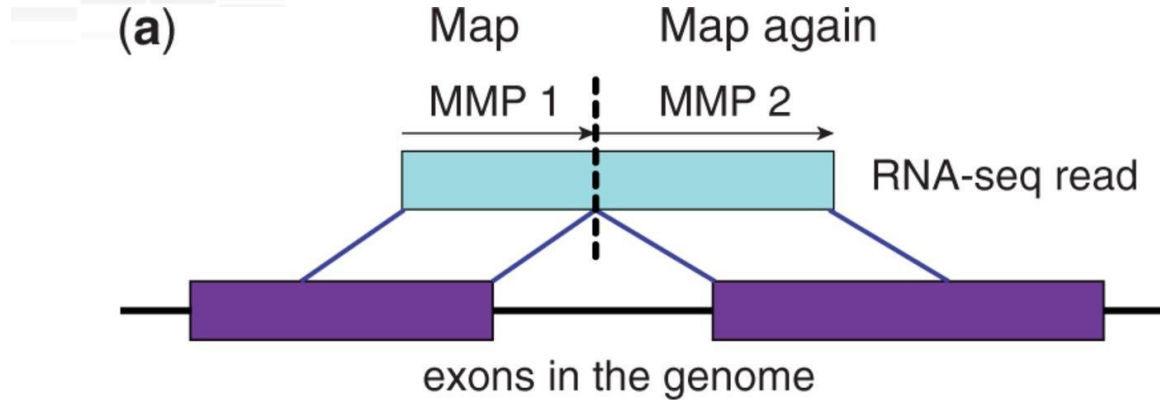
Tophat2 est constitué de beaucoup d'étape pour résoudre chaque cas difficile.

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

*Kim et al, Genome Biology, 2013*



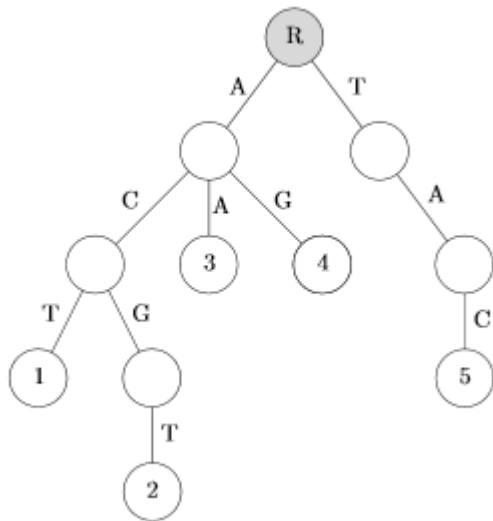
# STAR is an ultrafast universal RNA-seq aligner



*Dobin et al, Bioinformatics, 2011*

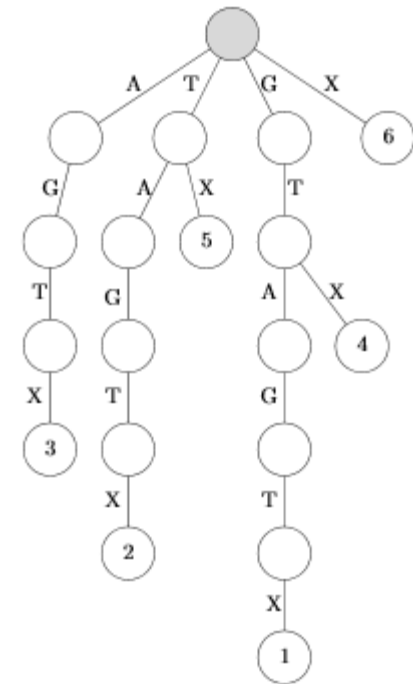
# STAR is an ultrafast universal RNA-seq aligner

Aligneurs index BWT  
(BWA, Bowtie, SOAP)



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

STAR



*GTAGT.*



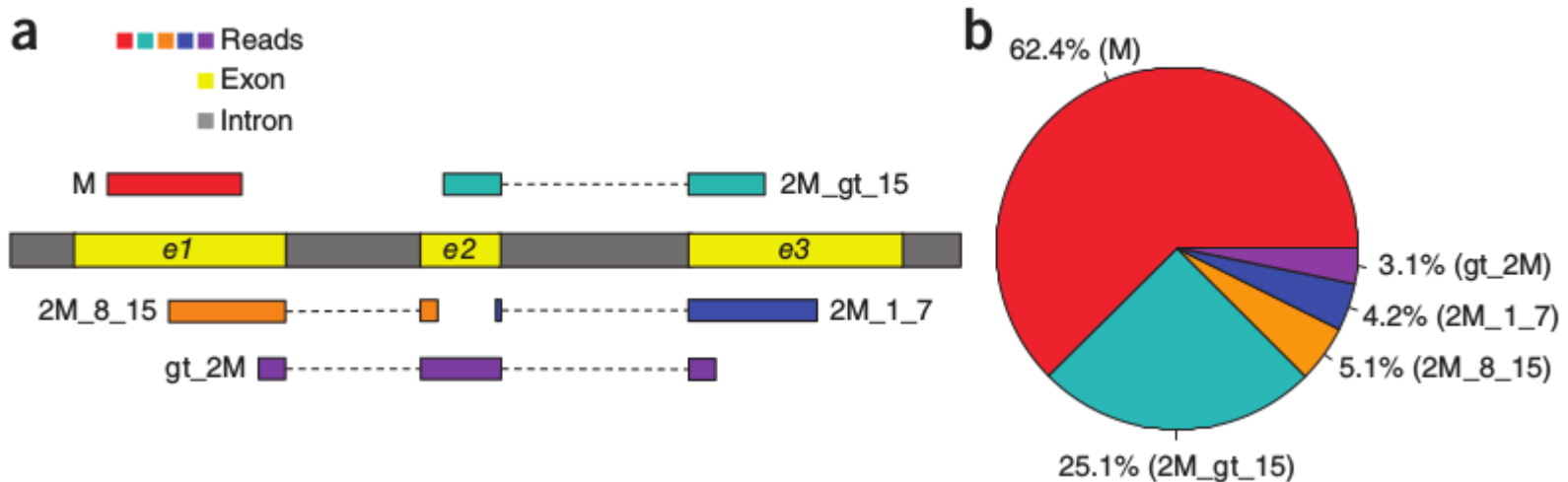
# STAR is an ultrafast universal RNA-seq aligner

- Préconisé par Djebali et al, Methods in Molecular Biology 2017
- Ds galaxy Sigenae: utiliser l'option gtf :
  - Indexation avec gtf: transcriptome de ref
  - STAR --quantMode TranscriptomeSAM

With `--quantMode TranscriptomeSAM` option STAR will output alignments translated into transcript coordinates in the `Aligned.toTranscriptome.out.bam` file (in addition to alignments in genomic coordinates in `Aligned.*.sam/bam` files). These transcriptomic alignments can be used with various transcript quantification software that require reads to be mapped to transcriptome, such as RSEM or eXpress. For example, RSEM command line would look as follows:

# HiSAT2

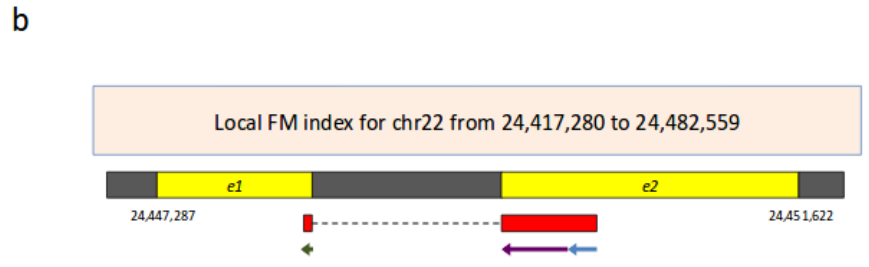
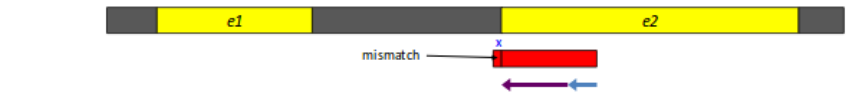
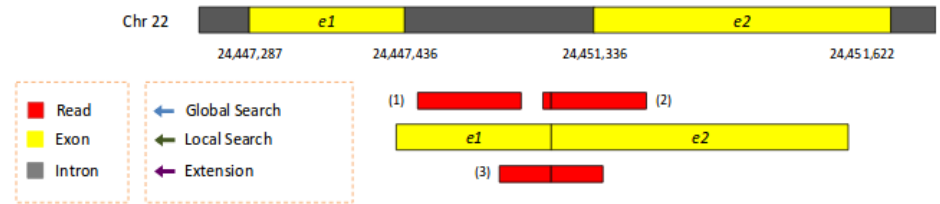
- ❖ “We recommend that the HISAT and TopHat2 users switch to HISAT2.”
- ❖ 2 FM index : tout genome + regions de 64kb



*Kim et al, Nature, 2015*

# HiSAT2

- ❖ 3 types reads
- ❖ a) en plein
- ❖ b) épissé avec petite region
- ❖ c) épissé 2 grandes régions

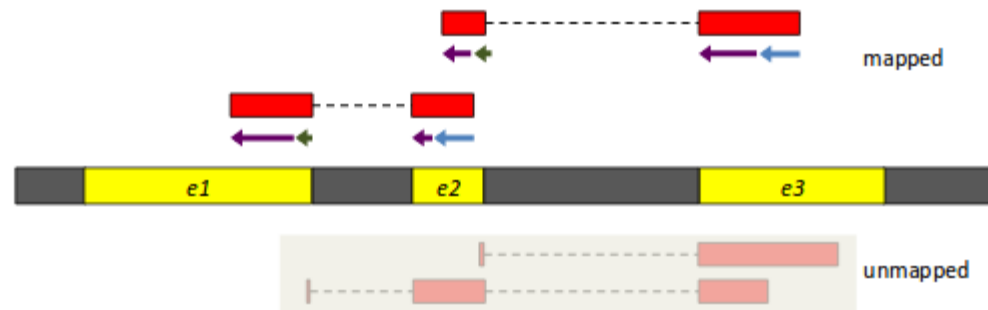


*Kim et al, Nature, 2015*

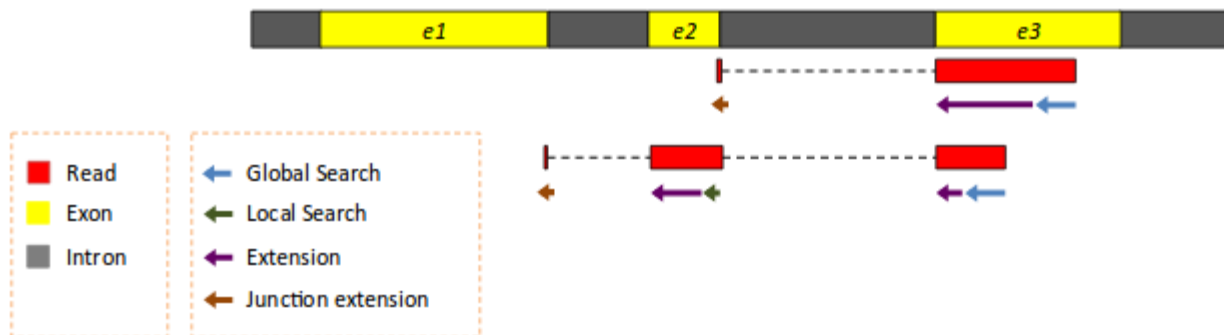
# HiSAT2

- ❖ Two-step approach version of HISAT to allow alignment of junction reads with small anchors

1<sup>st</sup> run of HISAT to discover splice sites



2<sup>nd</sup> run of HISAT to align reads by making use of the list of splice sites collected above



# Quel logiciel utiliser ?

## 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.

# Benchmark of RNAseq aligners

Recall: measures the fraction of all bases that were aligned correctly,  
 Precision: measures the fraction of all aligned bases that were aligned correctly

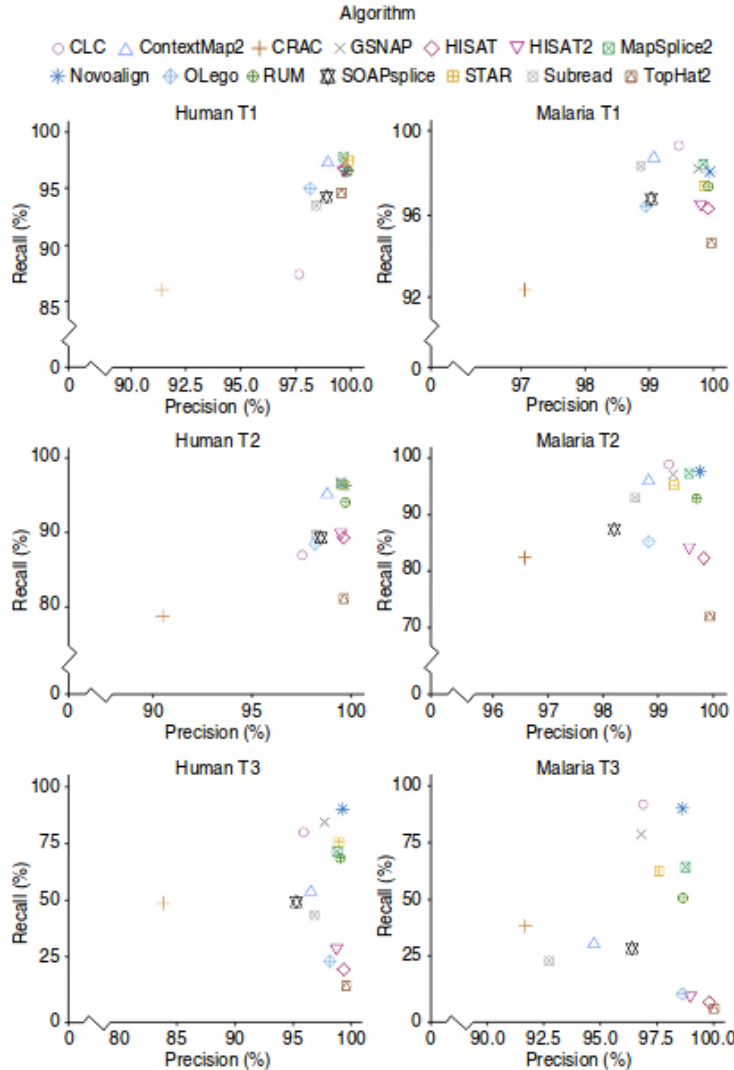


Figure 1 | Base-level precision and recall for human and malaria data sets.

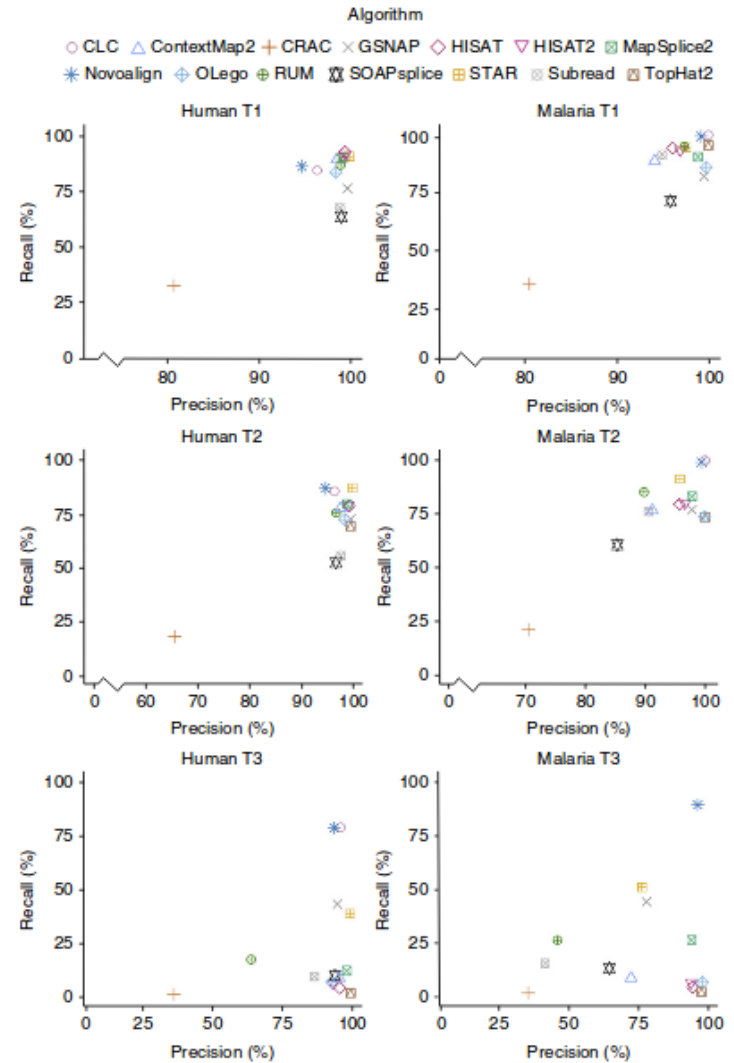
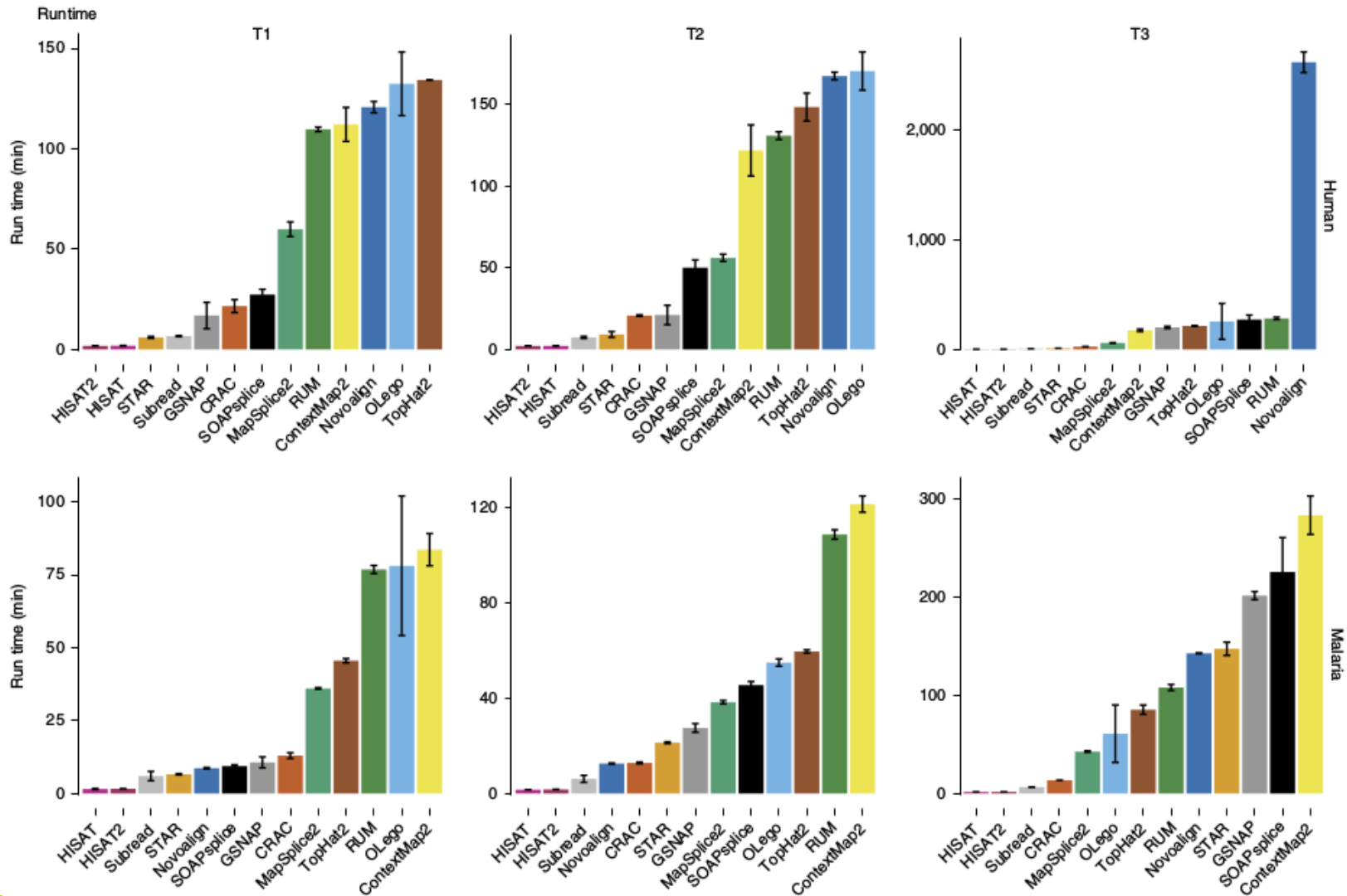


Figure 2 | Junction-level precision and recall for human and malaria data sets.



# Benchmark of RNAseq aligners





# Alignement : données initiales

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

# Format GTF : Gene Transfert Format

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

*Format :*

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

*Exemple :*

```
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 **1<sup>ère</sup> colonne** du **GTF**) doivent être les **mêmes**

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

# 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 ?

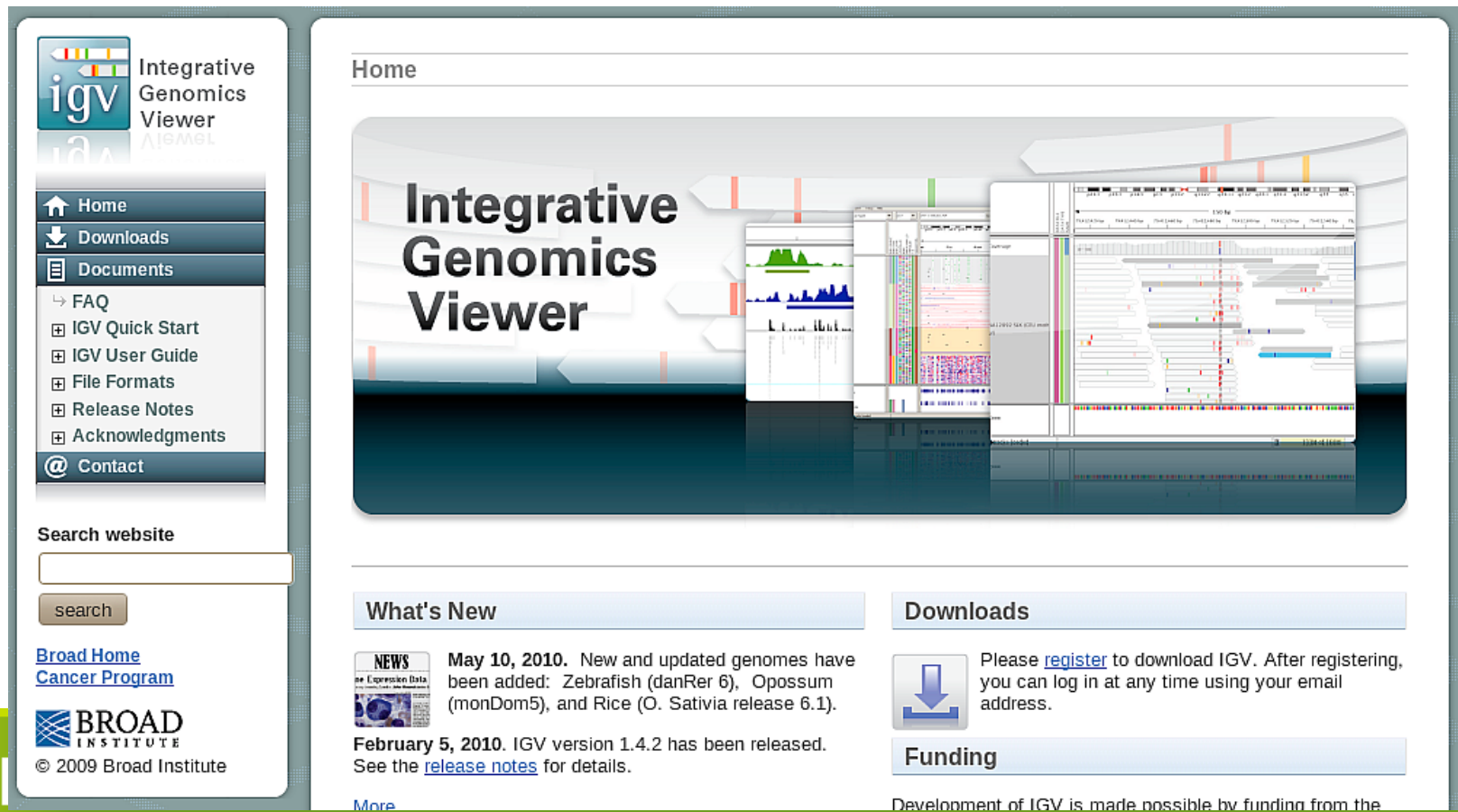
[http://genoweb.toulouse.inra.fr/~formation/2\\_Galaxy\\_SGS-SNP/.formats/sam.html](http://genoweb.toulouse.inra.fr/~formation/2_Galaxy_SGS-SNP/.formats/sam.html)



# TP: lancer l'alignement

# Visualisation des alignements avec IGV

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



The screenshot shows the homepage of the Integrative Genomics Viewer (IGV). The page features a navigation menu on the left, a search bar, and a main content area with a large banner for the IGV application. Below the banner, there are sections for 'What's New' and 'Downloads'. The 'What's New' section includes news items from May 10, 2010, and February 5, 2010. The 'Downloads' section provides instructions on how to register and download the software. The footer includes the Broad Institute logo and the text '© 2009 Broad Institute'.

**Integrative Genomics Viewer**

**Home**

- Home
- Downloads
- Documents
- FAQ
- IGV Quick Start
- IGV User Guide
- File Formats
- Release Notes
- Acknowledgments
- Contact

Search website

search

[Broad Home Cancer Program](#)

**BROAD INSTITUTE**

© 2009 Broad Institute

**What's New**

**NEWS** May 10, 2010. New and updated genomes have been added: Zebrafish (danRer 6), Opossum (monDom5), and Rice (O. Sativa release 6.1).

February 5, 2010. IGV version 1.4.2 has been released. See the [release notes](#) for details.

[More](#)

**Downloads**

Please [register](#) to download IGV. After registering, you can log in at any time using your email address.

**Funding**

Development of IGV is made possible by funding from the

# 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

---

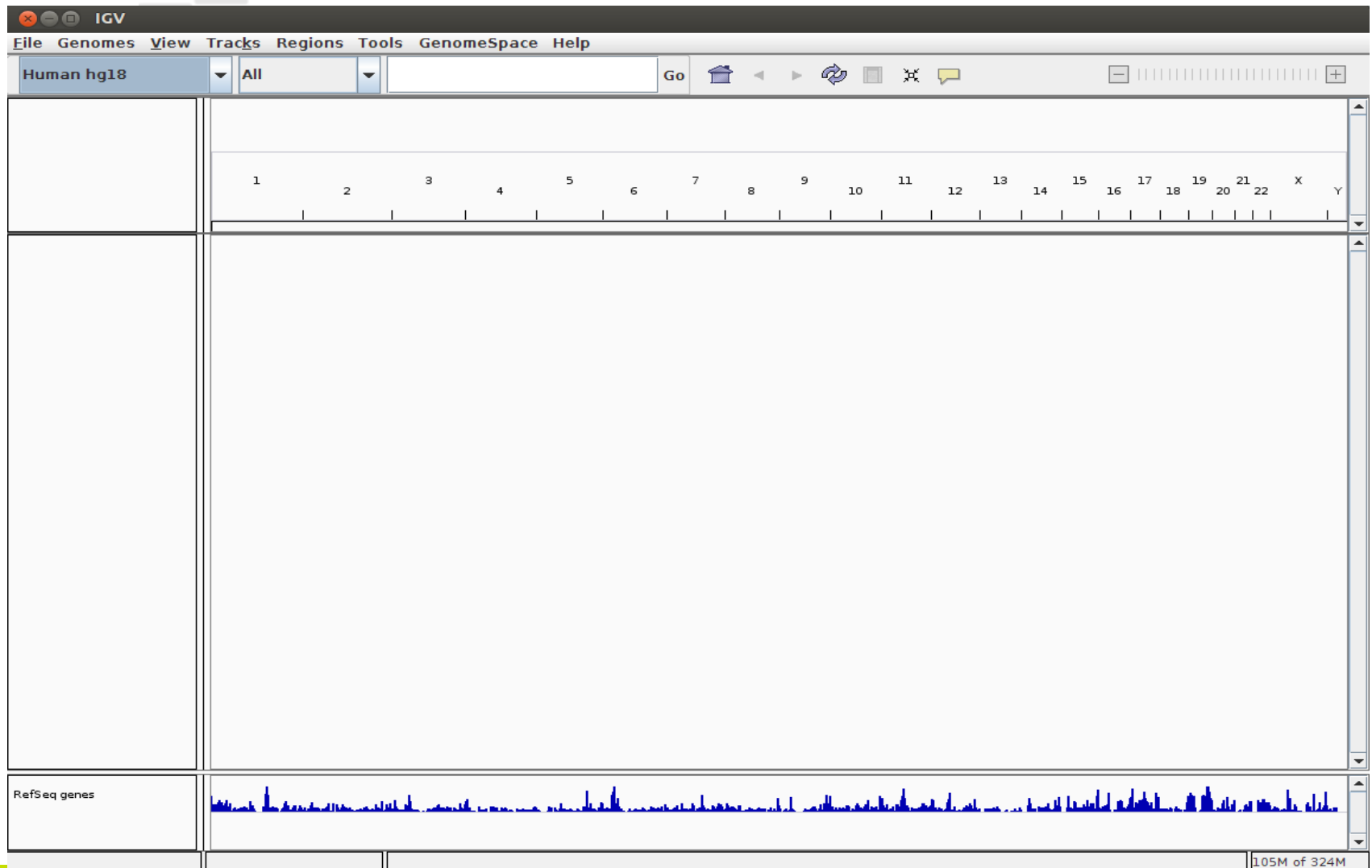
## File Formats

---

- [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

The screenshot shows the IGV application window with the 'File' menu open. The 'Load Genome from File...' option is highlighted. A 'Load Genome' dialog box is displayed, showing a file browser with a list of files and folders. The files listed are 'initrd.img', 'initrd.img.old', 'vmlinuz', and 'vmlinuz.old'. The dialog box also has a search field, a file type dropdown set to 'Tous les fichiers', and 'Ouvrir' and 'Annuler' buttons.

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

152M of 324M

# IGV : Chargement de l'annotation

The screenshot shows the IGV interface with the 'File' menu open. The 'Load from File...' option is highlighted, and a red box is drawn around the 'File' menu and the 'Load from File...' option. The main window displays a genomic track for chromosome 1, with a 10 mb scale bar and a 90 mb scale bar. The track shows various bands and features, including a red arrow pointing to a specific location. The text 'Charger le fichier GTF, pour avoir la piste d'annotation' is overlaid on the main window, with a blue arrow pointing to the 'Load from File...' option in the menu.

**Charger le fichier GTF, pour avoir la piste d'annotation**

2 tracks loaded | chr1:88 899 520 | 106M of 480M

# IGV : Chargement des alignements

The screenshot displays the IGV interface with a file selection dialog open. The dialog shows a list of files in the 'CORRECTION' directory. The file 'ERR003037.bam' is selected, and its index file 'ERR003037.bam.bai' is also visible. The main interface shows a genomic track with a 10 mb scale and a RefSeq genes track below it.

Rechercher dans :

- bam.intervals
- empty.vcf
- empty.vcf.idx
- ERR000017.bam**
- ERR000017.bam.bai
- ERR000017.fastq
- ERR000017.sai
- ERR000017.sam
- ERR000017\_rmdup.bam
- ERR000017\_rmdup.bam.bai
- ERR000017\_rmdup\_realign.bai
- ERR000017\_rmdup\_realign.bam
- ERR000017\_rmdup\_realign\_re...
- ERR000017\_rmdup\_realign\_re...
- ERR003037.bam**
- ERR003037.bam.bai
- ERR003037.fastq
- ERR003037.sai
- ERR003037.sam
- ERR003037\_rmdup.bam
- ERR003037\_rmdup.bam.bai
- ERR003037\_rmdup\_realign.bai
- ERR003037\_rmdup\_realign.ba...
- ERR003037\_rmdup\_realign\_re...

Nom de fichier : "ERR000017.bam" "ERR003037.bam"

Fichiers du type :

Ok Annuler

RefSeq genes

PKN2 GBP4 LRRC8D BARHL2 HFM1 BRDT GF1 MIF2 BCAR3 ABCD3 ALG14 PTBP2 DPVD MIR2682

2 tracks loaded chr1:88 899 520 106M of 480M

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

# IGV : Chargement des alignements

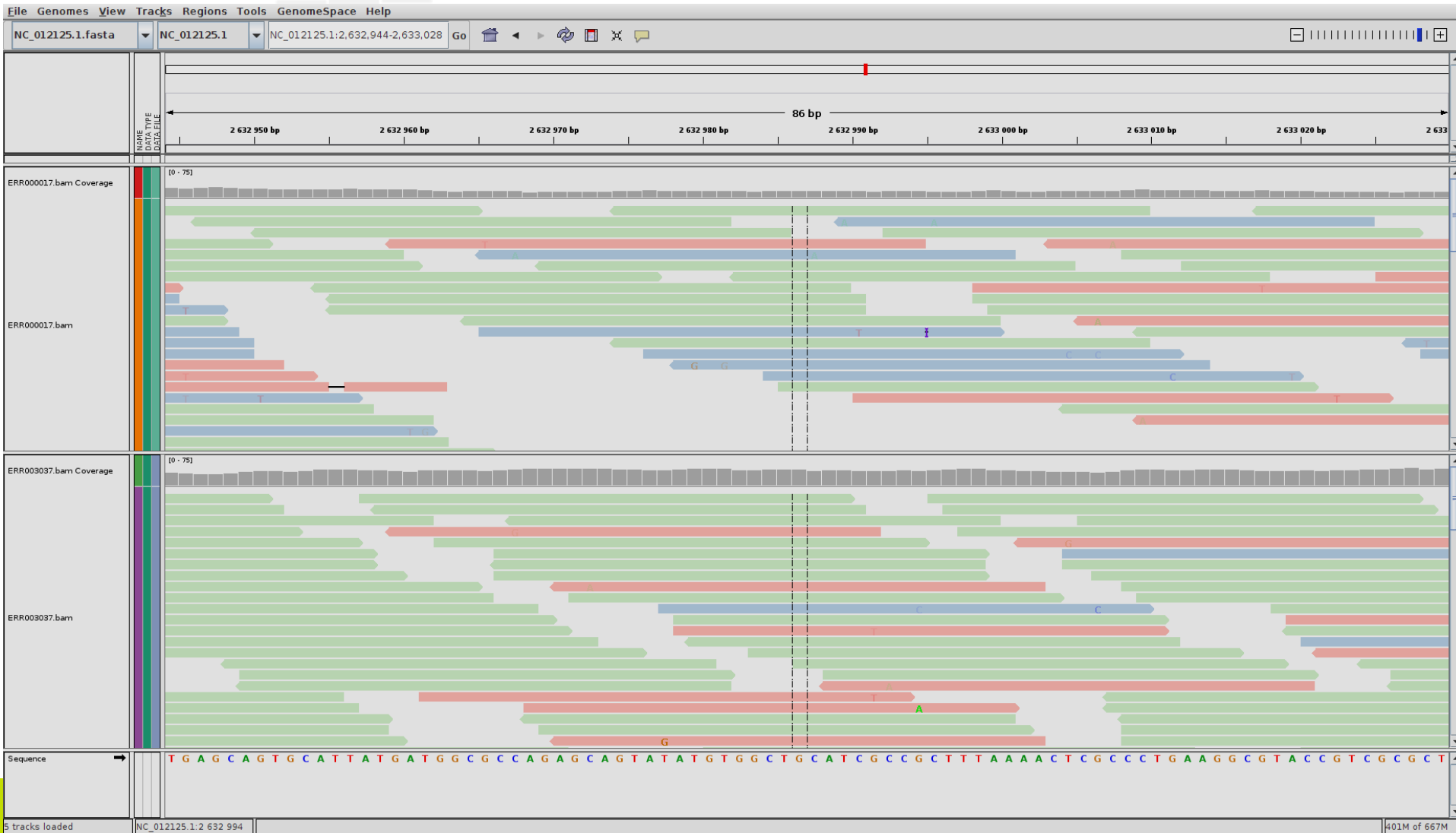
The screenshot displays the IGV interface with the following components:

- Menu Bar:** File Genomes View Tracks Regions Tools GenomeSpace Help
- File Path:** NC\_012125.1.fasta | NC\_012125.1 | NC\_012125.1 | Go
- Genomic Scale:** 0 kb to 4822 kb, with markers at 1000 kb, 2000 kb, 3000 kb, and 4000 kb.
- Tracks:**
  - ERR000017.bam Coverage [0 - 69]
  - ERR000017.bam (Zoom in to see alignments.)
  - ERR003037.bam Coverage [0 - 93]
  - ERR003037.bam (Zoom in to see alignments.)
  - SRR007327.bam Coverage [0 - 30]
  - SRR007327.bam (Zoom in to see alignments.)
- Status Bar:** 7 tracks loaded | NC\_012125.1:26 069 | 200M of 486M

# IGV : Chargement des alignements



# IGV : Chargement des alignements

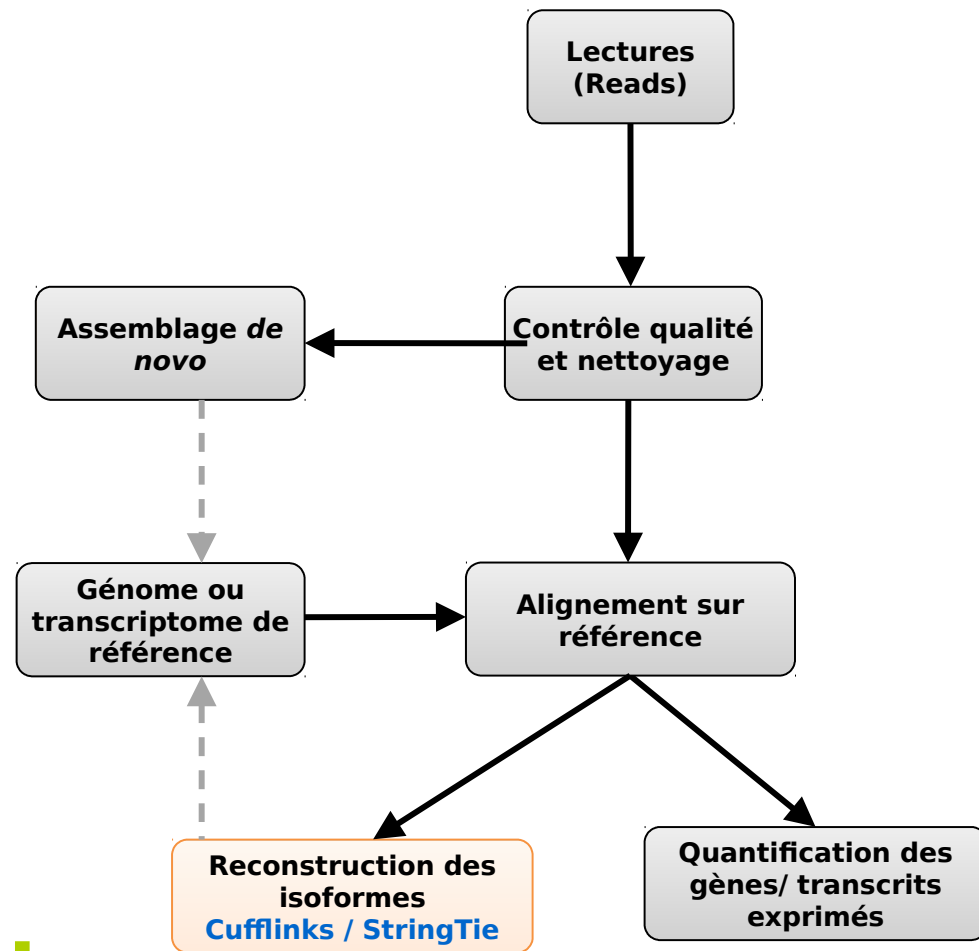




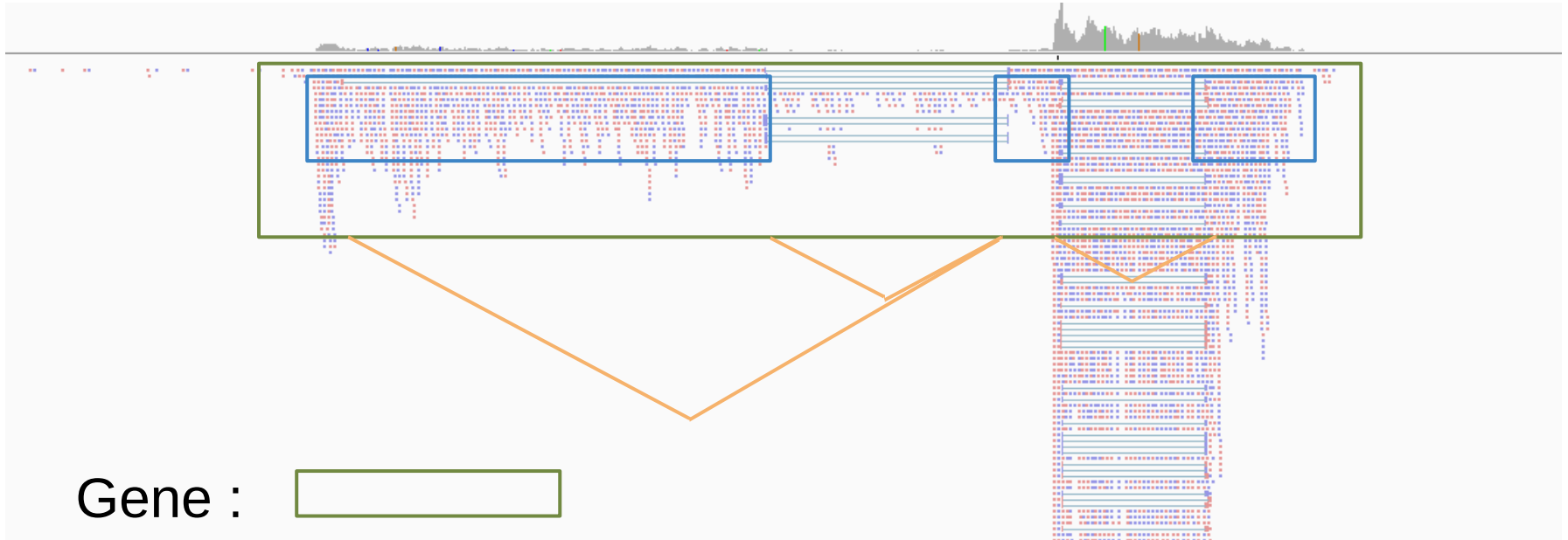
# TP – Visualisation



# 05 Reconstruction de transcript



# Modélisation



Gene :

Exons :

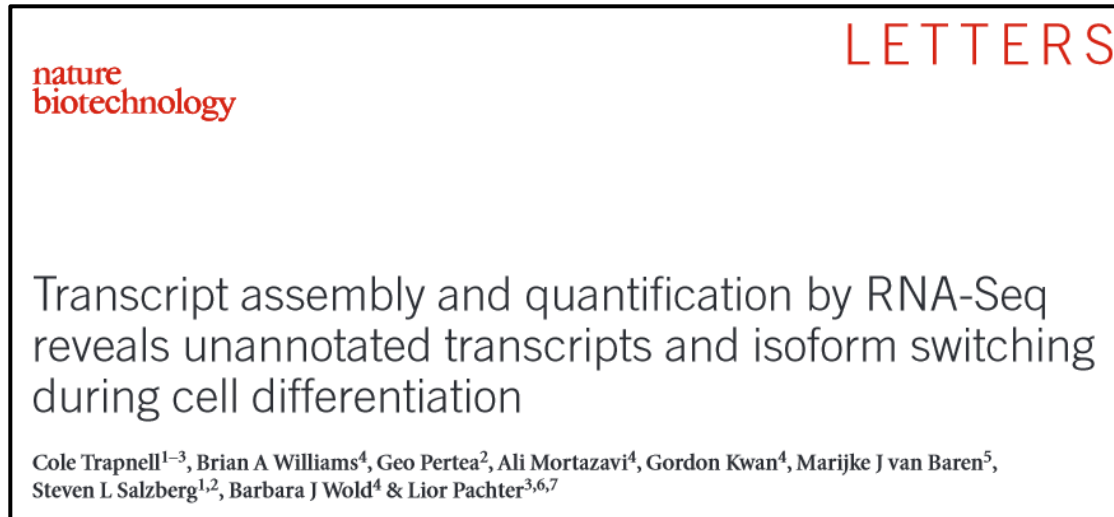
Jonctions (dans les paires & les Reads)



# 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)

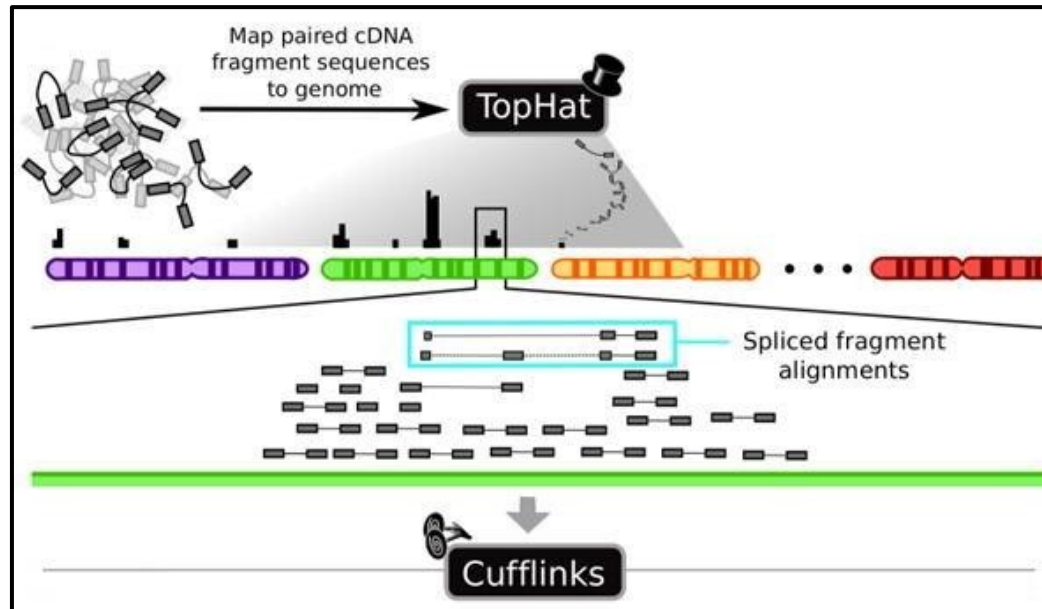


<http://cufflinks.cbcb.umd.edu/>

# Cufflinks

## Reconstruction de transcrits

- ❖ Fragments divisés en **loci non chevauchants**
- ❖ Chaque **locus** est **assemblé indépendamment**

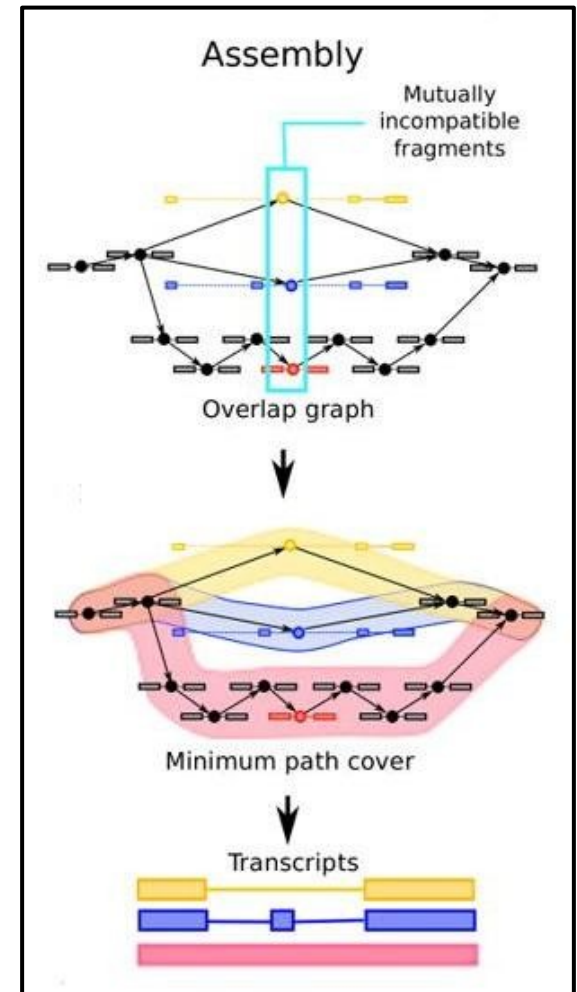


*Trapnell et al. Nat Biotechnol. 2010*

# Cufflinks

## Reconstruction de transcrits

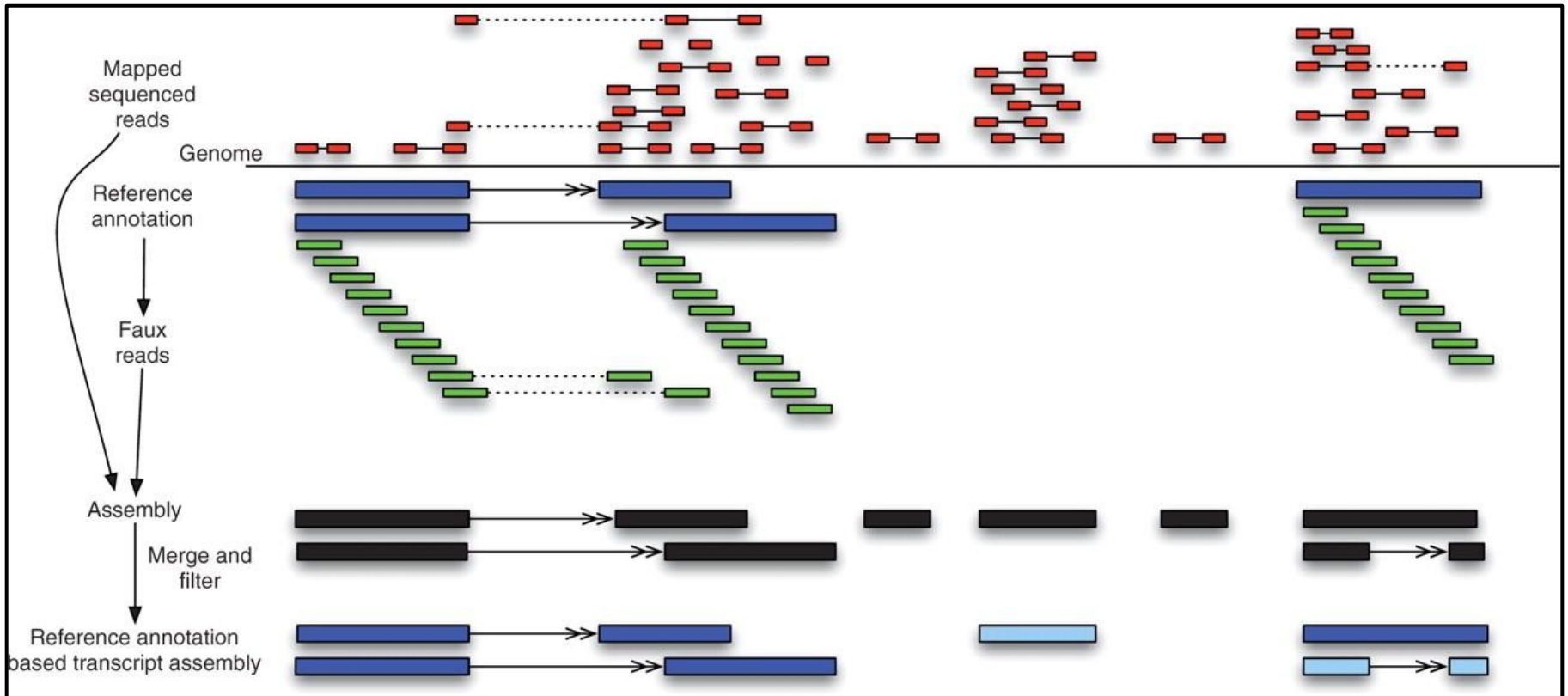
- ❖ **Les différents chemins :**
  - trouver les **positions** des **gènes**
  - trouver les **exons**
  - trouver les **jonctions** :
    - **entre les paires**
    - **dans les séquences**
- ❖ **Stratégie de construction du modèle :**
  - trouver le **nombre minimum de modèles** qui expliquent les lectures :
    - **minimum de chemins**
    - **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 (génomome)
- ❖ Référence gtf (transcriptome)
- ❖ 1 bam par échantillon
- ❖ Quelles sont les stratégies possibles pour identifier le **maximum** de transcrits ?

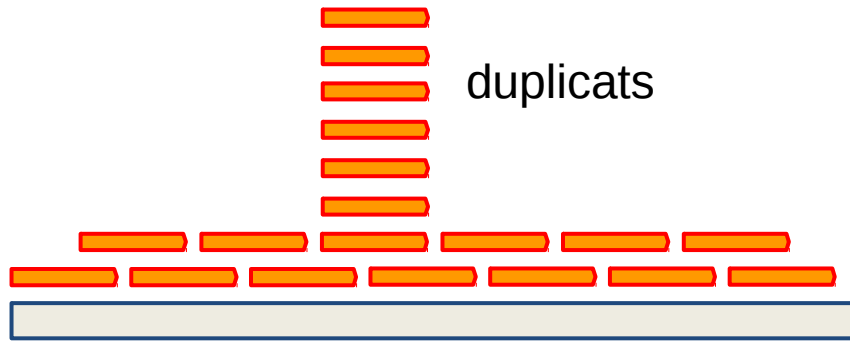
# Fusion d'alignements

- ❖ Samtools : suite logicielle permettant la manipulation de fichiers SAM/BAM/CRAM
- ❖ **Samtools view** : visualisation / conversion
- ❖ **Samtools merge** : fusion de fichiers d'alignement
- ❖ Il existe aussi samtools index, flagstats, rmdup ...



# Données redondantes

❖ Que faire dans ce cas ?



Les duplicats sont dus à des erreurs de préparation ou séquençage.

❖ Cas en pair-ends.



## Reconstruction des transcrits

### ❖ En **entrée** :

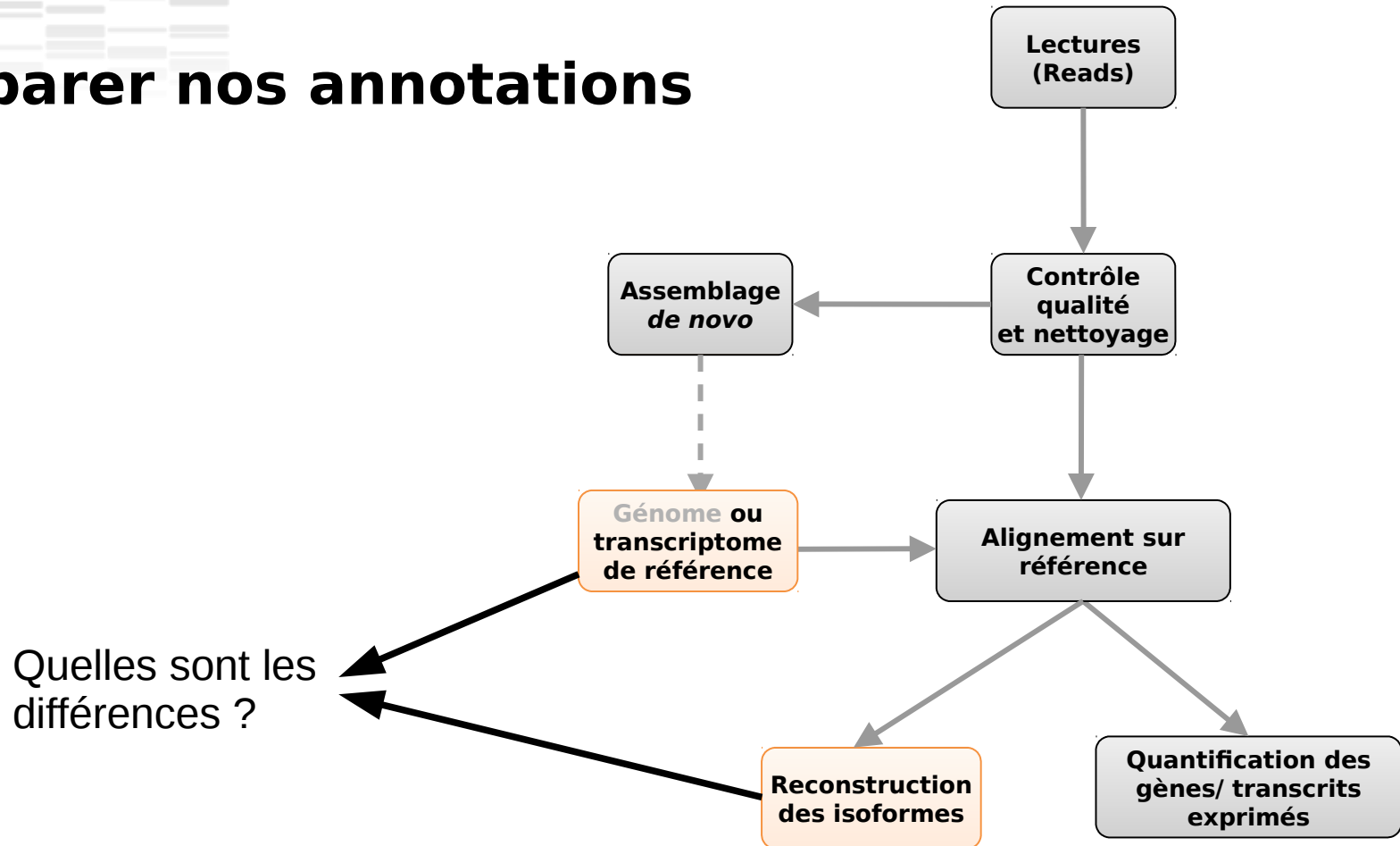
- **lectures (.sam/.bam)**
- Use guide transcript assembly : **annotations (.gtf)**

### ❖ 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

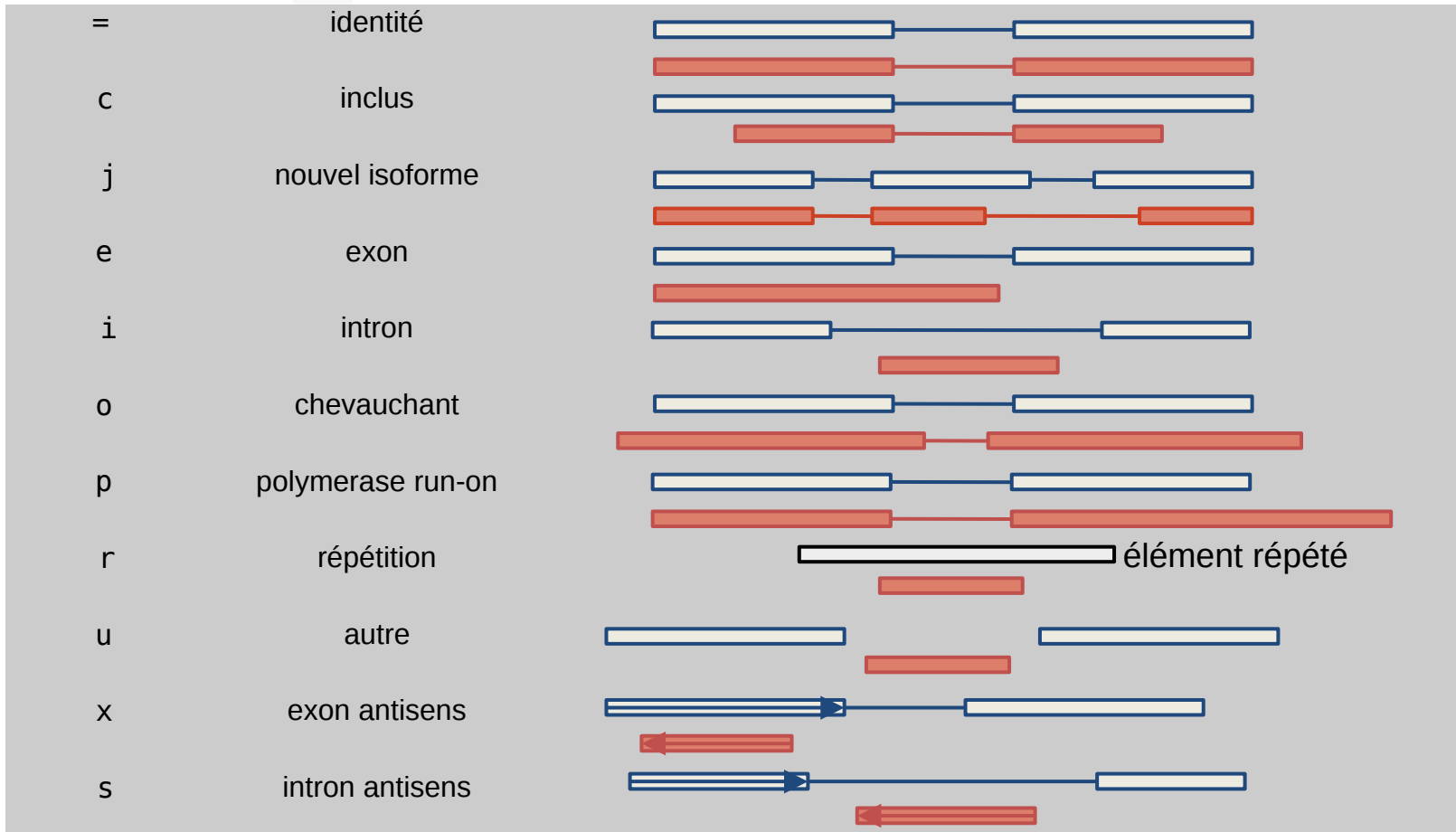
## Comparer nos annotations



Quelles sont les différences ?

# Cufflinks - Cuffcompare

## Class code de cuffcompare



[http://cufflinks.cbc.b.umd.edu/manual.html#class\\_codes](http://cufflinks.cbc.b.umd.edu/manual.html#class_codes)

# StringTie

RNA-Seq reads



Step 1: assemble reads into "super-reads" (optional)

Super-reads

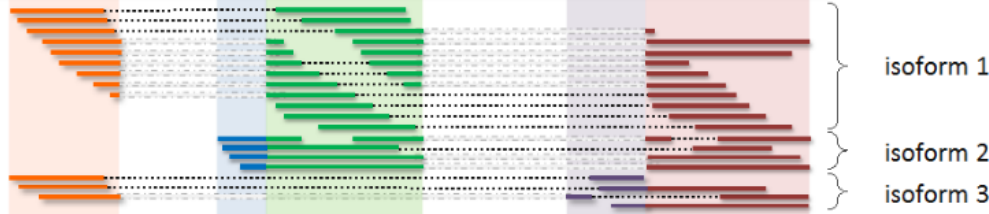


Step 2: map super-reads to the genome

Genome

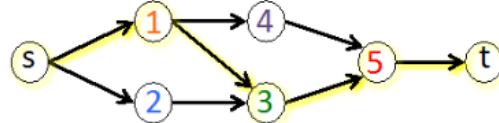


Mapped (super)-reads

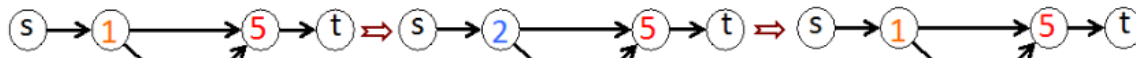


Step 3: build alternative splice graph

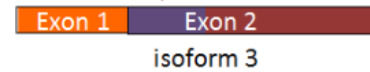
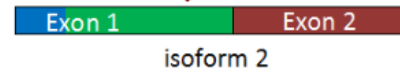
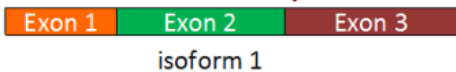
Splice graph with heaviest path highlighted



Step 4: construct flow network for path in splice graph with heaviest coverage



Step 5: assemble transcripts and update coverage



Pertea et al.  
Nature  
Biotechnology  
2015



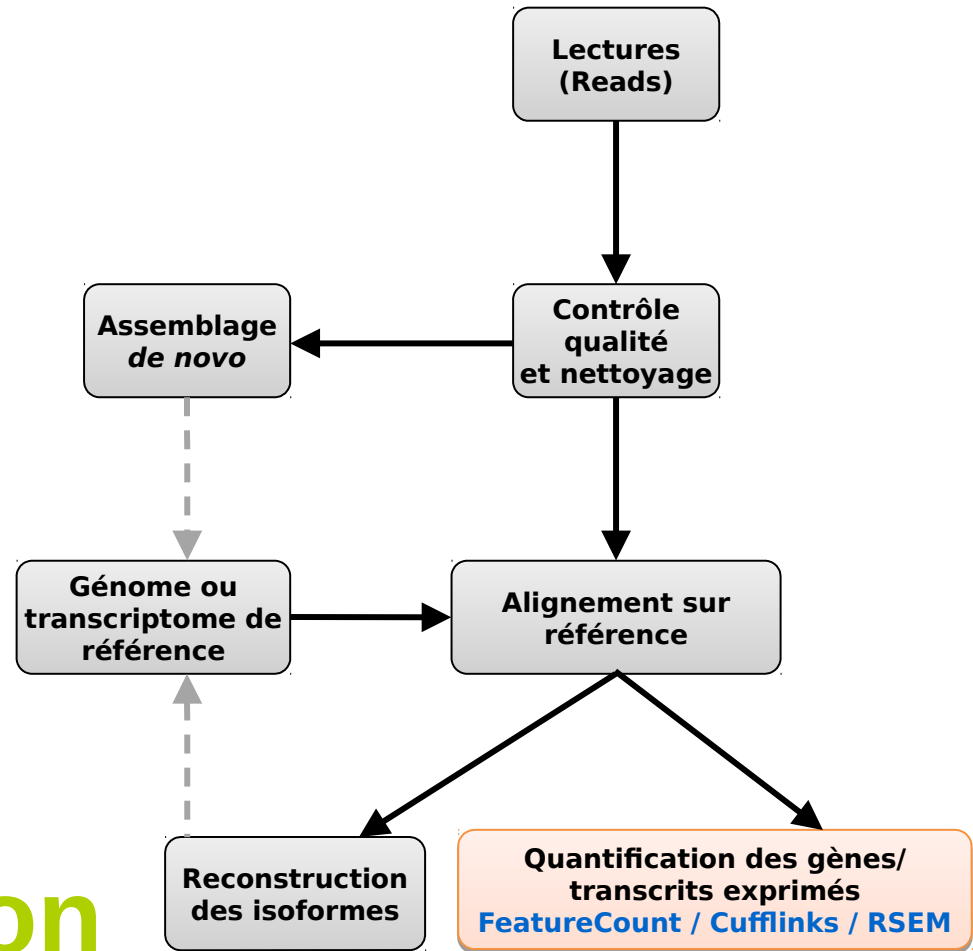
# Découverte de transcrit: quelle methodologie?

- ❖ Fusionner les alignements
- ❖ Supprimer les duplicats
- ❖ Détecter les nouveaux transcripts
- ❖ Ouvrir le nouveau transcriptome dans IGV



# TP - Découverte de transcrit

# 06 Quantification





# Quantification

## Que cherche-t-on à compter ?

### ❖ Quel *feature* compter ?

- gènes
- exons
- transcrits

chr2L	Drosophila_melanogaster.BDGP5.25.62.gtf.gz	aggregate_gene	7529	9484	.	+	.	gene_id "FBgn0031288"
chr2L	Drosophila_melanogaster.BDGP5.25.62.gtf.gz	exonic_part	7529	8116	.	+	.	transcripts "FBtr0300609+FBtr0300690"; exonic_part_n
chr2L	Drosophila_melanogaster.BDGP5.25.62.gtf.gz	exonic_part	8193	8589	.	+	.	transcripts "FBtr0300609+FBtr0300690"; exonic_part_n
chr2L	Drosophila_melanogaster.BDGP5.25.62.gtf.gz	exonic_part	8590	8667	.	+	.	transcripts "FBtr0300609"; exonic_part_number "083";
chr2L	Drosophila_melanogaster.BDGP5.25.62.gtf.gz	exonic_part	8668	9484	.	+	.	transcripts "FBtr0300609+FBtr0300690"; exonic_part_n
chr2L	Drosophila_melanogaster.BDGP5.25.62.gtf.gz	aggregate_gene	9836	21372	.	-	.	gene_id "FBgn0021211"
chr2L	Drosophila_melanogaster.BDGP5.25.62.gtf.gz	exonic_part	9836	11344	.	-	.	transcripts "FBtr0078170+FBtr0078171+FBtr0078169+FBt
c_part_number "001"; gene_id "FBgn0021211"								
chr2L	Drosophila_melanogaster.BDGP5.25.62.gtf.gz	exonic_part	11410	11518	.	-	.	transcripts "FBtr0078170+FBtr0078171+FBtr0078169+FBt
c_part_number "002"; gene_id "FBgn0021211"								
chr2L	Drosophila_melanogaster.BDGP5.25.62.gtf.gz	exonic_part	11779	12221	.	-	.	transcripts "FBtr0078170+FBtr0078171+FBtr0078169+FBt
c_part_number "003"; gene_id "FBgn0021211"								
chr2L	Drosophila_melanogaster.BDGP5.25.62.gtf.gz	exonic_part	12286	12928	.	-	.	transcripts "FBtr0078170+FBtr0078171+FBtr0078169+FBt
c_part_number "004"; gene_id "FBgn0021211"								
chr2L	Drosophila_melanogaster.BDGP5.25.62.gtf.gz	exonic_part	13520	13625	.	-	.	transcripts "FBtr0078170+FBtr0078171+FBtr0078169+FBt
c_part_number "005"; gene_id "FBgn0021211"								
chr2L	Drosophila_melanogaster.BDGP5.25.62.gtf.gz	exonic_part	13683	14874	.	-	.	transcripts "FBtr0078170+FBtr0078171+FBtr0078169+FBt

### ❖ Comptage brut sur les gènes ou les exons ou les transcrits:

- **featureCount**

### ❖ Estimation de l'abondance des transcrits reconstruits :

- **Cufflinks, RSEM**

### ❖ Comptage brut des multimaps

- **mmquant**

### ❖ Dépend des données disponibles

gene_id	untreated1	untreated2	untreated3	untreated4	treat
FBgn0000003	0	0	0	0	1
FBgn0000008	92	161	76	70	140
FBgn0000014	5	1	0	0	4
FBgn0000015	0	2	1	2	1
FBgn0000017	4664	8714	3564	3150	6205
FBgn0000018	583	761	245	310	722
FBgn0000022	0	1	0	0	0
FBgn0000024	10	11	3	3	10
FBgn0000028	0	1	0	0	0
FBgn0000033	1446	1712	615	672	1608

# featureCounts

- ❖ Mieux que 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 read 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

## Minimum 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 ▾

**Minimum 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:**

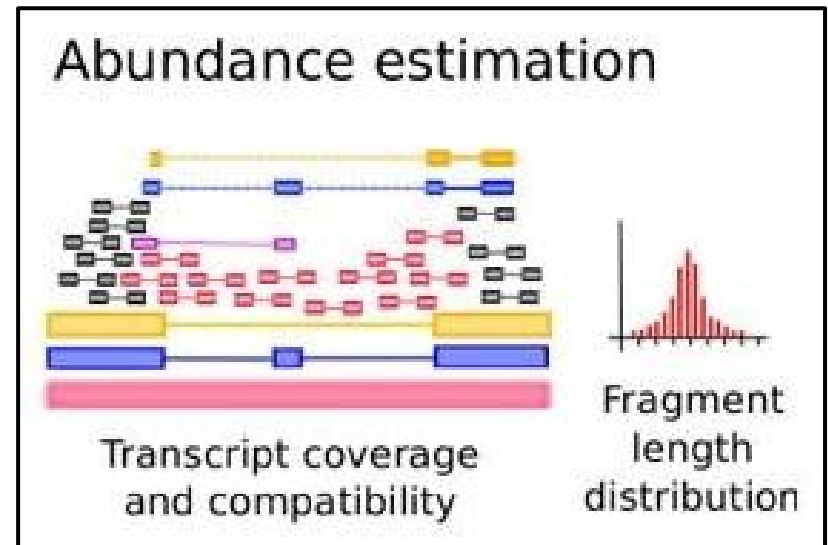
The chimeric fragments will NOT be included ▾

Execute

# Cufflinks

## Principes

- **Assignation des lectures** à un transcript
- **Estimation** de l'**abondance** de **chaque transcript** mesurée en :
  - **RPKM** (*single reads*)
  - **FPKM** (*paired-end reads*)



*Trapnell et al. Nat Biotechnol. 2010*

❖ 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

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

### ❖ **FPKM** :

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

❖ Pas de possibilité d'utiliser les packages R: EdgeR ou Deseq. (Utiliser cuffdiff)

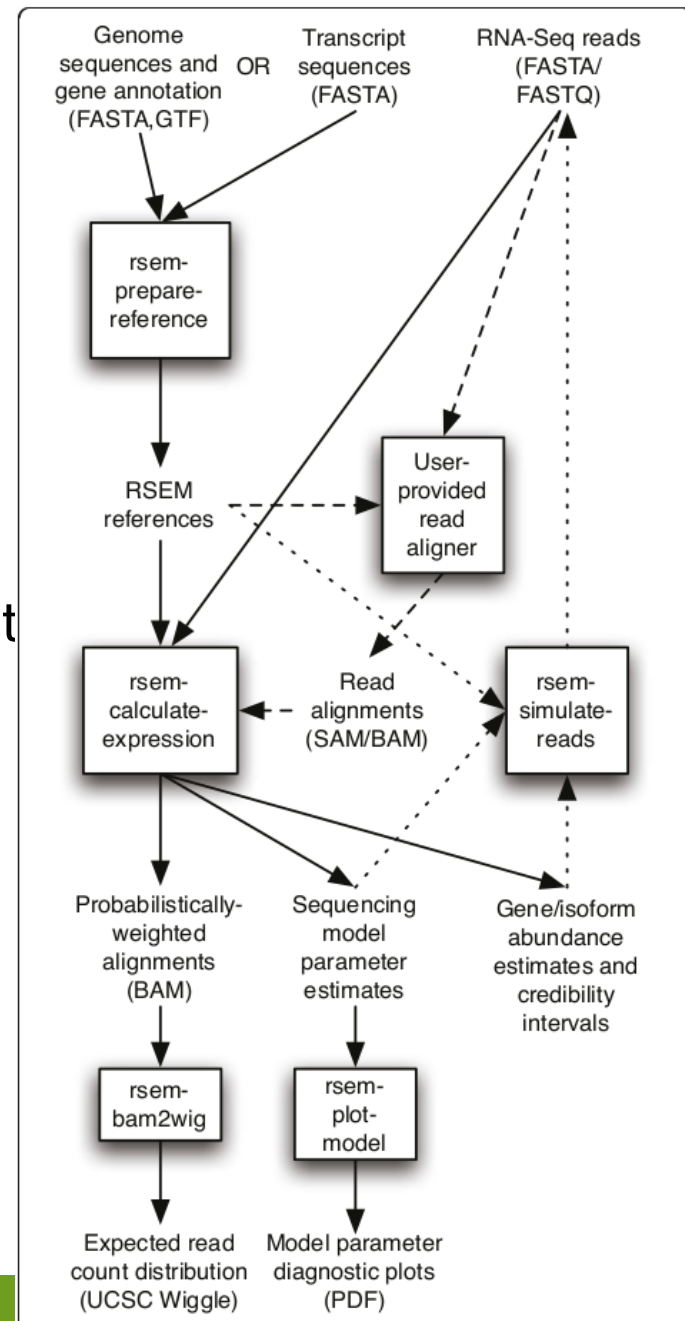
*Mortazavi et al. Nature Methods 2008*

# RSEM



- ❖ Logiciel qui fait :
  - l'alignement
  - l'estimation des isoformes
- ❖ Travaille uniquement sur un alignment contre le transcriptome.
- ❖ Une fois les estimations arrondies: EdgR , Deseq
- ❖ Préconisé par ENCODE3

*Li & Dewey. BMC Bioinformatics, 2011*





# TP - Quantification



# **\_07**

## **Partie stat**

# **\_07**

# **Conclusion**

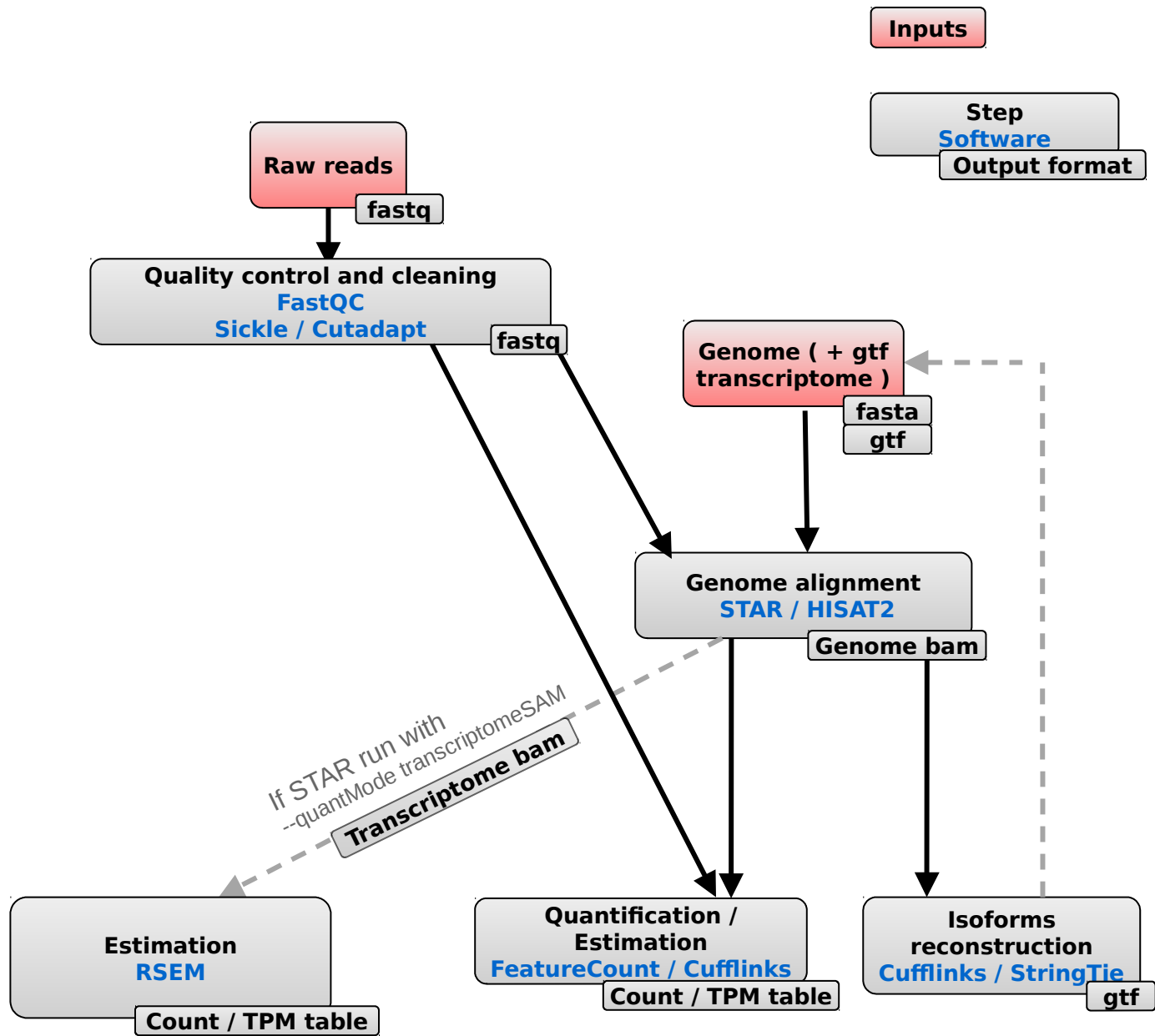
# Conclusion générale

- ❖ Workflow galaxy à construire
- ❖ Choix des outils dépendent des données disponibles et de la question biologique

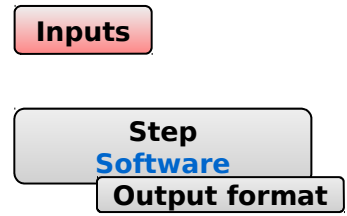
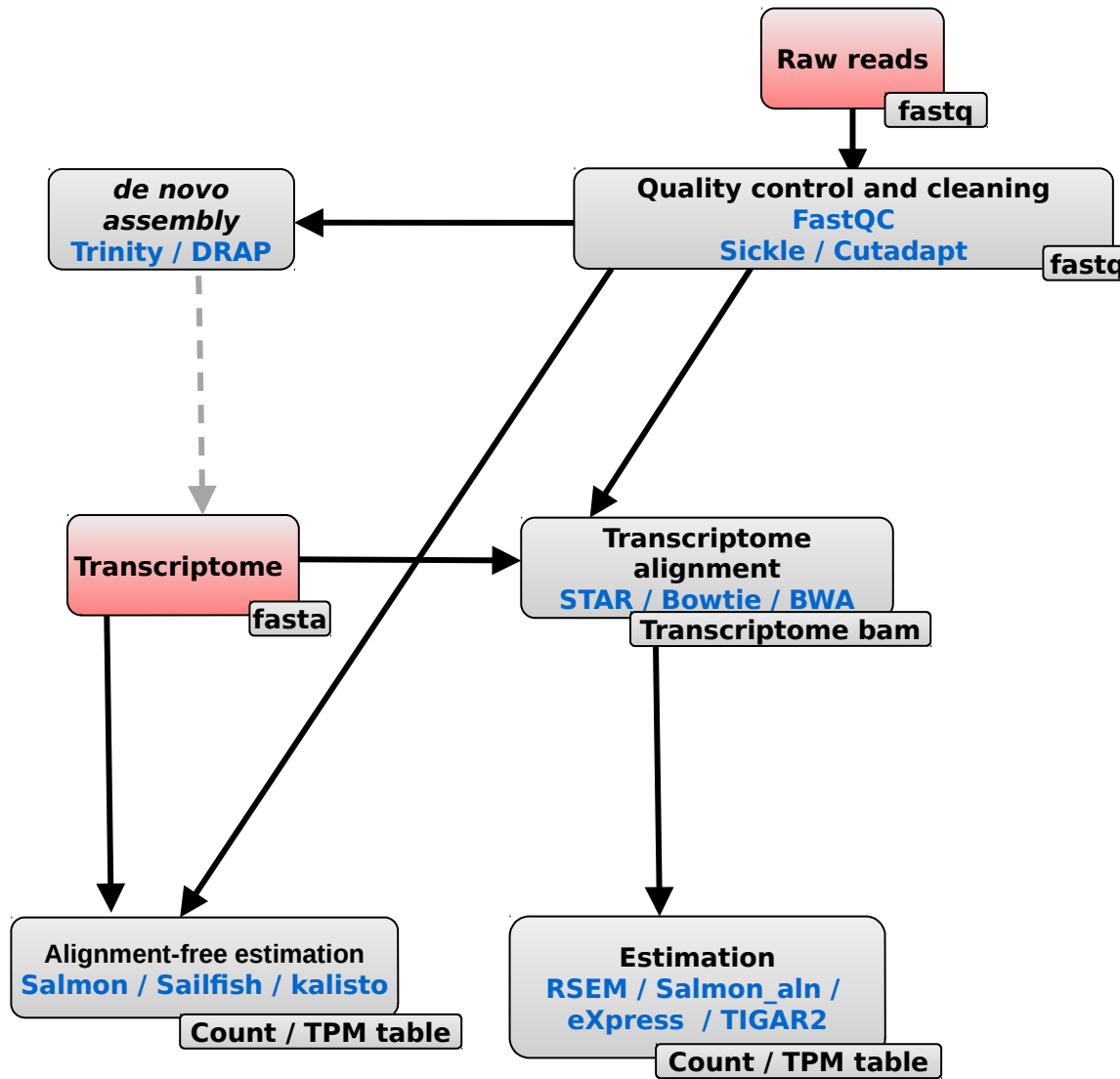
Tous les outils sont dispo sur Migale et Galaxy

- ❖ Et maintenant en avant pour les stats !

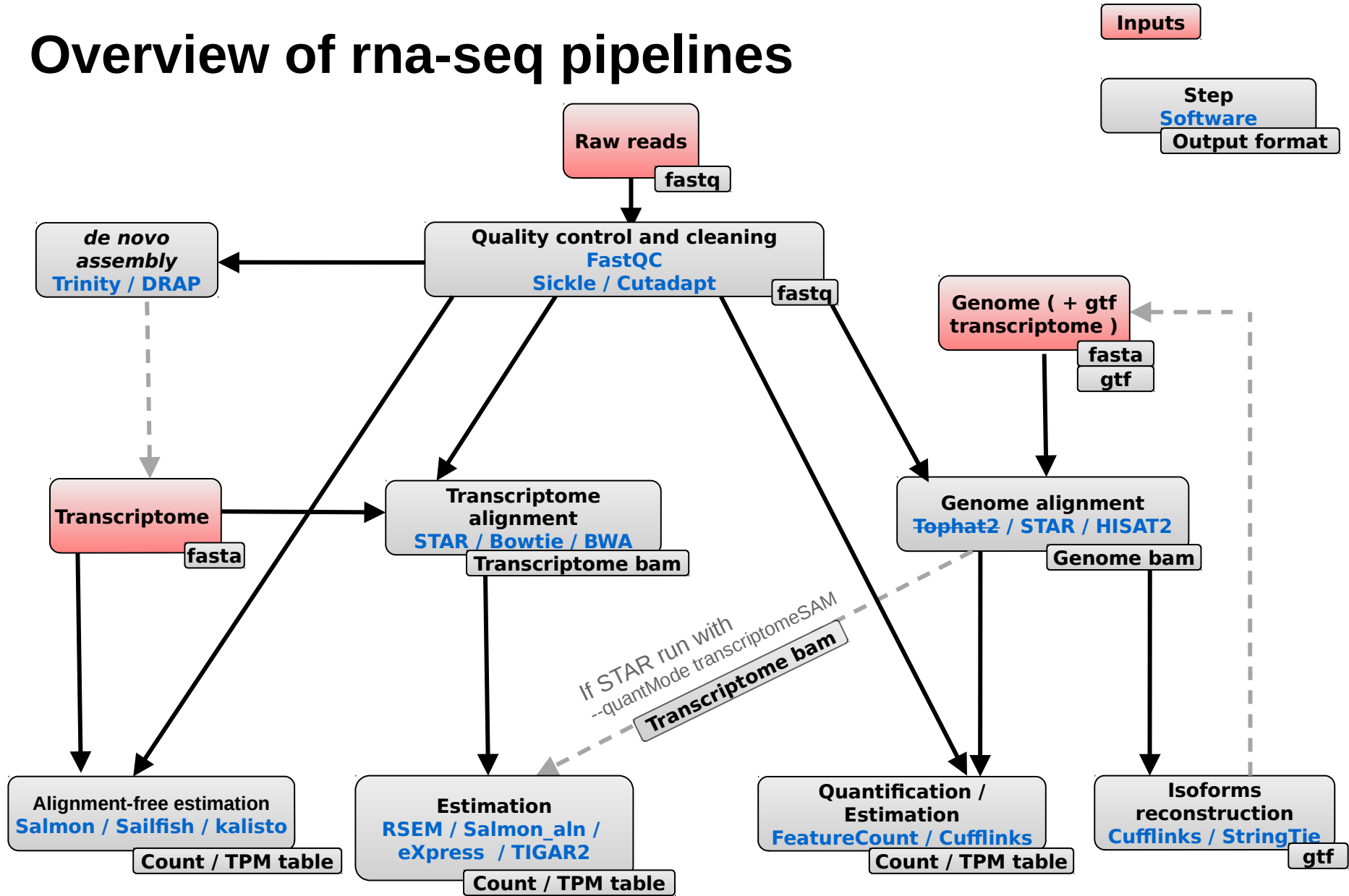
# What we did



# If no genome



# Overview of rna-seq pipelines



## Liens utiles

- ❖ **Seqanswers** : <http://seqanswers.com/>
- ❖ **Biostar** : <https://www.biostars.org/>
- ❖ **RNA-Seq blog** : <http://rna-seqblog.com/>

## Remerciements

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

Satisfaction form :

[http://genoweb.toulouse.inra.fr/~formation/4\\_Galaxy\\_RNAseq/2018/doc/questionnaire.html](http://genoweb.toulouse.inra.fr/~formation/4_Galaxy_RNAseq/2018/doc/questionnaire.html)